The evolution of *Campylobacter*

Submitted in partial fulfilment of the requirements
for the PhD in Statistical Genetics

mEpiLab, Infectious Disease Research Centre
Institute of Veterinary, Animal and Biomedical
Sciences
at Massey University,
New Zealand.

Shoukai Yu

2012

Abstract of : The evolution of *Campylobacter*

Author : Shoukai Yu

Date : 2012

The genus *Campylobacter* is a major cause of human gastroenteritis worldwide, so understanding the evolution of *Campylobacter* has important implications. This multidisciplinary project unifies developments from statistics, genetics, bioinformatics and computer science and creates a good opportunity to investigate the evolution of *Campylobacter* by focusing on the factors which affect genetic exchange.

In order to understand how *Campylobacter* evolves, a mathematical method is put forward to estimate the relative rates of recombination and mutation in generating new alleles that lead to single locus variants (SLVs), and examine the effect of selection, recombination and mutation. This analysis shows the importance of recombination in the evolution of *Campylobacter* and larger contribution made by recombination, compared to mutation, in the evolution of *Campylobacter jejuni*, and *Campylobacter coli*. In addition, this research demonstrates that purifying selection plays an important role in the evolution of *Campylobacter*. For comparison, this analysis also examined the role played by recombination in the evolution of other bacteria. This application highlighted the importance of recombination for creating diversity in closely related isolates.

A range of phylogenetic and population genetic tools were applied to investigate the effect of geographical isolation on the evolution of *Campylobacter* by comparing datasets from two geographically separated countries, New Zealand and the United Kingdom, this is the first time this has been attempted. Analysing sequence data at different levels of resolution provided evidence that geographical isolation affects the evolution of *Campylobacter* genotypes over short time-scales, but that this effect diminishes over longer time-scales. Furthermore, this analysis estimates the time for divergence of NZ specific lineages of *Campylobacter* strains.

In New Zealand, *Campylobacter jejuni* strain type 474 (ST-474) is responsible for more than a quarter of human campylobacteriosis notifications, but has been rarely found outside NZ. Knowing the clonal relationships of ST-474 strains is helpful for inferring the origin and the evolutionary mechanism of *Campylobacter*. This research accessed 59 isolates of *Campylobacter*. It applied a range of phylogenetic tools to targeted gene reference set to compare estimations of the clonal genealogy inferred for *Campylobacter* datasets.

These findings have implications for identifying the origin of *Campylobacter*, developing disease intervention strategies, predicting the emergence of pathogens, and reducing the occurrence of campylobacteriosis in the food supply chain.

Acknowledgements

Firstly, I would like to thank my excellent supervisors, Professor Nigel French, Dr Barbara Holland, Dr Patrick Biggs, Prof Paul Fearnhead, and Dr Grant Hotter for their help, encouragement and all the guidance throughout the PhD project, which made my time in New Zealand wonderful and meaningful.

Secondly, I am grateful to Marsden project 08-MAU-099 (Cows, starlings and Campylobacter in New Zealand: unifying phylogeny, genealogy and epidemiology to gain insight into pathogen evolution) for funding. I would like to show my gratitude to Institute of Veterinary, Animal and Biomedical Sciences for several times funding for conference and research. I also would like to thank the financial support from the Maurice & Phyllis Paykel Trust.

Thirdly, I appreciated all of the PhD support from Massey University Doctoral Research Committee.

Fourthly, I am truly thankful for all of the effort made by the mEpiLab staff to produce the comprehensive datasets. It is a great pleasure to thank to all the administrative staff and my colleagues for these fantastic years.

In addition, I am grateful to World Health Organization for the internship opportunity.

At last, I want to thank my parents for their love, encouragement and all the support in my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 General background

The genus *Campylobacter* is a leading cause of gastroenteritis worldwide [4, 37, 197, 363]. The species *Campylobacter jejuni* and *Campylobacter coli* are the main causes of bacterial food-borne disease in developed countries, compared to other members of the family *Campylobacteriaceae* [137, 208]. The genome of *Campylobacter* is relatively small (1.6 to 1.7 million basepairs) [42, 285, 286]. However, *Campylobacter* is a prominent human pathogen. Understanding the evolution of *Campylobacter* has important implications in a wide range of areas, such as epidemiological investigations, bacterial speciation, and policy development to minimise the impact of emerging pathogens [71, 145].

The evolution of *Campylobacter* has been affected by both mutation and recombination [72, 99, 154, 194, 243, 250, 251, 325, 357, 390], and *Campylobacter jejuni* evolves rapidly [334, 403]. Recombination plays a dominant role in the evolutionary process of diversity of *Campylobacter* genotypes, although mutation is the way to create a new allele or alter a gene [71, 403]. Recombination within and between *Campylobacter* spp. can occur by natural transformation, conjugation and/or transduction [71, 368]. It has been recently proposed that frequent recombination between *C. jejuni* and *C. coli* has resulted in the convergence of the two closely related zoonotic pathogenic species [335]. This speciation reversing process could be a result of the change in environment caused by human farming activities [334, 335].

Within a bacterial species, a clonal complex is a cluster of closely related bacterial strains that group around a founder (or ancestral) strain. Estimates of the rate of DNA sequence evolution at the clonal complex level have been found to be faster than estimates obtained using more distantly related isolates [71], perhaps due to the lack of time for purifying selection to act. By making use of DNA sequence data

gathered as part of multilocus sequence typing (MLST) schemes, we can focus on pairs of strains within clonal-complexes that share a very recent common ancestor. These pairs of closely related strains are known as single locus variants (SLVs) as they differ at only one gene of the seven genes used in the MLST scheme, and most SLVs occur within a clonal complex. The relative rates of recombination and mutation in generating SLVs can be estimated using model-based methods. These estimates can reflect the evolution of more closely related sequence types. Preliminary analyses on estimating the relative rate of recombination to mutation are based on a limited number of sequences [101, 388]. This PhD project will put forward a mathematical method to estimate the relative rates of recombination and mutation in generating SLVs, and examine the effect of selection, recombination and mutation.

The availability of worldwide *Campylobacter* databases has facilitated research on the evolution of *Campylobacter* [69, 403]. It has been demonstrated that host association plays a more important role than geographical separation in the evolution of *Campylobacter* [332], but little is known about what effect geographical location has had on the evolution of this globally distributed bacteria (*Campylobacter*). As important zoonotic pathogens, *C. jejuni* and *C. coli* have caused gastroenteritis internationally, and research has been done separately in different countries (New Zealand (NZ) [123, 246, 260, 262], the United Kingdom (UK) [122, 239, 333, 403], Africa [199, 200], European countries (Finland [193], Norway [163, 318], Switzerland [209], Denmark [232]) and the US [255, 367]). These studies show that prevalence and incidence rates varied between different countries. However, the differences in the sampling process among sampling areas, human population settlements, and sampling methods make the comparison among countries very difficult.

The unique geographical location of NZ and its distinctive history with the introduction of European wild life and livestock provide a good opportunity to investigate the role played by geographical isolation on the evolution of *Campylobacter* by comparing the features/characters of current *Campylobacter* spp. in NZ and an equivalent dataset from the UK. This PhD project will apply a wide range of tools from phylogenetic networks, coalescent theory and population genetics to investigate the effect of geographical separation by comparing equivalent datasets from NZ and the UK. These two datasets were sampled equivalently according to several factors such as the size of the area from which the isolates were obtained, the mix of urban and rural areas, and the time over which the sample was collected.

The development of molecular typing methods also offers the chance to further investigate the evolution of *Campylobactor* [102]. With the development of sequencing technology and bioinformatics, whole genome sequence data is becoming available. MLST schemes were a good start pointing to build a worldwide database for the

investigation on the evolution of the globally distributed bacteria. The availability of whole genome sequence data allows the investigation of how the analysis and results for MLST type data extend across the genome. This comparison creates an opportunity to compare the results of the existing analytical tools and gain more information about the evolution of *Campylobactor*.

MLST datasets are the main data available for analysis of population structure and molecular epidemiology of *Campylobacter* species. MLST schemes only cover 400 to 500 basepairs (bp) for each of seven genes, and only cover around 0.2% of the genome [118]. Furthermore, seven housekeeping genes are selected for MLST, but different types of genes may undergo significantly different selection pressure. Therefore, the opportunity of working on large scale datasets can make it possible to draw inferences about the clonal genealogy of *Campylobactor*, based on more genes and information. After 15 years, the MLST technique has become popular in bacterial studies, and this technique allows comparison across laboratories worldwide. At the time of writing (November 27, 2012), there are 6194 strain types for *Campylobactor jejuni/coli* available in the public MLST database.

For this PhD project, the aim is to improve understanding of the evolution of *Campylobacter* species, by combining the recent development from both the sequencing and statistical analysis areas. This is a multidisciplinary project. Since the 1960s, statistics, genetics, bioinformatics and computer science have developed rapidly. Unifying all the developments together creates a good opportunity to investigate the evolution of *Campylobacter*. Two factors make this PhD project unique.

Firstly, this PhD project can access the full genome of 59 isolates and MLST data sets from NZ and the MLST datasets from UK for comparison. These 59 isolates include several ST-474 genomes. ST-474 is defined by MLST scheme, and it is commonly found in NZ occurrence of campylobacteriosis but infrequently in any other countries. NZ data were mainly produced by Massey University's Molecular Epidemiology and Public Health Laboratory (mEpiLab). This laboratory has the largest *Campylobacter* dataset in the Southern hemisphere and this dataset contains the five year Manawatu Sentinel site study (>5000 isolates) for a range of host sources: humans, livestock animals, poultry, environmental water and wild birds.

The second relates to NZ's unique historical and geographical location and high rate of *Campylobacter* infection [260] during the period 1984 to 2010. This thesis will compare the features and the evolutionary paths of NZ and the UK *Campylobacter* datasets. This analysis is the first attempt to study the role played by geographic isolation on the evolution of *Campylobacter*. The findings could be helpful in disease control and intervention and are useful in determining the origin of *Campylobacter*.

There are two main threads going through the whole thesis: three comparisons and

three methodological stages. The three comparisons are the comparison between the NZ and UK *Campylobacter* datasets, the comparison between the analysis on MLST and the full genome sequencing datasets, and the comparison among a range of analytical methods and models from phylogenetics, population genetics and statistics.

The comparison between NZ and UK *Campylobacter* datasets is like comparing the results of a huge experiment, one that has been set up in two distant locations, thousands of years ago which continues to the present. One task of this project is to trace back the distinct experiment and report the difference of the two locations. The comparison between MLST and extended genome sequencing datasets provides an opportunity to utilise the data produced by advanced sequencing techniques to answer some questions which could not be answered before, such as inferring the clonal genealogy or phylogenetic relationships for different sequence types. The comparison among analytical tools can compare the methods, based on different model assumptions, and can test for consistency among different analytical theories.

The three methodological stages mean that firstly, the SLV model analysis is only on the MLST dataset, then the phylogenetic methods are on both MLST and extended genome sequencing datasets, and lastly the phylogenetic methods are used on the extended genome sequencing datasets only. Compared to the limited number of genes (seven loci) for MLST data, the current extended genome sequencing datasets have been greatly updated and extended by using genome resequencing technology to provide more information about over one thousand loci across the whole *Campylobacter* genome.

If being able to control epidemics is a long journey, this PhD project is part of that journey, and it can make a contribution to achieve the One World One Health goal, which is considering the human, animal and environmental health (ecosystems) as a unity. Using the evolution of *Campylobacter* as a paradigm, similar analytical methods could be extended to other bacteria. From this starting point, with the ultimate goal of investigating how and why *Campylobacter* emerged to become such a prominent human pathogen, we will improve our understanding of the evolutionary mechanisms of pathogens, be able to predict the emergence of pathogens, and reduce the occurrence of campylobacteriosis in the food supply chain.

1.2 Objectives

In this thesis, three main objectives are achieved.

1. A new statistical method is put forward to estimate the ratio of recombination rate vs. mutation rate to generate single locus variants (SLVs) in the evolution of *Campylobacter*. Through simulation, the accuracy of the new method is evaluated. The new method is then applied to different bacteria. Knowing the relative rates of mutation and recombination will improve the understanding of how important recombination is relative to mutation for the generation of new strains.

2. This PhD project investigates the role played by geographical separation to the evolution of *Campylobacter* by applying a range of phylogenetic and genealogical methods to data. This objective makes use of the unique geographical isolation and the distinctive history of New Zealand. This analysis will help develop understanding of the effect of geographical isolation on the evolution and diversity of globally distributed bacteria.

3. Unifying all the most recent developments together from a multidisciplinary perspective, we can access and analyse sequence data derived from 59 whole genomes of *Campylobacter*. These 59 isolates include ST-474, which is a sequence type relevant to the occurrence of campylobacteriosis in New Zealand. These isolates help estimate the refined clonal genealogy of ST-474, a sequence type commonly found in NZ, but rarely found anywhere else in the world. The availability of whole genome sequence data allows the comparison of the results for MLST type data and for more gene sets.

1.3 Organisation of the thesis

This thesis has seven chapters. The next chapter is a literature review related to the area of the evolution of *Campylobacter*. The third chapter puts forward a method to estimate the ratio of recombination to mutation for closely related strains. Chapter three is the basis of the research for chapter four. The fourth chapter extends this method to different bacteria. The fifth chapter analyses the different evolutionary paths for New Zealand (NZ) *Campylobacter* data and United Kingdom (UK) *Campylobacter* data. The sixth chapter compares the existing five methods to infer the genealogy of the given sequence types (STs), then compares the results of different methods on the whole genome data set. This whole genome data set is like an extension of current widely-used multi locus sequence type (MLST) datasets. Chapter

seven provides the summary of this research and further suggestions. Chapters 3, 4, and 5 are either published or in a form where they are about to be submitted and therefore each chapter has been written to stand alone so the introductions in the three chapters contain some overlapping material.

Chapter three has been published, I thank all of the co-authors (Prof Paul Fearnhead, Dr Barbara Holland, Dr Patrick Biggs, Prof Martin Maiden, and Prof Nigel French) for sharing their valuable insights and expertise. Their comments and help greatly improved my original manuscript. The majority of the work reported in Chapters four to six has been done by myself, but these chapters benefited from my supervisors (Dr Barbara Holland, Prof Paul Fearnhead, Dr Patrick Biggs, and Prof Nigel French), who provided valuable ideas and assistance to the undertaking of the research summarised here.

Chapter 2

Literature review

2.1 *Campylobacter*

2.1.1 General information

The genus *Campylobacter* belongs to the family *Campylobacteraceae*, order *Campylobacterales*, class *Epsilon Proteobacteria* of the phylum *Proteobacteria*. In 1963, the genus *Campylobacter* was created (*Campylobacter* means "curved rod" in Greek) [117, 387]. *Campylobacter* spp. are gram-negative, spiral-shaped, microaerophilic bacteria, and they are the major cause of human bacterial gastroenteritis worldwide [267]. They inhabit the intestinal tracts of warm-blooded animals. *Campylobacter* can be spread among animal populations through drinking at a common water source or by contact with infected faeces [3, 92, 130, 133, 397]. The World Health Organization (WHO) reports that *Campylobacter* is one of the most common causes of zoonotic enteric infections worldwide [399]. The species *C. jejuni* and *C. coli* are the main causes of bacterial food-borne disease in developed countries, compared to other members of the family *Campylobacteraceae* [208]. In 2002, the reported campylobacteriosis rate of New Zealand was ten times higher than that of the US, and more than two or three times than that of other industrialized countries including the UK [10, 269]. The differences in reporting systems and methodology can only partially explain the observations ([222]; cited by [10]).

History

The first reported description of *Campylobacter* can be traced back to 1886, when a nonculturable spiral-shaped bacterium was observed by Theodor Escherich [383]. It was not until 1957 that *Campylobacter* spp. was identified as a cause of human

enteric disease [205], though at that time, *Campylobacter* was referred to as "*Vibrio*" spp. [387, 383]. In 1963, the new genus *Campylobacter* was proposed by Sebald and Véron (1963, cited by Véron [387]), due to some of their biological structures differing from other *Vibrio* species. *Campylobacter* was included in the family *Spirillaceae* in 1973, due to the morphological and physiological likeness between the genera *Campylobacter* and *Spirillum* [387]. In the 1980s, *Campylobacter* species were determined as one of the most common causes of human enteric disease worldwide [7]. Since 1990, genomic sequencing techniques have been applied to differentiate and build the phylogenies of microorganisms. In 2000, the first *Campylobacter jejuni (C. jejuni)* genome was sequenced (NCTC 11168) [286], a milestone in *Campylobacter* genetics.

In 1991, the taxonomy of genus *Campylobacter* was revised [386] and the new bacterial family *Campylobacteraceae* was put forward [384]. The *Campylobacteraceae* family contain *Campylobacter*, *Arcobacter*, and *Sulfurospirillum* [267]. In 2010, the misclassified *Bacteroides ureolyticus* was reclassified into *Campylobacter ureolyticus* [385]. The genus *Campylobacter* currently includes the species: *C. fetus*, *C. hyointestinalis*, *C. lanienae*, *C. sputorum*, *C. mucosalis*, *C. mucosalis*, *C. concisus*, *C. survus*, *C. retus*, *C. gracilis*, *C. rectus*, *C. hominis*, *C. jejuni*, *C. coli*, *C.lari*, *C. insulaenigrae*, *C. canadensis*, *C. upsaliensis*, and *C. helveticus* [267] and 13 others[1].

Morphological characteristics

The family *Campylobacteraceae* have the following characteristics: cells are spirally curved rods that are 0.2 to 0.8 μm wide, 0.5 to 5 μm long, are gram-negative, and nonsaccharolytic [267]. Most members of the genus *Campylobacter* are motile, using unipolar or bipolar flagella [286], though some of species are nonmotile (*Campylobacter gracilis*) [267].

Campylobacter are microaerophilic: they require low oxygen concentrations to survive [196, 267]. Some members of the genus *Campylobacter* can cause human and animal infections.

2.1.2 *Campylobacter* epidemiology

Campylobacteriosis is the disease caused by *Campylobacter* bacteria. Approximately five to ten percent of campylobacteriosis cases result in hospital admission [339]. In Australia and the UK, campylobacteriosis is responsible for most hospitalizations caused by bacterial infection. In the US *Campylobacter* is only second

[1]The genus *Campylobacter* currently contains 32 species and 13 subspecies. URL: http://www.bacterio.net (March 16, 2012)

to salmonellosis as the cause of hospital admission [151]. The incidence of *Campylobacter* spp. enteritis is probably underestimated, because the report rates in many countries are quite low [393]. There is a surveillance pyramid to describe the reported infection rate. The diagnosed and reported true infection rate is only a fraction of the successfully grown *Campylobacter* cells recorded in hospital. Those successfully grown cells are only a fraction of cultured specimens, and those specimens are only a fraction of the sick people who go to hospital, those who attended hospital is only a fraction of the people who are infected by *Campylobacter* bacteria [267].

Out of 32 species in the *Campylobacter* genus, *C. jejuni* and *C. coli* are the two main human gastroenteric pathogens, and they are estimated to be responsible for more than 95% of food-borne diseases caused by *Campylobacter* spp. [7, 72, 124].

The seasonality of campylobacteriosis has been researched by several studies in different countries [10, 267, 279]. In many studied countries, such as UK, USA, and New Zealand, there is a peak for incidence of campylobacteriosis in the warmer months, but the shape and the size of the peaks vary between latitudes and regions [267]. The reason for the seasonality in *Campylobacter* infection remains elusive, but it may occur because human activities differ with the seasons and exposure to bacteria increases in summer. Additionally, the prevalence rates at non human reservoirs may be affected by temperature and humidity [156, 319], or the transmission medium, such as flies, may vary among the seasons [85, 272].

A range of factors that might impact on the infection rate of *Campylobacter* have been considered. The age and gender distribution of *Campylobacter* infection has been investigated [267]. Generally, males have a higher infection rate compared to females [267, 340]. The incidence rate of campylobacteriosis is much higher for preschool children than adults [**?** 124, 379]. In developing countries, the situation of *Campylobacter* infection is worse than that in developed countries [280]. The incidence rate is two to eight times as much as in the developing countries for different age groups [7]. Compared to other developed countries, NZ has a higher reported incidence rate of *Campylobacter* infection [11]. Even after 1990, when *Campylobacter* culturing became routine, the number of infections reported in NZ in 2003 was three times higher than 1991 [11, 12]. In 2003, the incidence rate of *Campylobacter* infection in UK and USA was around 50 cases per 100,000, while in NZ the incidence rate was 396 per 100,000.

It has also been shown that an infectious dose of *Campylobacter* can be as low as 500 [307] or 800 [267] organisms. Previous research also indicates a naive population tends to get infections more easily at a lower infectious dose, than a previously exposed population [29, 365, 366]. Despite this research apart from the improvement

of detection and reporting systems, the reported incidence rates of *Campylobacter* continued to rise in many developed countries, including NZ between the 1980s and mid 2000s.

The symptoms of *Campylobacter* infection

The symptoms of *Campylobacter* infection include: abdominal cramps, abdominal pain, diarrhea (with or without blood), fever, headache, nausea, and vomiting [271]. Some symptoms are quite similar to appendicitis, and usually last three to six days [264, 392]. In rare cases, infection in very young children or elderly patients could be fatal, and with serious cases can lead to several months painful inflammation of the joints or neurological disorders such as Guillain-Barré syndrome (GBS) [6]. GBS was first described in 1916 [406]. The symptoms of GBS are similar to ascending paralysis, dysaesthesias usually below the waist at the early stage, and may result in respiratory and severe neurological dysfunction or death in a number of cases [6]. Mortality is rare, although the mortality rate for those who had a *Campylobacter* infection within one year was three times higher than those who did not get infected. In addition, most fatal cases occur in the elderly or those suffering from other serious diseases [159].

Source, risk factors and transmission

Campylobacter exists widely in most warm-blooded domestic and wild animals. Both *C. jejuni* and *C. coli* have been found in wide range of hosts, such as cattle, sheep, poultry, cats and dogs [243, 271]. There are multiple sources and pathways for the transmission route of *Campylobacter* spp., including food, water, contaminated soil, animal contact and person to person transmission. The food-borne transmission route is regarded as the primary route [**?** 13, 81], such as consuming undercooked meat products and contaminated milk and water.

The relative contribution of each of these sources (poultry, bovine, ovine and environment) to the overall burden of human disease has been studied [191, 257, 260]. There is a significant association between campylobacteriosis and contact with raw or undercooked poultry products. *Campylobacter* spp. has been found in many food types, such as raw milk, beef, lamb and seafood. Cross contamination can occur in the slaughtering process of red meat producing animals when muscle tissue is in contact with intestinal contents [353]. Cooked and frozen meat products are not reservoirs of *Campylobacter* spp., except for sporadic situations caused by cross-contamination from raw meat products [353].

Most *Campylobacter* infections are reported in sporadic infections, rather than outbreaks. Sporadic infections have been documented from contact with animals, such as poultry [86, 87, 302], livestock [347], and pets [318]. Some cases are associated with drinking raw milk or untreated water [81], and even swimming in natural water [267].

The contamination of *Campylobacter* to surface water and the terrestrial environment are through the faeces of infected animals and birds. *Campylobacter* can be found in most rivers in New Zealand [322]. In NZ, children aged 1-4 years were more at risk than other age group [11]. The fecal material from wild birds in children's playgrounds could be a contributor to that high risk, as pre-school children may ingest infective material, through their frequent behaviour of hand-mouth contact. Wild birds inhabiting public areas are recognized carriers of *Campylobacter* [123]. The environmental exposure to faeces from livestock, such as ruminants [321], could also be one of important contributors to human infection, although it is not a dominant contributor compared to food-related exposure [13, 81, 123]. Secondary transmission from human to human is not common in *Campylobacter* infections.

Burden of *Campylobacter* infections and economic effect

The research into the disease burden for *Campylobacter* infections incorporates epidemiology, statistical modelling, and assessment of risk factors, and can be measured by morbidity and mortality. The research findings can advise policy makers by increasing their understanding of the disease. This is important for the development of better prevention strategies, more efficient allocation of resources and more accurate measures for food safety. A proper estimation of the global burden caused by foodborne diseases is needed for reallocating resources related to control and prevention of diseases, and policy making. It was not until 2012 that the initiative for reliable epidemiological measurements was launched by WHO [213, 330]. However, the effect of the global burden of foodborne diseases is still unclear.

Campylobacter causes a great burden to public health. The cases of campylobacteriosis cause a huge economic impact in Australia, the UK, the US, and NZ. It was estimated that the cost of campylobacteriosis is around $40 million dollars each year in NZ [326, 407], and this accounts for 73% of food borne illness costs [326].

Control and prevention of *Campylobacter* infections

Control and prevention methods can be applied at different levels of food processing, and the prevention of infection can be applied at all stages of the food chain, from farms to factories. A hygienic abattoir environment can be helpful to reduce the

transmission from faeces to carcasses, but cannot reduce the presence of *Campylobacter* in the meat products. Bactericidal treatment, like heating well or irradiation have been proven to be useful for eliminating *Campylobacter* from food production [17, 19, 21]. The prevention of *Campylobacter* infection also requires the attention from both the commercial and household kitchens. At the household level, eliminating *Campylobacter* from contaminated food is a solution by such methods as heating the food well before consumption.

Only a small number of *Campylobacter* bacteria can make most people produce the *Campylobacter* infection symptoms. For individuals, consuming pasteurized milk, well-heated red meat and poultry, and drinking treated water can reduce the risk of *Campylobacter* infection.

Intervention

Since 2000, the reported cases of campylobacteriosis have been stable or even slightly in decline in some countries after prevention efforts, especially in the poultry industry [134, 311].

In NZ, there has been a significant decline in campylobacteriosis notifications since 2006 [327]. Considering the unchanged laboratory practices, hospitalization rate of other enteric diseases, such as salmonellosis and cryptosporidiosis, this very likely reflects a true decline in underlying disease incidence [327]. It also indicates the possible reasons for this decline: the cooperation between scientific area, policy and specific industry actions [327]. Research on source attribution of *Campylobacter* and surveillance have played an important role for the development of regulatory actions to reduce campylobacteriosis [13]. The decline can be highly associated with the research [13, 81, 405] which identified poultry as the primary source of campylobacteriosis in New Zealand via foodborne transmission [12].

The New Zealand Food Safety Authority developed a risk management policy in 2006, and this policy aimed at reducing campylobacteriosis attributed to NZ poultry industry [328]. A significant decline in campylobacteriosis was observed after these interventions [328], and a source attribution study show a marked decrease attributed to poultry by 2010 [121].

2.1.3 Molecular biology of *Campylobacter*

The availability of whole genome sequencing technology had increased our understanding of the evolution, genotype, and phenotype of *Campylobacter*. Since 2000, the whole genome sequence for different species of *Campylobacter* became available:

C. jejuni [286], *C. coli*, *C. lari*, *C. upsaliensis* [118], etc. Because *C. jejuni* and *C. coli* are the main causes of bacterial food-borne disease in developed countries, compared to other members of the family Campylobacteriaceae [208], the following part will gave a brief introduction of molecular biology of *C. jejuni* and *C. coli* only.

Campylobacter jejuni genome biology

In 1990, Nuijten et al. proposed the genome sizes of *C. jejuni* and *C. coli* are about 1.7 Mb each, and they built the first physical map of *C. jejuni* (UA580) [277]. Parkhill et al. [2000] published the complete sequence of the *C. jejuni* (NCTC 11168) genome. Fouts et al. [2005] in one of the early studies on *Campylobacter* genome sequencing provided a core genetic blueprint of the genus by comparing five sequenced *Campylobacter* genomes. The analysis of *Campylobacter* genomes also included the development of the systems for strain typing, which can benefit further research in phylogenetics, epidemiology, source tracking, and public health [118]. One example of the genome characteristics from some reference genomes of *C. jejuni* was given in Biggs et al. [2011], such as genome ID, genome length (Mb), the number of genes, the number of sequences.

Molecular evolution of *C. jejuni* and *C. coli*

Dingle et al. [2005] reported the extension of the multilocus sequence typing (MLST) technique to include *C. coli*, which allows *C. jejuni* and *C. coli* to be compared. The gene *flaA* alone cannot tell the difference between *C. jejuni* and *C. coli*. It is crucial to understand the molecular evolution of *C. jejuni* and *C. coli*, since both of them are responsible for a large percentage of gastroenteritis worldwide. *C. jejuni* and *C. coli* share 86.5% nucleotide sequence identity at the MLST housekeeping gene level. Dingle et al. [2005] also point out that there is no apparent clustering of STs by source in *C. coli*, which contradicts previous research [223, 230] on the association of *C. coli* strains and host sources using amplified fragment length polymorphism typing (AFLP).

Sheppard et al. [2008] put forward the possible convergence of *C. jejuni* and *C. coli*. Wilson et al. [2009] applied several statistical models to reveal the importance of recombination in *C. jejuni*, and state that the divergence of *C. coli* from *C. jejuni* occurred around thousands of years ago rather than millions of years ago. In terms of biochemical characteristics, *C. coli* are quite similar to *C. jejuni*, except that *C. coli* cannot hydrolyse hippurate, although some *C. jejuni* strains cannot either [383].

2.1.4 Flagella and the major outer memberane proteins

Campylobacter pathogenesis can cause gastrointestinal disease. It colonizes the mucus lining of the gastrointestinal tract. The polar flagella of these pathogens provide the necessary motility for intestinal colonization. Early research on the importance of flagella in the evolutionary process of *Campylobacter* can be traced back to 1990 [148, 160, 204, 278]. In 2007, Guerry [147] summarized the multifaceted role of the polar flagella in *Campylobacter* virulence.

Flagellin genes (*flaA* and *flaB*)

The flagellin gene of *Campylobacter* has two similar copies: *flaA* and *flaB*. The length of coding regions for the *flaA* and *flaB* sequences are both around 1.7 kilobases, and *flaA* and *flaB* sequences locate about 180 bases apart from each other [249]. In *Campylobacter,* previous research indicated that the evolution of the *flaA* and *flaB* genes is coordinated [249].

Concerted evolution occurs when the expected divergence of copies of genes within an individual is less than the divergence of the gene from other species [227]. It was shown that segments of *Campylobacter fla* show concerted evolution occurring at a rate that is larger or equal to the rate of clonal divergence. Compared to the expected diversity for most of the *Campylobacter* genome, flagellin clearly has a greater diversity, although the information on the divergence rate of other genes in *Campylobacter* is not currently available [249].

The major outer memberane proteins (MOMPs) and its encoded gene (*porA*)

As mentioned previously, *Campylobacter* spp. are gram-negative bacteria, and have outer memberane proteins (OMPs) [35, 60, 181, 283, 417]. The major outer memberane proteins (MOMPs) have unique structural features, and function as porins which are helpful for linking up the bacteria and their environment. In 2000, it was put forward that *Campylobacter*'s MOMP may be crucial for the bacteria to adapt to various host environments, and it was proved that a single locus gene (later, it was defined as *por*A [48]) encodes the MOMP [417]. Research into MOMP and its encoding gene (*por*A) will be useful for the development of diagnostics and vaccines that are based on MOMP [417].

Research to date has focussed on different aspects of its structure and function. Clark et al. [2007] put forward the phylogenetic relationship obtained by *por*A sequencing. In 2007, three separate lineages of *por*A were found and defined [48]. The absence

of recombination within *por*A clade 1 and 2 suggests there are constraints on the MOMP structure, and the existence of a purifying selection. Different clusters of MOMP sequences have different functions in their biological properties [48]. This may be useful for research into bacterial ecology or virulence [48].

In conclusion, combined with multilocus sequence typing (explained in section 2.3), the *porA* gene can provide additional and useful information to the further research in epidemiology area. The *porA* gene can be helpful for dividing *Campylobacter* into subgroups by the important functional or virulence properties, which will enable the research on *Campylobacter* evolution to reach a new level of phylogenetic differentiation that has not been found by other typing methods.

2.2 Typing methods

This section outlines the molecular techniques that can be used to provide data for determining evolutionary processes. Further background of the data source can be helpful for making sensible assumptions in a biological model. Molecular typing techniques can be applied to investigate the short-term (or epidemic) epidemiology of disease; they can be also applied to study the long-term or global epidemiology [315]. For the short-term epidemiological investigation, molecular typing techniques are used to recognize the common strain types and possible sources of the infection. For the long-term epidemiology investigation, molecular typing techniques are used to identify emerging or re-emerging strains [315].

There are many molecular typing techniques available, and they can be divided into two broad categories: phenotyping and genotyping. Serotyping [49, 125, 231, 290] is one of widely used phenotyping methods for *Campylobacter*. The occurrence of nontypeable organisms limits the application of serotyping. Furthermore, serotyping is labour-intensive and costly, which makes this method impractical for many clinical laboratories [315].

In general, genotyping methods have greater discriminatory power than phenotyping methods [135, 284, 288, 349, 393]. Genotyping methods can be further divided into two categories: band-based methods and sequencing methods [114]. Depending on the length of the gene regions that are used, band-based methods are separated into three categories: single locus, multiple loci and whole genome. *fla*A restriction fragment length polymorphism typing (*fla*A RFLP) [154, 152] is an example of a single locus genome typing method. Flagellin locus restriction fragment length polymorphism (*fla*-RFLP) method is a band-based method, which is difficult to standardise among different laboratories. The *fla*-RFLP method has been proved

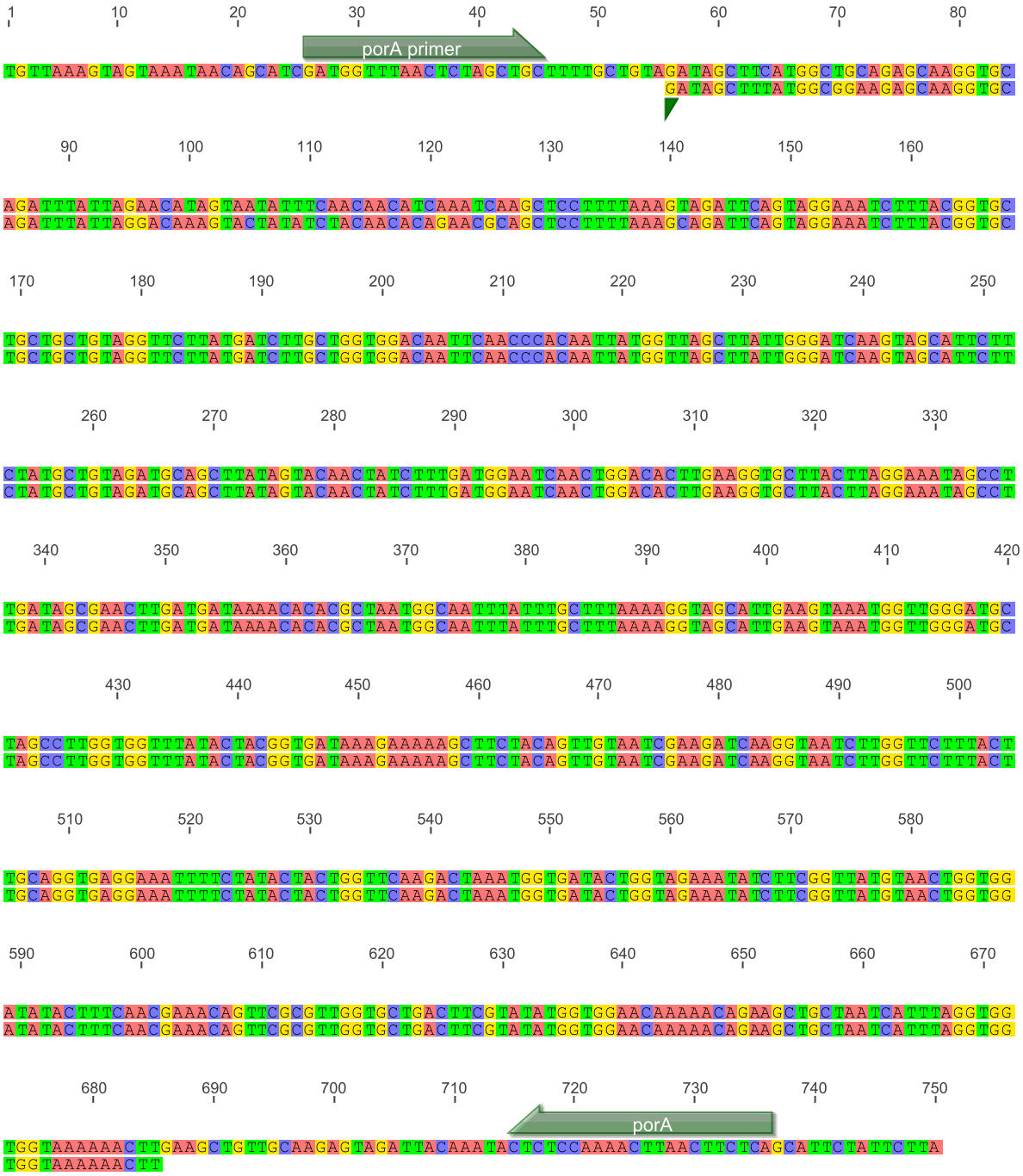


Figure 2.1: A portion of the full gene *por*A on *Campylobacter jejuni* strain NCTC11168

to be useful in epidemiological investigation and strain discriminating in outbreak analysis [49, 153, 248, 266, 295]. The advantages of the *fla*-RFLP method are low cost, rapid results, high throughput, and these properties are required in epidemiological analysis when an outbreak has just occurred. Furthermore, it could be used to predict clonal complexes [267].

Multiple loci genome typing methods include multilocus enzyme electrophoresis (MLEE) [1, 142, 329] and whole genome methods include amplified fragment length polymorphism typing (AFLP) [8, 26, 66] and pulsed-field gel electrophoresis (PFGE) [138, 283, 393]. Compared to AFLP, PFGE is more labour intensive [49]. PFGE highlights all of the variations between strain types, which is difficult to extract cluster information from pathogens. Thus PFGE is too discriminatory for long term epidemiology because its results cannot indicate that isolates have come from clonal lineage [237, 268]. Therefore, PFGE is more useful over very short-time scales but is not an appropriate method for evolutionary studies. PFGE is useful in defining population structure, but is labour intensive [49].

There are several types of sequencing methods, including multilocus sequence typing (MLST), flagellin short variable region (*fla*-SVR) and *porA* sequencing methods. Compared to PFGE, the *porA* sequencing method can provide equivalent typeability, discriminatory power, and more accurate results when allocating isolates into different groups [417]. The *fla*-SVR sequencing method has a higher discriminatory power than MLST, and similar to PFGE, because this method subtypes strains within sequence types. The portability and easy interpretation of *fla*-SVR enables this method to be applied widely in the analysis of surveillance networks. Compared to *fla*A-SVR typing, *fla*B-SVR typing is less discriminating, but more suitable for outbreak investigation [266, 267]. After polymerase chain reactions (PCR), *fla*-RFLP and *fla*-SVR are two options for further investigation. *fla*-SVR overcomes some of the limitations in *fla*-RFLP, but *fla*-SVR detects less diversity than *fla*-RFLP [267]. Clark et al. [49] suggest *fla*-SVR sequence typing and *fla*-RFLP methods can both useful in outbreak analysis, and large-scale surveillance. The following sections will describe MLST in detail as that was the methods used for the analysed datasets.

2.3 Multilocus sequence typing (MLST)

There are several methods which could be used for molecular typing [356]. For this project, the data are produced by one molecular typing technique: multilocus sequence typing (MLST). A major advantage of MLST that is the results are comparable among different laboratories [237], and information can therefore be shared worldwide.

Although MLST was originally designed for identifying lineages of pathogens, many other applications have shown a significant usefulness of this method in a wide range of bacterial related study areas [237, 267].

2.3.1 Selection of MLST

As MLST provides data that can be used in long-term or short-term studies, it was selected for this study on *Campylobacter*. MLST is now a universally accepted system to characterize many bacterial species, including *Campylobacter*. MLST has many advantages over other methods for this research project, for example: a large number of *Campylobacter* are not typeable by serotyping [325]. Although the operation of AFLP is easier, faster and cheaper, AFLP is difficult to compare the results between laboratories due to variation of DNA sequence data caused by experimental differences [325].

MLST and multilocus enzyme electrophoresis (MLEE) operate on the same principle, as both of them sequence fragments of multiple housekeeping genes selected across the genome [89]. MLEE is a band-based typing method, and it measures electrophoretic mobilities of selected metabolic enzymes. MLST is a sequence-based method. The advantage of MLST over MLEE is that the results from MLST are comparable between laboratories. Compared to MLEE, MLST is more precise and convenient [72, 237, 329]. Compared to *fla*-RFLP, MLST has more discriminating power [75]. The combination of MLST and flagellin A short variable region (*fla*A SVR) can discriminate *C. jejuni* in outbreak investigations as well as PFGE [315]. Thus the introduction of MLST provided a useful tool for population genetic analysis [74, 237].

For evolutionary studies the sequence typing technique needs the following properties [237, 369]:

- suitable for evolutionary modelling
- a highly discriminatory method
- the results should be unambiguous
- the process can be repeatable across laboratories

MLST has the above properties, and it can be applied widely, and the results can be compared worldwide.

How MLST works

The aim of MLST is to provide an accurate and highly discriminating typing system that can be used for most bacteria and is particularly helpful for the typing of bacterial pathogens. This unambiguous procedure characterizes isolates of bacterial species using the DNA sequences of internal fragments of multiple (usually seven) housekeeping genes. This method uses approximately 450-500 bp internal gene fragments, which can be accurately sequenced on both strands using an automated DNA sequencer. At the gene level, each unique housekeeping gene sequence is assigned a distinct allele number within a bacterial species. At the isolate level, the alleles (usually seven) at each of the loci define the allelic profile or sequence type (ST) [74, 237].

Data from many MLST analyses have been stored and can be accessed from the public database PubMLST [72]. Seven loci are chosen for *Campylobacter* MLST studies: *aspA* (aspartase A), *glnA* (glutamine synthetase), *gltA* (citrate synthase), *glyA* (serine hydroxymethyl-transferase), *pgm* (phosphoglucomutase), *tkt* (transketolase), and *uncA* (ATP synthase a subunit). The reason for choosing seven housekeeping genes is because the information from these seven housekeeping genes provides discriminatory information equivalent to the 15 to 20 loci required by multi locus enzyme electrophoresis analyses [74, 237]. Because the positions of these seven housekeeping genes on the chromosome are far enough apart, it is unlikely that two of them will be changed in one recombination event (Figure 2.2) [74].

How MLST has been applied

MLST has been successfully applied to a wide range of bacteria [2, 88, 343, 358] and was first used to analyse *Campylobacter* in 2001 [74]. Schouls et al. [2003] commented that the typing of *Campylobacter* strains only works well when identifying an outbreak but may fail in source tracing and global epidemiology due to the enormous variation in strains and the carriage of multiple types in animals. But they mentioned several future research areas, such as analysis of the *Campylobacter* host source. Later studies [72, 243, 260, 335] have indicated MLST's ability to identify different sources of infections of human diseases.

In 2004, MLST was applied to *C. coli* [72], which allowed *C. jejuni* and *C. coli* to be compared. Compared to *C. jejuni*, *C. coli* does not have a large diversity, but this may be due to the limited sample size for *C. coli* [72]. It is crucial to understand the molecular evolution of *C. jejuni* and *C. coli*, since both of them are responsible for a large percentage of gastroenteritis worldwide.

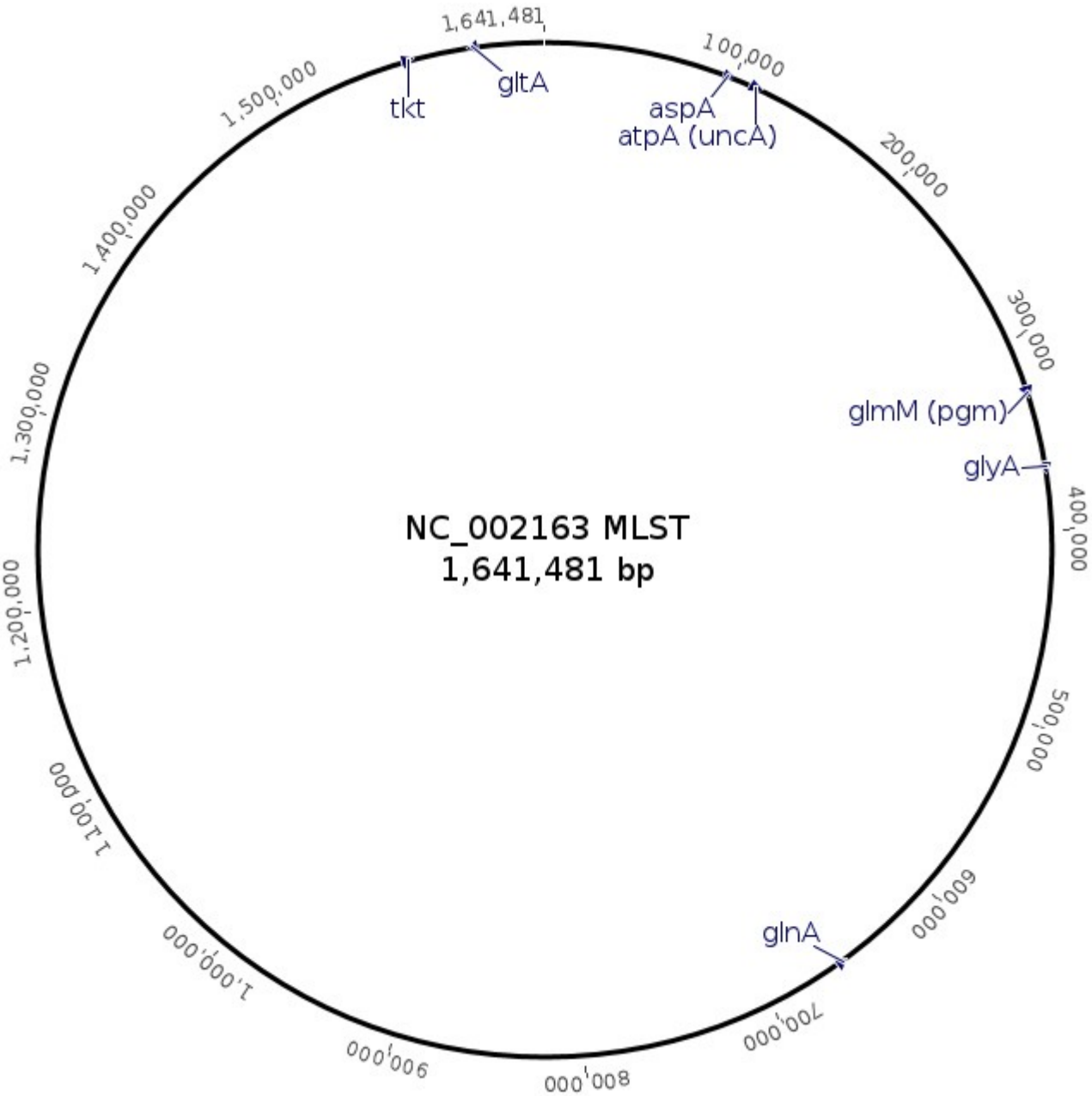


Figure 2.2: The positions of MLST loci on the chromosome of one *C. jej* strain NCTC 11168 (GenBank accession number NC002163) [286].

Because MLST can be repeated across laboratories, the results are comparable worldwide. MLST has been applied in different countries, such as Canada [49], Australia [66, 151], the UK [74], the US [237] and NZ [260]. Extended MLST (10-locus typing scheme) [73] represents a highly discriminatory typing scheme, which combines MLST, *fla*A-SVR typing, *fla*B-SVR, and *por*A typing systems. This extended typing sequence method can also be useful in both long-term epidemiology and outbreak analysis [73].

PubMLST is a publicly accessible dataset which stores MLST typing results for several bacteria species, including *C. jejuni* and *C. coli*. In the PubMLST dataset for *C. jejuni* and *C. coli*, the genes, such as *fla*A, *fla*B and *por*A, related to cell surface antigens have been integrated.

2.4 Evolutionary methods and phylogenetics

Phylogenetics is about inferring the evolutionary relationships among groups of species. In particular, molecular phylogenetics uses molecular sequences to address the evolution of organisms. The methodology developed in this area has been widely used to infer how species are related, and to understand the evolution of life [296, 410]. For bacterial research, phylogenetics can be used to find the origin of pathogens, transmission paths, and adaptions to a specific host. This thesis will employ a range of phylogenetic network tools to investigate the evolution of *Campylobacter*. The methods that I will discuss in this section have been selected as they are the most applicable to the large MLST dataset we have available for analysis. Further, some of the processes we are interested in such as recombination are not tree like, so phylogenetic networks methods will be applied.

2.4.1 Phylogenetic networks

Classification

Phylogenetic networks can be classified into four groups: phylogenetic trees, splits networks, reticulated networks (including hybridisation networks and recombination networks), and other types of phylogenetic networks [179]. The estimation methods of phylogenetic trees usually require the input file to be sequences or distances, but supertree methods take trees as input and summarise them into a single tree. Methods for constructing split networks can be divided into three subgroups depending on their input type: 1) median networks [15], 2) splits decomposition [14] or neighbour-net methods [15], and 3) consensus networks [166] or super networks.

For the input file, median networks require sequences as inputs, splits decomposition and neighbour-net methods [34] use distances, and consensus networks and super networks use trees. Reticulated networks use trees as input for hybridisation networks, and use sequences as input for recombination networks.

Phylogenetic networks can be also classified into two groups by whether the methods produce an explicit result for the evolutionary processes [179]. Implicit networks are used to reflect incompatible signals, and explicit networks are used to provide an explicit model of evolutionary history which includes hybridisation and recombination networks. According to this rule, phylogenetic trees and reticulated networks are classified as explicit methods; while splits networks and other types of phylogenetic networks are classified as implicit networks.

Both implicit and explicit networks are useful for understanding the evolutionary process. Explicit methods are probably more useful and interpretable in practice, but they are also harder to construct. Compared to explicit networks, implicit networks have currently been widely applied and have proved to be robust and helpful for visualising and exploring the conflicts and information contained in data.

When events like gene transfer or recombination occur, tree-based methods cannot explain complex evolutionary scenarios well, and networks methods seem more realistic [34]. In a reticulate network, there is an implied time direction from node to node. In a split networks graph, the external nodes represent taxa and sets of parallel edges represent splits, but unlike the reticulated networks graph, the internal nodes do not represent the hypothetical ancestors [179]. The following paragraphs will briefly introduce some well known methods in the first two groups: phylogenetic trees and splits networks.

Phylogenetic trees are the most widely used way to represent the evolutionary history of a set of taxa. A phylogenetic tree is defined as a tree (either unrooted or rooted) that represents the evolutionary history of a given group of taxa, with leaves labelled by a set of taxa and with or without branch lengths reflecting the evolutionary parameters, such as genetic distance or time [179]. Evolutionary trees can be reconstructed using many methods. There are two main groups of methods: 1) distance-based methods, including Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [344] and neighbour joining related methods [33, 127, 316]; 2) sequence-based methods, including Maximum Parsimony [82], Maximum Likelihood [82], and Bayesian inference [175].

However, the simplified evolutionary model (phylogenetic tree) cannot deal with complex but more realistic evolutionary scenarios, especially when reticulate events occur [179]. Reticulate events could include hybridisation and horizontal gene transfer. Hybridisation refers to the process in which different species of organisms com-

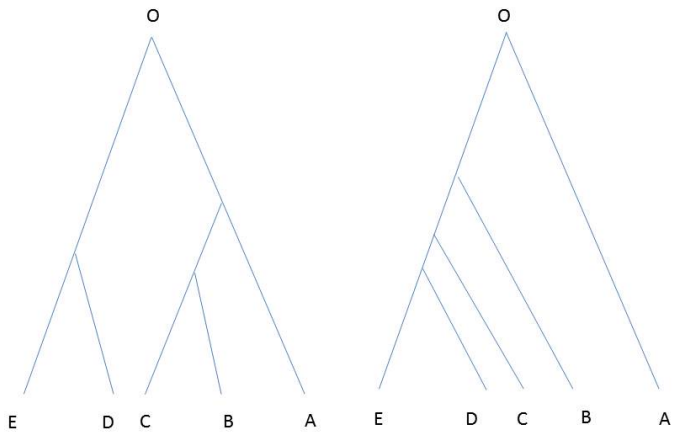


Figure 2.3: Two trees of the same set of taxa, but with different tree shapes.

bine to create another organism. Horizontal gene transfer (HGT) refers to events that transfer genetic material from one organism to another. *Campylobacter* evolve via a number of mechanisms including horizontal gene transfer. These reticulate events can be represented by phylogenetic networks. Moreover, even when the true history is a tree-like process, it could be difficult to reconstruct the tree-like history, therefore, the network methods in phylogenetics can be employed to display the conflicting signal in the statistical estimation methods [179]. The network methods also can add additional evolutionary events to the phylogenetic inference [180].

Based on a set of taxa X and a set of splits S, splits networks are a connected graph, and in this graph, taxa and splits are represented by some nodes and edges [180]. Splits networks are based on a given set of splits, and are designed to reflect the incompatibility in those given sets of taxa by sets of parallel edges. This design means splits networks can combine and represent the conflicting signals in the given splits. However, this design also means the true ancestral progress can not be clearly inferred. Therefore, in the biological area, the application of splits networks is limited to the early stage of research, and other techniques are required to infer the true history of the evolution of the given taxa. Figure 2.3 shows two trees of the same data set. Figures 2.4a and 2.4b show the differences when representing the same data in (a) a split network and (b) a reticulated network. A split network represents all the splits shown in Figure 2.3; a reticulated network uses reticulation events to demonstrate the differences between two trees in Figure 2.3.

Consensus networks [166] and neighbour-net methods [15] are both types of splits networks. Consensus networks are built from a set of input trees; all splits that

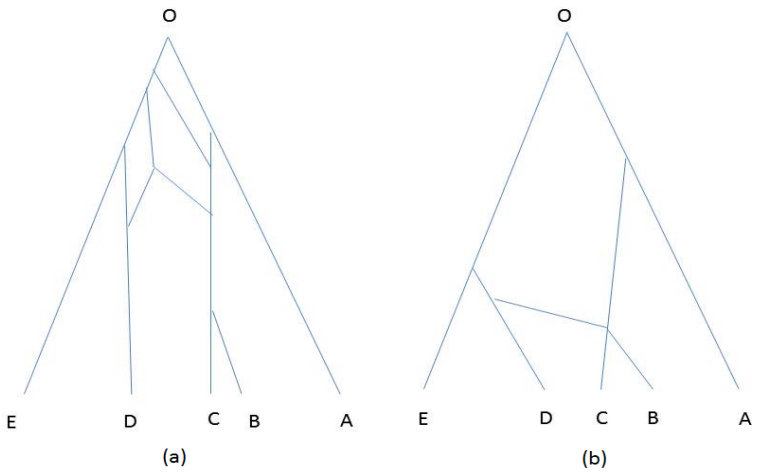


Figure 2.4: (a) represents a split network and (b) represents a reticulated network. Figure (a) represents all the splits that appeared in Figure 2.3; Figure (b) uses reticulation events to demonstrate the differences between two trees in Figure 2.3, and in Figure (b), there are three reticulation events.

occur in the consensus network should appear at a given proportion. Like other splits networks methods, consensus networks are also a way to show the conflicts among a given set of splits/trees, but not for conclusive phylogenetic inference. The neighbour-net method is fast and informative [34]. Neighbour-nets can represent the different signals due to both sampling error and recombination [34]. Neighbour-net efficiently yields a snapshot of the data, and produces a way to visualise the conflicting signals. Although neighbour-nets are not a final solution to the questions asked by biologists, they are a step further in building understanding the evolutionary history. Neighbour-nets highlight specific portions of the network for further formal statistical investigation. More detail on these is included in the following section.

2.4.2 Assessing confidence in phylogenetic trees

Like any other inference methods, there are potential estimating errors in the methods. Sampling error and systematic error are the two main types of errors. Sampling error occurs through random error due to the limited number of sites (sample size), while systematic error is the error reflecting the biases in the method's assumptions which could misinterpret data [179]. Sampling error can be dealt with by the non-parametric bootstrap or multiple samples obtained from the posterior distribution [179]. With the development of sequencing technology, the sampling error attributed to sample sizes has been improved [64, 179]. However, the development of sequen-

cing cannot improve the systematic error. In fact, increases in sequence length make the systematic error larger [64].

In order to test the robustness of the given parts of the trees, bootstrapping was introduced into phylogenetics [111]. A high percentage (at least 70%, suggested by Felsenstein [110]) of bootstrap support is required for statistically reliability and inference. The splits with low bootstrap support are recognised as very sensitive to the exact sequences in the input file [110].

2.4.3 Specific phylogenetic methods

Unweighted Pair Group Method with Arithmetic Mean (UPGMA) UPGMA [34, 344] is an agglomerative method used to produce phylogenetic trees. UPGMA yields a rooted phylogenetic tree based on the distance matrix of a set of taxa. Each node on the tree is assigned a height and the edge length represents the difference in heights of two connected nodes; edge lengths can be thought of as proportional to time or to genetic distance. UPGMA reduces the distance matrix by joining the nearest taxa together, and is based on the molecular-clock hypothesis. In molecular evolution, a molecular clock measures the genetic diversity accumulated through time for different species, and assumes that the divergence of two nucleotide sequences accumulates at a constant rate. UPGMA works well only when the situation is consistent with clock-like distance:

$$d(x, y) \leq d(x, z) = d(y, z) \tag{2.1}$$

where d represents the distance. The formula, known as the three-point condition, implies that the two longest routes are equal, and larger than or equal to the third [180].

Neighbour Joining Saitou and Nei et al. [34, 316] proposed the neighbour-joining method to construct phylogenetic trees. The neighbour-joining method is more widely used than UPGMA, becauseunlike UPGMA, the neighbour-joining method does not rely on the molecular clock hypothesis. Neighbours are defined as two taxa connected by a single node, usually, in an unrooted tree. At each step, the neighbour-joining method [128] joins the two closest sub-trees that are not already joined. The result of neighbour-joining is a single unrooted tree. Rooted trees can also be constructed by adding outgroup edges. This method is quite efficient and also suitable for large datasets [362].

Neighbour Net A neighbour-net is a variant of the neighbour-joining method, and it is an agglomerative method for building phylogenetic networks [34].

In 2004, Neighbour-net was put forward by Bryant & Moulton [34], and it uses a distance matrix as input. The differences between Neighbour-net and other distance-based phylogenetic networks methods are: neighbour-net does not immediately join the closest neighbours, instead, it waits until two closest nodes share a common neighbour, then joins the three together, and the matrix decreases by one element, and so on, until all of the nodes are joined together in a circular ordering. At the final stage, neighbour-net uses non negative least squares to find weights for the set of splits implied by the circular ordering. It can be mathematically proved [34] that compatible sets of splits are circular collections of splits. Splits can be calculated by the circular order [34], and then converted to a graphical representation in SplitsTree [177]. The advantages of neighbour-net are that this method is quick, the output is not too complex (planar), it can be applied to distances, and easily extended to sequences.

2.4.4 Sequence based methods

Maximum Parsimony

In Maximum Parsimony (MP), the phylogenetic tree requires the smallest number of evolutionary events to explain some observed set of aligned sequences [113]. The MP method is a widely used sequence-based non-parametric statistical method to reconstruct a tree. Based on MP, there are several trees which shared the same minimal number of changes (called equally parsimonious) [144]. In this case, no single tree among those can be inferred. A strict consensus tree or majority-rule consensus tree can be used to produce a consensus tree, which summarize the information from a set of equally parsimonious trees.

The MP method does not make explicit assumptions about the underlying evolutionary process. When the number of analysed sequences is small and the degree of divergence is not large, the MP methods can work well [144, 292, 348]. However, when some branches of the underlying tree are much longer than others, which means, some sequences evolve much faster than others, the MP method tend to group the long branches together to produce a wrong tree. This is called long branch attraction ([108, 112]), which is a systematic error [411, 418]. This phenomenon can occur to likelihood or Bayesian methods as well, when the underlying model is assumed to be too simple [411, 409].

In addition, with the increase of analysed sequences, the number of possible trees

increased dramatically, because the number of trees (unrooted) for n sequences is given by:

$$Nt = (2n-5)!/2^{n-3} \times (n-2)!$$

in which, Nt is the number of possible trees for a given set of n aligned sequences. Practically, we use heuristic searches to examine a subset of all possible trees. For a heuristic search, we start one initial tree, and compare the trees with similar topology to find a better tree, then use the better tree as an initial tree to start comparison, and so on.

Maximum likelihood

Maximum Likelihood (ML) method is a parametric statistical method, which uses all the information in the given set of data and requires explicit assumption on the evolutionary process, like the evolutionary model for nucleotide substitution. The likelihood of the parameter with data is the probability of the observed sequence pattern given the parameters. ML method contains two step optimization: branch lengths and tree (topology) model. The topology can be viewed as a model, and the branch lengths and parameters for various nucleotide substitution model are the parameters for a given model (tree).

The ML tree inference is a comparison of statistical models, and these statistical models contain the same number of parameters. Thus, the tree that has the highest probability of producing the observed data is the most likely tree. The properties of maximum likelihood estimates include consistent (when the number of data increases, the estimation approach to the true value) and efficiency (smallest variance among estimates). These properties hold, only when the chosen model is the true tree.

The main weakness of ML is the time-consuming computation, because ML needs consider all the possible alternative trees and the maximum likelihood value for all the possible trees [144]. Furthermore, if the underlying model is a wrong model or if the divergence among chosen sequences is large then ML may result in a wrong tree.

2.4.5 Bayesian methods

Bayesian statistics is an alternative method to frequentist/classical method. The main difference between the two methods can be summarised as follows:

• In the frequentist approach, data are repeatable, parameters are not. A frequentist will consider what values for the data are plausible conditional on a particular value of the parameters. This can be interpreted as $P(data|parameters)$.

• In the Bayesian approach, the parameters are uncertain, but the observed data are not. They will consider the probability distribution of the parameters conditional on the observed data $P(parameters|data)$.

Bayes' theorem relates to a formula which is used to calculate the conditional probabilities of events A and B: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$. The equivalent version for the density functions is applied more widely in Bayesian statistics: $f(\theta|x) = \frac{f(\theta)f(x|\theta)}{f(x)}$. This equation is used to derive the density function for the unknown parameter θ, conditional on the given data x. In this equation

• $f(x|\theta)$ is the likelihood of the data conditional on the parameters.

• $f(\theta)$ is the prior distribution of θ. It states what the decision-maker knows about the parameter \theta without taking into account any of the data. Prior knowledge can be made available from either experts or similar historical data.

• $f(\theta|x)$ is the posterior distribution of θ, given x . It expresses uncertainty about θ after taking the data.

• $f(x)$ is the normalising constant, which ensures that the posterior distribution integrates to 1.

Markov Chain Monte Carlo Methods Monte Carlo methods use Monte Carlo simulation to solve various statistical and mathematical problems. This term represents a large and widely-used collection of algorithmic approaches rather than a single Monte Carlo method. Given the sample $\theta_1, \theta_2, ..., \theta_N$, if N is sufficiently large, any posterior summary of θ can be obtained by Monte Carlo Integration. The main idea behind Monte Carlo Integration is to use the sample averages to approximate the posterior expectations. Furthermore, most of the summaries of the posterior distribution can be interpreted by the expectations:

$$E[f(\theta|x)] \approx \frac{1}{N}\sum_{i=1}^{N} f(\theta_i).$$

Markov Chain Monte Carlo (MCMC) can be applied to estimate properties of probability distributions that are very difficult to obtain analytically. This algorithm is based on the Markov Chain and uses the Monte Carlo way of thinking, so it is called the Markov Chain Monte Carlo. The MCMC methods construct a Markov chain with θ as the state variable and desired distribution $f(\theta|x)$ as the stationary distribution. Then we simulate the chain from the arbitrary starting values. If the

sample is sufficiently large, the stationary distribution will converge to the target distribution $f(\theta|x)$. Since the samples from posterior distribution can be always obtained, this method allows us to do inference, no matter how complex the prior distribution or likelihood is.

Bayesian phylogenetics

Bayesian phylogenetics uses Bayes' theorem, and generates a posterior probability of a tree, which contains the information from the prior distribution of the tree and the data (likelihood). The outcome of Bayesian analysis is a credible sample of trees. The posterior probability of a tree can be interpret as a tree is correct, given the sequence data that we observed. It can be state as [411]:

$$P(T, \theta|x) = \frac{P(T, \theta)P(x|T, \theta)}{P(x)}$$

in which, $P(T, \theta)$ represents the prior probability of the tree T, and parameter θ, $P(x|T, \theta)$ represents the probability of the observed data x, given the tree T, and parameter θ; also as the likelihood of the tree T, and parameter θ. $P(x)$ is a normalizing constant. The choice of prior probability is difficult to make and the impact on the posterior probability is unexpected. The posterior tree probability is also sensitive to model violations [411].

2.4.6 Consensus trees and consensus split networks

A consensus tree provides a summary of different trees [400]. Different trees can be obtained for different reasons, such as a multi-gene study, different inference methods or a Bayesian method which produces multiple trees [180]. There are different methods for producing a consensus tree, for example:

1. The semi strict consensus tree: this method only show the clades which are not contradicted by other trees.

2. Majority rule: the threshold can be varied by customising, but is usually larger than 50%, which means the clades present on more than half of the whole tree sources will be the clades on the consensus tree. A strict consensus tree is an example of majority rule where the threshold equals 100%, which means only the clades which appear in all trees will be present on the consensus tree [402].

A consensus tree represents the agreed part of a set of phylogenetic trees, and it represents the mostly likely evolutionary history of the given taxa. The remaining

part can be demonstrated by the consensus split networks, and this incompatible signal can be further investigated. For multi-gene datasets, consensus networks can combine trees as the input datasets. The input of consensus split networks is a set of trees, based on the same set of taxa and a threshold to filter the groups of trees. The output from consensus networks is a splits-graph [165]. There are software which can be used to produce consensus split networks, including versions of SplitsTree [177, 178]. The most frequently used of SplitsTree is SplitsTree4 [178].

2.5 Population genetics

Population genetics studies the distribution and the change of allele frequencies [155, 372]. Allele frequencies can be influenced by several main factors: natural selection, genetic drift, mutation, gene flow, recombination, and population structure. The coalescent model is an extension of traditional population genetics, and has some advantage over phylogenetics for investigating genetic polymorphisms. The following section will briefly introduce coalescent theory.

2.5.1 Coalescent theory

Coalescent model

With the development of genotyping technologies, genetic polymorphism data can be used to infer population phenomena, like migration; however traditional population genetic methods, like deterministic models based on the Hardy-Weinberg principle, cannot handle the property of genetic polymorphism data as they come from a unique, complex, non-repeatable evolutionary history [310]. Since the 1980s, the coalescent model has been proposed to cope with polymorphism data [170, 171, 202, 203, 360]. Coalescent theory is an extension of population genetics and it has a close relationship with classical population genetics.

In classical population genetics, one begins by simulating the entire population, then waiting until it reaches equilibrium to take a sample. Unlike traditional deterministic methods in population genetics, coalescent models go back in time to find the parents for the current samples, until the most recent common ancestor (MRCA) is found, and then add the mutations or recombination events on to the genealogy. The focus of coalescent models is the inference of the past evolutionary process on the current genetic diversity, rather than the effects of different initial conditions on the evolutionary process.

Coalescent methods use polymorphism data obtained from natural populations to infer evolutionary processes, rather than estimating them directly from laboratory experiments. This difference leads to difficulty in data analysis, because the real evolutionary process in history is like the result from a single experiment, and this experiment has no replication results and an obscure starting condition [310]. Coalescent models are more suitable for dealing with these difficulties.

In order to model polymorphism data, the randomness of genealogy and mutation in the evolutionary process must be considered, because the pattern of the polymorphism (the nucleotide position where the variant occurs and the frequency of this variant) is affected by the mutation history and the genealogy (the branches where the variants are). Mutations create new genetic material [71], the coalescent process gives rise to the genealogical tree [310], and the recombination breaks the linkage between loci [275], and increases the diversity of the sequence types [71].

In Rosenberg and Nordborg's review [2002], they gave an example to illustrate the dependence in the samples of haplotype data. This paragraph will restate and explain their example. Haplotypes are the allelic combinations of multiple loci on one given individual's chromosome [310]. In other words, a haplotype is a set of closely related (not easily separated) genetic loci [282]. This set of genetic markers tends to be inherited together and work together to decide one type of trait. For each haplotype, from the horizontal view (the different combination of allelic states from different loci), the genetic linkage brings the dependence to those loci; from the vertical view (different allelic states for one locus), the common ancestor causes the dependence in that locus. These two dependencies rely heavily on the randomness inherent in the evolutionary process. The coalescent model was introduced to deal with this dependence as the results of an irreversible and unique history, whereas phylogenetic methods do not consider these uncertainties. In short, coalescence is a stochastic process, as going back in time, the haplotype randomly picks its ancestral linkage; when two sequences pick the same "parent" (ancestral linkage), one coalescence occurs.

Factors that affect coalescent modelling There are many factors can affect this process [310]: some factors can affect the coalescent rate such as reproductive ability, age and sex structure; others can change the genealogy structure, such as recombination. The reason is that recombination can introduce the net-like (non tree-like) shape into the genealogy, which creates different genealogical trees for different linked genes.

Population structure and selection pressure [173, 192, 270] also play a role in shaping the genealogy. These factors are not difficult to incorporate into the coalescent

model. The speed of coalescent processes depends on the number of parents (the ancestors) and children (the lineages). The more children and fewer parents there are, the faster the process is.

Application of coalescent models

The coalescent model can be applied to study molecular ecology, phylogeography and the divergence time of species. The coalescent model also can be applied to investigate the evolutionary process and epidemic transmission [275, 310].

When applying the coalescent model to the analysis of the relationships among or within different groups of species, it can cope with the variation within and between species, as well as, the different gene histories of the genome. The coalescent model can also be applied to investigate how rapidly bacteria are evolving and whether certain sequence types are under selection pressure. Different parts of the genome could be under different selection pressures. Therefore, different genes across the genome may have a different genealogy, and the coalescent model can easily incorporate this kind of analysis.

From a statistical point of view, the coalescent model can be used to construct test statistics based on the analysis of experimental data. One of the widely used tests is Tajima's D statistic [361], which compares the average observed number of pairwise differences among DNA sequences in a sample to the expected number of mutations in the coalescent model. This statistic provides an opportunity to explore any deviation from the null model.

Coalescent methods can also be used to simulate data under some scenarios/assumptions, which can be used to compare with the observed data. Compared to the classic population-genetic simulation, the coalescent method is more efficient and convenient. In exploratory data analysis, it is now accepted as a good way to test some hypotheses [172, 212].

Allowing full likelihood analysis of evolutionary and demographic models is one of the exciting but under-developed aspects of the coalescent model [310, 351]. Some efficient numerical computational tools for calculation of likelihoods have been developed to cope with a simplified model, because integrating all the possible genealogies in a complicated (maybe more realistic) model is still difficult. Two well known techniques are Markov Chain Monte Carlo (MCMC) [61, 136, 214] and importance sampling [139, 352]. However, computational requirements still limit full likelihood analysis on the whole genome scale. Alternatively, some methods based on summary statistics were introduced to simplify the calculation, but still maintain the crucial

information [275, 310], though in practice, a thoughtful choice of summary statistics is required for more accurate analysis [310].

2.5.2 The comparison between phylogenetic model and coalescent methods

The difference from the phylogenetic model

Phylogeny aims to estimate how species are related; population genetics focus on relationships within species. As described in Rosenberg and Nordborg's paper [310], phylogenetic methods try to estimate a tree, whereas coalescent models are not limited by the tree assumption, and can take into account crucial events in history, such as recombination, migration, and selection. There are some limitations in the phylogenetic methods: firstly, one single estimated gene tree cannot reflect the possible species tree, because different genes across the genome can produce different evolutionary trees. Therefore, only considering one estimated gene tree will ignore other possibilities. If recombination is prevalent, then even different parts of a gene might have different histories. Secondly, it makes more sense to consider the likelihood of estimated tree under different models, because genealogical trees are generated by a random process. Thirdly, the tree assumption might be invalidated by large effects caused by migration.

The difference in the mathematical formulae for the likelihood of obtaining the data:

There are two equations commonly used to infer the likelihood of obtaining analysed data from given parameters. Rosenberg and Nordborg [310] discuss two equations.

$$L = P(D|G, \mu), \tag{2.2}$$

where L is the likelihood (the probability) of obtaining the observed data, based on the given parameters, D is the observed data, G represents the genealogical tree and μ is a set of parameters determined by the simple evolutionary (mainly mutation related) models [310, 112].

In comparison, the equation for likelihood of inferences in the coalescent setting is

$$L = \sum_{G} P(D|G, \mu)P(G, \alpha), \tag{2.3}$$

where α is a set of parameters that we are interested in for the population information, such as population structure, population sizes, and the rates of migration or mutation [310].

Equation 2.2 treats the the genealogical tree as a parameter G, and tries to estimate it. In contrast, the aim of equation 2.3 is to estimate parameter α. The tree or genealogy, G, is a nuisance parameter, which is not the primary interest and will not be estimated by equation 2.3, because it is removed by taking the average value for all possible trees.

The difference in procedures

For phylogenetics [112], the procedure is 1) collect samples of sequence data; 2) estimate a tree based on the given sample, regardless of the non-tree like situation; 3) make inferences from the species tree, based on the estimated gene tree.

For coalescent models [112, 310], the procedure is 1) collect samples of sequence data; 2) consider all the possible genealogies under different evolutionary models, taking into account the recombination event; 3) calculate the total likelihood under each model, which equals the weighted sum of the likelihoods of all genealogies produced by that model; 4) estimate the parameters of interest that maximise equation 2.3; 5) repeat step 3 and step 4 to compare the likelihood of data from different evolutionary assumptions.

As described in Section 2.4, phylogenetics can produce an exploratory data analysis for further investigation. Phylogenetics is a reasonable choice if the interest is the evolution of specific locus, rather than the parameters in the evolutionary model. Coalescent models allow standard statistical comparisons among different evolutionary scenarios, and can deal with complex evolutionary events, like admixture and recombination.

Conclusion of differences between phylogenetic model and coalescent methods

Coalescent models overcome the phylogenetic limitations in analysis of polymorphism data and are more consistent with statistical frameworks. The coalescent models cover more possibilities than phylogenetic models, because more genealogical trees are considered in the analysis. Furthermore, it is much easier to handle recombination in coalescent models than models borrowed from phylogenetics. The coalescent models can take into account the random process for generating genealogy tree-like and non-tree-like situations.

Compared to traditional population genetic methods, coalescent models provide more detail about the evolution of *Campylobacter*. Therefore, coalescent models will be applied in this project, in addition, to the phylogenetic and traditional population genetic methods. The results can provide useful information for the future application of coalescent models.

2.6 Software

This section provides a list and brief description of the software used in the investigation.

PAUP* [359]: (Phylogenetic Analysis Using Parsimony): a phylogenetic analysis package. It can perform maximum likelihood, parsimony, and variety of distance methods.

ModelTest [300]: software that can select and rank among 56 models of DNA substitution that best fit the data. ModelTest is based on three criterion hierarchical likelihood ratio tests (hLRTs), Akaike information criterion (AIC) and Bayesian information criterion (BIC). ModelTest produces likelihood values through running PAUP*.

SplitsTree4 [177, 178]: this software is used for visualising and exploring a set of taxa. It can work on sequences, distances and trees. This software can implement a range of phylogenetic network methods, such as Neighbour Net [34], UPGMA [34, 344], Neighbour Joining (NJ) [34, 316] and its variant, BioNJ.

Arlequin [93]: a software package for population genetics analysis. It can be applied to calculate Fst, Analysis of Molecular Variance (AMOVA) and Tajima's D. It can calculate a variety of distance metrics: such as, Jukes-Cantor, the Kimura 2-parameter distance, and the Tamura-Nei distance. It also can perform coalescent-based methods, such as Tajima's D test, as well as several population genetic analyses, such as genetic diversity.

DanSP [228]: a program for the polymorphism data analysis. The required input is aligned DNA sequence data. It can perform coalescent based methods, such as Tajima's D test, as well as several population genetic analysis, such as linkage disequilibrium, recombination, and gene flow. It also can measure the variation of DNA sequences within and between populations by several statistics.

RDP3 [240]: an integrated program designed for detecting and identifying recombination events.

BEAST (Bayesian Evolutionary Analysis Sampling Trees) [80]: a program for Bayesian inference for parameters in the coalescent theory. It can perform a variety of sub-

stitution models for nucleotide and codon data under different population models. In phylogenetics, it is mainly used to estimate trees under Bayesian theory and estimate time of divergence for different bacteria or species.

ClonalFrame [69]: a suitable tool to estimate the clonal relationships, which are usually represented by a genealogy. ClonalFrame is a model-based method, which can be used to infer the clonal pattern of bacteria and the recombination location on the chromosome. This method is suitable for analysing MLST data alignments of multiple bacterial genomes.

SimMLST [70]: a package for simulating MLST type data based on a neutral model. SimMLST can generate MLST type data under the given recombination and mutation rate model and the population growth model also can be specified.

Structure [303]: software for Bayesian model-based cluster, and it can be applied to infer population structure or assign individuals to populations.

BAPS (Bayesian Analysis of Population Structure) [58]: software for Bayesian model-based clustering.

Chapter 3

Estimating the relative roles of recombination and point mutation in the generation of single locus variants in *Campylobacter jejuni* and *Campylobacter coli*

This chapter comprises a paper published in 2012 from this research. However some additional background information is provided as context on single locus variants (SLVs) for the paper.

3.1 Background

The definition of SLVs

Single locus variants are defined as a pair of strain types that differ at exactly one of the seven alleles that make up the MLST profile [103]. The extensive use of MLST datasets in many laboratories offers a good opportunity to estimate what proportion of SLVs arise by recombination. Many SLVs correspond to a single event e.g., a mutation, recombination, or at least a small number of events. SLVs are closely related strains on the evolutionary path, and can be assumed to share a recent common ancestor. The clonal complexes are typified by a group of isolates sharing identical alleles at six loci, plus minor clonal variants which differ from this group at only one out of the seven loci [101].

Table 3.1: Example one for an SLV

ST	*asp*A	*gln*A	*glt*A	*gly*A	*pgm*	*tkt*	*unc*A	clonal complex
19	2	1	5	3	2	1	5	ST-21 complex
21	2	1	1	3	2	1	5	ST-21 complex

Table 3.2: Example two for an SLV

ST	*asp*A	*gln*A	*glt*A	*gly*A	*pgm*	*tkt*	*unc*A	clonal complex
1326	104	7	10	4	1	7	1	ST-45 complex
3071	184	7	10	4	1	7	1	ST-45 complex

The classification of SLV

As SLVs are pairs of STs that most likely share a very recent common ancestor, analysis of SLVs can be helpful in understanding the evolution and molecular epidemiology of pathogens. In this chapter, mutation is defined as a single nucleotide change (a point mutation), whereas recombination represents the transfer of several adjacent nucleotides from one DNA source to another.

Table 3.1 shows that strains ST-19 and ST-21 only differ at one allele. The alleles *gltA*-5 and *gltA*-1 only differ by one nucleotide. This is most likely the results of a mutation. Table 3.2 demonstrates that strains ST-1326 and ST-3071 only differ at one allele. The alleles *aspA*-104 and *aspA*-184 differ by 60 nucleotides, so the SLV most likely arose due to a recombination event.

Previous research [104, 101, 105, 100] considered SLVs occurred with in clonal complexes, but this research has shown this assumption is not accurate. For *Campylobacter*, SLVs mostly occur within clonal complexes, but sometimes they cross clonal complexes, e.g., clonal complex-48 (CC-48) and CC-206.

The rationale of the SLV analysis

1. The aim is to study closely related strains, therefore, the analyses are based on single locus variants, rather than variants for more than one locus, e.g., double locus variants. The reason for considering SLVs is because they are close on the evolutionary path, and can be assumed to have a most recent common ancestor (MRCA). MLST is a highly discriminating technique for *Campylobacter* and isolates that have identical alleles (sequences) at all seven loci, and the variants that differ at only a single locus, are almost certainly closely related and are descended from a recent common ancestor [101].

2. There are a large number of SLVs available for analyses. The difference in one locus can contain several nucleotide differences. At present, analysis of these

variants at the gene level is sufficiently informative to locate the STs which are close to each other in the evolutionary path.

Using eBURST

BURST is an algorithm for MLST data clustering. BURST uses a simple but appropriate model of bacterial evolution to display the relationships between closely-related isolates of a bacterial species or population. The basic idea of the model is that an ancestral genotype diversifies to a cluster of closely-related genotypes. In other words, all descended genotypes come from the founding genotype. A "clonal complex" is defined as a cluster of genotypes, and these genotypes are assumed to have a common ancestor [100, 103]. The eBURST method was described in a paper by Feil et al. [2004]. eBURST is the advanced version of BURST. eBURST is a new updated version of BURST and eBURST was integrated with the MLST databases from the websites. The main difference between BURST and eBURST is the way of displaying the relationships between closely related STs. The eBURST algorithm is implemented as a Java applet. eBURST produces one possible way that each clonal complex may have emerged and diversified, and the information on phenotypic, genotypic, or epidemiological data should be taken into account to check the proposed ancestry and patterns of descent.

The principle of this method is that eBURST allocates the related STs into the same group, and then predicts the ancestral (founding) genotype of each group, which means, eBUSRT identifies the most parsimonious patterns of descent of related sequence types from the predicted founders. Bootstrap resampling procedures are used to calculate the confidence intervals of a ST being the group founder.

There is a great distinction between an eBURST group and a real clonal complex. An eBURST group refers to a collection of STs selected by a group definition in the algorithm, whereas a clonal complex refers to a biologically meaningful cluster of STs. The group definition can be set by different criteria. For example, the default setting of belonging to one group is that all strains share six out of seven loci. But this criterion can be relaxed to all strains share five loci belong into one group. No matter which criterion is being used, no ST is assigned into two groups. At the centre of each clonal complex is a common founder. eBURST tries to estimate the most likely founder within each clonal complex, based on the given group definition and the graphic results could be affected by the different group definitions. These results are unlikely to represent the original real genotype of the entire group, due to the large number of possible genealogies. For example, in an eBURST diagram (Figure 3.1), all STs are linked as SLVs to at least one other ST, but these could be individually separated from the real clonal complex [103].

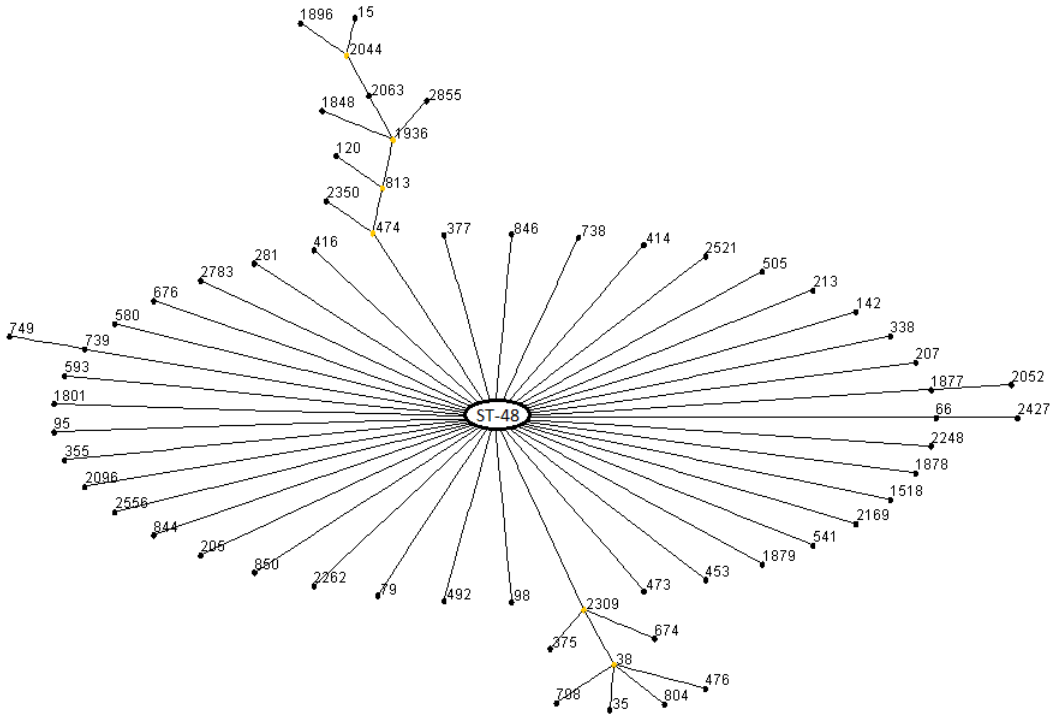


Figure 3.1: Nodes represents STs, numbers beside the nodes in the diagram represent ST numbers (the middle one is ST-48), lines connects all the SLVs, there is no indication of distance, the centre of star is the estimated common ancestor for this eBURST group.

Estimating the relative roles of recombination and point mutation in the generation of single locus variants in *Campylobacter jejuni* and *Campylobacter coli*

Authors: Shoukai Yu, Paul Fearnhead, Barbara R. Holland, Patrick Biggs, Martin Maiden, Nigel French,

Journal of Molecular Evolution, June 2012, Volume 74, Issue 5-6, pp 273-280

Abstract

Single locus variants (SLVs) are bacterial sequence types that differ at only one of the seven canonical MLST loci. Estimating the relative roles of recombination and point mutation in the generation of new alleles that lead to SLVs is helpful in understanding how organisms evolve. The relative rates of recombination and mutation for *Campylobacter jejuni* and *Campylobacter coli* were estimated at seven different house¬keeping loci from publically available multilocus sequence typing (MLST) data. The probability of recombination generating a new allele that leads to an SLV is estimated to be roughly seven times more than that of mutation for *Campylobacter jejuni*, but for *Campylobacter coli* recombination and mutation were estimated to have a similar contribution to the generation of SLVs. The majority of nucleotide differences (98% for *Campylobacter jejuni* and 85% for *Campylobacter coli*) between strains that make up an SLV are attributable to recombination. These estimates are much larger than estimates of the relative rate of recombination to mutation calculated from more distantly related isolates using MLST data. One explanation for this is that purifying selection plays an important role in the evolution of *Campylobacter*. A simulation study was performed to test the performance

of our method under a range of biologically realistic parameters. We found that our method performed well when the recombination tract length was longer than 3kb. For situations in which recombination may occur with shorter tract lengths, our estimates are likely to be an underestimate of the ratio of recombination to mutation, and of the importance of recombination for creating diversity in closely related isolates. A parametric bootstrap method was applied to calculate the uncertainty of these estimates.

Introduction

The genus *Campylobacter* is the major cause of gastroenteritis in many industrialized countries [363], with approximately 1% of the population throughout the western world being affected by campylobacteriosis every year (The World Health Organization, cited in [176]). The species *Campylobacter jejuni* and *Campylobacter coli* are the main causes of bacterial food-borne disease in developed countries, compared to other members of the family Campylobacteriaceae [208]. Substantial evidence for the presence of recombination at specific genes has been found in several studies [99, 357]. The relative contributions of recombination and point mutation to ge-netic diversity have also been investigated [101, 102, 104, 320]. Although most research indicates that recombination contributes more to genetic diversity than mutation, there is considerable uncertainty about the relative number of events and the number of nucleotide differences that may be attributable to these two processes [99, 306, 325]. This paper is focused on estimating the relative contributions of recombination and point mutation to the generation of new alleles that lead to single locus variants (SLVs), based on *C. jejuni* and *C. coli* from the seven gene multilocus sequence typing (MLST) scheme. An SLV is a pair of sequence types (STs) that differ at exactly one of the seven alleles that make up the MLST profile [103]. SLVs are pairs of STs that most likely share a very recent common ancestor and the analysis of SLVs can be helpful in understanding the evolution and molecular epidemiology of pathogens. The large collections of isolates that have been characterized by MLST provide a good opportunity to study SLVs in detail. This research is based on distinct STs of *C. jejuni* and *C. coli* in the PubMLST database (http://pubmlst.org/campylobacter). In order to understand whether there are differences in the mechanisms that produce SLVs across the genome, SLVs were divided into groups depending on the locus at which the STs differ. The distribution of nucleotide differences within SLVs was explored. The nucleotide differences between two STs that form an SLV can be generated by two different kinds of events: recombination or mutation. Intuitively, SLVs that comprise two STs which differ at many

nucleotide positions are more likely to be due to recombination, whereas those that differ at only a few nucleotide positions may be the result of point mutations. In this study, an EM algorithm was applied to allocate SLVs into either a point mutation only model or a recombination model. Two key parameters were estimated: the probability that an SLV arose due to point mutation(s) only, and the relative rate of recombination to mutation. In order to test the performance of our method, a simulation study was performed under a range of biologically realistic parameters. When the recombination tract length was longer than 3kb, our method performed well. 3kb is the average tract length suggested by previous research on Campylobacter [25, 99, 325, 403]. When recombination occurs with shorter tract lengths, our estimates may underestimate the ratio of recombination to mutation.

Material and methods

Campylobacter Data

The data were taken from the PubMLST database (September 27, 2010), at this time the PubMLST database contained 4676 distinct *C. jejuni* and *C. coli* STs. MLST is a way of typing strains that is based on nucleotide sequences (Maiden et al., 1998). Using the MLST technique [74], these isolates are sequenced at seven housekeeping loci (*asp*A, *gln*A, *glt*A, *gly*A, *pgm*, *tkt*, and *unc*A). These seven loci are widely dispersed around the genome, which means there is a very low chance for one recombination to change two or more loci.

We separated ST datasets for *C. jejuni* and *C. coli*, and excluded the 22 STs found in both species. Furthermore, we separated *C. coli* by clades according to previous research [335, 336], and we chose *C. coli* clade 1 to investigate in detail because *C. coli* clade 1 contains more STs, and is more diverse, compared to the other two clades [335]. We selected clade 1 from *C. coli* by extracting all STs that are members of ST-828 clonal complex and ST-1150 clonal complex [334]. There are 3654 STs for *C. jejuni*, and 606 distinct STs for *C. coli* clade1.

Methods overview

Either mutation(s) or recombination(s) can generate SLVs. In this paper, mutation is defined as a single nucleotide change (a point mutation), whereas recombination represents the transfer of several adjacent nucleotides from one DNA source to another. An event is either a mutation or a recombination. An SLV can be generated by one or more events, however recombination will tend to mask mutation. We

model separately the mutation and recombination process in order to derive a probability model for the number of nucleotide differences between STs, under both the assumption that the SLV has been created solely by mutation, and that it has not. This then enables us to estimate the proportion of SLVs that have been caused solely by mutation, and also estimate the relative rate of recombination to mutation. More details of the analysis are given in the Supplementary Material.

Modeling SLV evolution

The data consists of, for each SLV, the locus at which the pair of STs differ, and the number of nucleotide differences at that locus. From this we aim to infer how likely it is that the differences observed at this locus arise from point mutation only, as opposed to being produced by recombination.

To do this we first model the distribution of nucleotide differences we would expect at an SLV at a given locus if these differences are solely due to mutation. This can be done by first calculating the probability of an SLV given the number of point of mutations that have occurred in one locus as the likelihood function, introducing a prior distribution for the number of mutations to occur between two STs in that locus. The former probability is based on the need for all mutation events to occur at the same locus. Under the coalescent theory, a geometric distribution is chosen to use as the prior distribution [158]. Under Bayesian theory, we can obtain the required conditional distribution (Equation 3.2 in Supplementary Material). The resulting conditional distribution of the number of nucleotide differences is concentrated on small numbers of nucleotide differences, and is robust to the choice of prior.

Secondly, the probability of observing h (h =1, 2, 3...) nucleotide differences introduced by recombination was estimated using Bayesian methods. It was calculated by sampling the alleles based on their frequencies in the current database. Two (simplifying) assumptions for the recombination model were made: (1) if recombination occurs between two alleles it affects an en¬tire locus rather than just part of a locus; and (2) we ignore the effect of any additional mutation events. Under these assumptions, our model suggests that in most cases recombination will introduce many more nucleotide differences than expected under the mutation only model. Note that our results are robust to the assumption in (1) unless recombination affects only small fragments of a locus, in these cases our assumption will tend to lead to overestimates of the proportion of SLVs due to mutation only. Hence, it will tend to underestimate the ratio of recombination to mutation.

Given these two models, we can then estimate the proportion of our SLVs at each locus that are due to mutation only. In practice we use an Expectation-maximization

(EM) algorithm [65] to infer this proportion. Lastly, based on the estimated proportion of SLVs at a given locus that is due to mutation only we estimate the probability that the single event that led to the generation of a new allele was a mutation. The above analysis was carried out by an R script (available by request).

To test the accuracy of our method for estimating the ratio of recombination to mutation, MLST data were simulated under different known ratios of recombination to mutation with different recombination tract lengths using SimMLST software [70].

We used a parametric bootstrap to assess uncertainty in estimates. We simulated 100 datasets for both *C. coli* and *C. jejuni*. These datasets matched the true data in terms of number of STs, relative rate of mutation to recombination, and overall mutation rate across the 7 gene loci. Within the simulations we assumed the mutation rate and recombination rate were the same across loci. For our simulated data we estimated the probability of an event being a mutation, and calculated the variability of estimates of this quantity across the simulations: both for estimates for a single locus, and for the estimate obtained by averaging across loci. We consider estimates of this quantity as the variance of the estimates changed little when we varied the true value of the relative rate of recombination to mutation. Confidence intervals where then calculated using a normal approximation, and transformed to confidence intervals for the relative rate of recombination to mutation.

Results and Discussion

SLV analysis on the Campylobacter MLST databases

From our downloaded dataset, there were 7417 SLVs (*asp*A: 992; *gln*A: 1045; *glt*A: 1250; *gly*A: 773; *pgm*: 1580; *tkt*: 1060; and *unc*A: 717) for *C. jejuni*, and 1842 SLVs (*asp*A: 110; *gln*A: 179; *glt*A: 128; *gly*A: 292; *pgm*: 325; *tkt*: 647; and *unc*A:161) for *C. coli* clade1. The difference in the number of SLVs at each locus suggests it is worthwhile estimating the relative mutation and recombination rates separately for each locus.

The distribution of nucleotide differences between each SLV for each locus

Each SLV relates to one pair of STs, and the plots (Figure 3.2 and Figure 3.3) show the nucleotide differences that occurred within those pairs of STs at each MLST locus for *C. jejuni* and *C. coli* clade 1. These plots show that SLVs with

a large number of nucleotide differences (> 45) occurred in every locus. The pairs of STs with a large number of nucleotide differences (50 to 80) are almost certainly due to recombination, as it is highly unlikely that more than 50 independent point mutations would occur at a single locus while the other six loci remained the same. These large differences are likely to be due to recombination between *C. jejuni* and *C. coli* [335, 403]. Species were designated according to the PubMLST data, and only those SLVs that comprised STs that were assigned 100% *C. jejuni* or *C. coli* were plotted. Even with this strict species designation, there were still large nucleotide differences visible between SLVs within species. There were second peaks in the range of 15 or 20 differences at the loci *gly*A, *pgm* and *tkt* for both *C. jejuni* and *C. coli* clade 1. These peaks are likely to be due to recombination as well. The first peak of most loci (except for *pgm* for *C. jejuni* and *tkt* for *C. coli* clade 1) represented approximately 100 to 200 SLVs for *C. jejuni* and around 100 SLVs for *C. coli* clade 1 with only one nucleotide difference; most of these are more likely to be due to mutation.

Relative contributions of recombination and mutation separately for *C.jejuni* and *C. coli* clade 1.

Table 3.3, and Table 3.4 demonstrate that recombination contributed more to the generation of SLVs than did mu¬tation for both of the groups (*C.jejuni*, and *C. coli* clade 1), but the range of estimates vary for the two groups. The average ratio of recombination events to mutation events from the seven loci is 6.96 (95% CI 6.08, 8.09) for *C. jejuni* (Table 3.3), and 1.01 (95% CI 0.78, 1.30) for *C. coli* clade 1 (Table 3.4).

For each locus we also estimated the proportion of nucleotide differences introduced by recombination as opposed to mutation, and this ranged from 97% (*glt*A and *gly*A) to 99% (*asp*A, *tkt* and *unc*A) for *C. jejuni*, and from 60% (*gln*A) to 98% (*asp*A) for *C. coli* clade 1.

We also investigated the robustness of the mutation model to different prior distributions of the probability of events caused by mutations only. These suggest that the results in Supplementary Table 1 and Supplementary Table 2 are conservative regarding the importance of recombination in producing new variation for *C. jejuni* and *C. coli* clade 1.

We see evidence for differences in the relative role of recombination to mutation across the genes (Table 3.3 and Table 3.4). In particular, the parametric bootstrap results show that there is evidence for a lower rate of recombination in *gln*A for *C. coli* and for *gly*A in *C. jejuni*, and for a higher rate in *asp*A in *C. coli*. To assess the

Table 3.3: Allele lengths for each locus; estimates for *C. jejuni* for each housekeeping locus of the probability of an SLV being caused by mutation only (p); the expected number of mutations for an SLV; the relative rate of recombination to mutation; 95% CI for the estimated relative rate of recombination to mutation; and the % of nucleotide differences of an SLV that were introduced by recombination.

Locus	Allele lengths (bp)	p	Expected Number Mut.	Relative Rate of Rec	95% CI	% Differences due to Rec
*asp*A	477	0.09	0.11	8.91	(5.98,16.06)	99
*gln*A	477	0.09	0.11	8.86	(5.96,15.91)	98
*glt*A	402	0.10	0.12	7.84	(5.43,13.12)	97
*gly*A	507	0.21	0.32	2.97	(2.40,3.77)	97
pgm	498	0.10	0.15	7.14	(5.06,11.41)	98
tkt	459	0.11	0.15	6.81	(4.87,10.66)	99
uncA	489	0.12	0.16	6.21	(4.53,9.37)	99
Average	472.71	0.12	0.16	6.96	(6.08,8.09)	98

strength of this evidence, we looked at the lowest (and highest) estimated value of the relative rate of recombination to mutation across the 7 genes in our simulated data divided by the average of estimated rate across the 7 genes. For both *C. coli* and *C. jejuni* we never observed an estimate as low as that for *gln*A and *gly*A respectively across the 100 simulations in each case (the lowest estimates were 0.36 and 0.62 for *C. coli* and *C. jejuni* respectively, compared to observed values of 0.23 and 0.43) or as high as *asp*A for *C. coli* (highest estimate was 2.04, compared to an observed value of 2.21).

Discussion

We have analysed SLVs to infer the relative importance of recombination and mutation to generate differences between closely related *C. jejuni* and *C. coli* clade 1isolates. The higher average estimates for *C. jejuni* compared to *C. coli* demonstrates higher recombination in *C. jejuni*, compared to *C. coli*. This is consistent with the existing population structure (three clades) of *C. coli*, but no apparent subclade structure in *C. jejuni* [335]. We estimate that recombination contributes between 2.97 and 8.91 times more than mutation to events that generate new alleles for *C. jejuni*, depending on the MLST locus, and between 0.23 and 2.23 for *C. coli* clade 1. The variations between housekeeping genes within species also show the different evolution pressure on different genes. For *C. jejuni*, *gly*A has less recombination contribution, compared to the other six genes. For *C. coli*, *gln*A has less recombination contribution, compared to the other six genes.

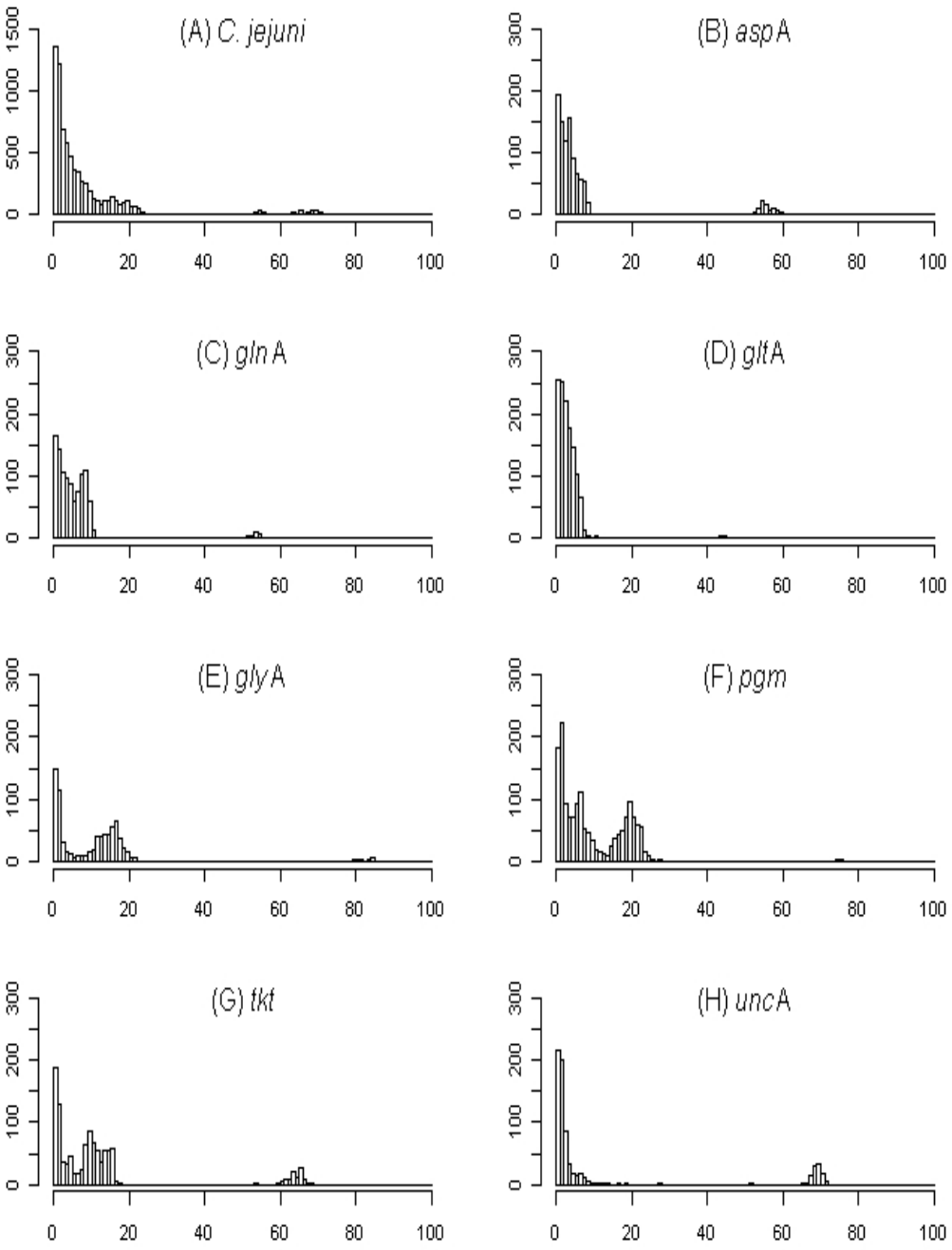


Figure 3.2: SLVs of PubMLST data. The x axes represent the number of nucleotide differences between STs that make up an SLV; y axes represent the number of recorded events. A represents the nucleotide differences for SLVs in the PubMLST database for *C. jejuni*; others are the nucleotide differences for SLVs by loci.

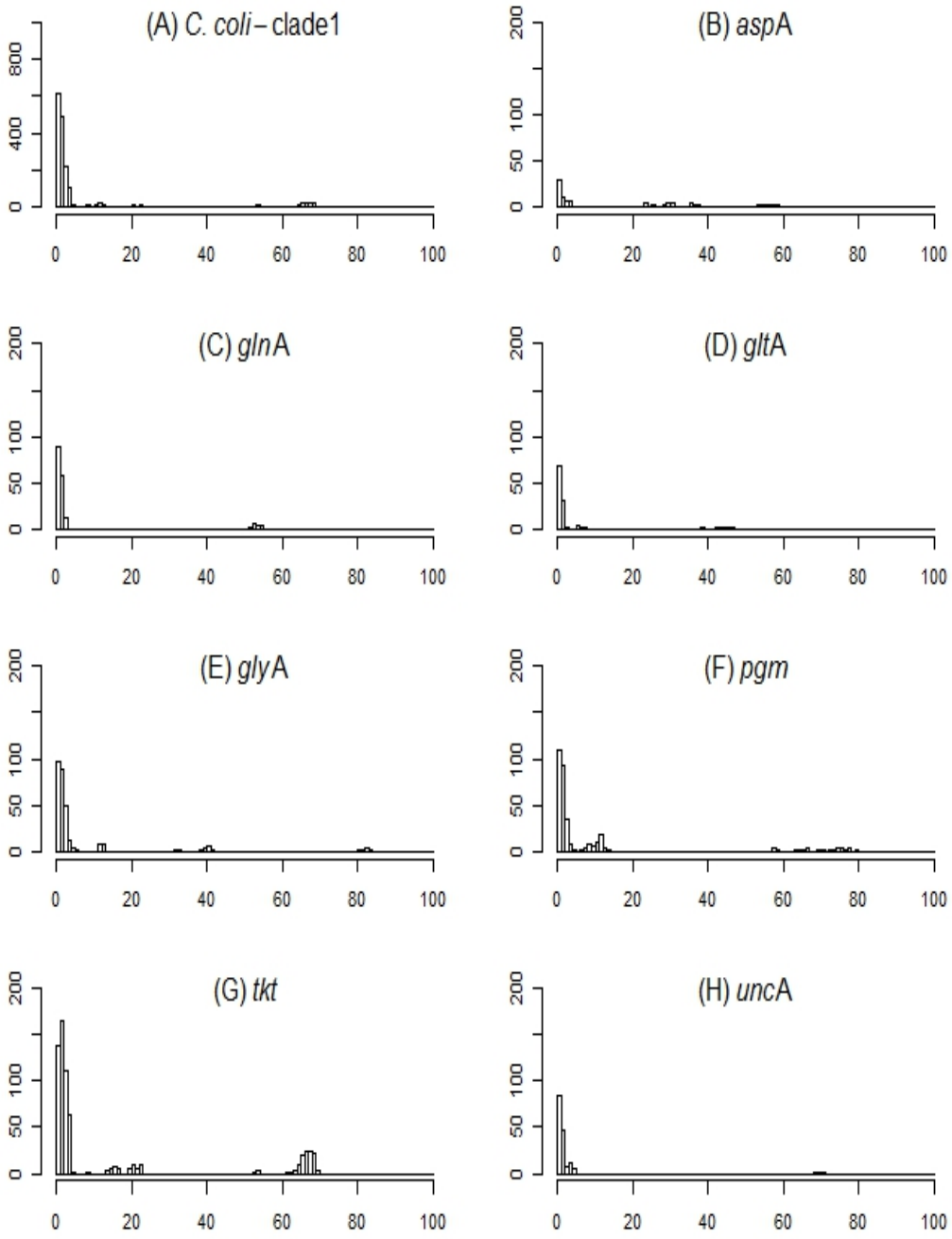


Figure 3.3: SLVs of PubMLST data. The x axes represent the number of nucleotide differences between STs that make up an SLV; y axes represent the number of recorded events. A represents the nucleotide differences for SLVs in the PubMLST database for *C. coli* clade 1; others are the nucleotide differences for SLVs by loci.

Table 3.4: Allele lengths for each locus; estimates for *C. coli* clade 1 for each housekeeping locus of the probability of an SLV being caused by mutation only (p); the expected number of mutations for an SLV; the relative rate of recombination to mutation; 95% CI for the estimated relative rate of recombination to mutation; and the % of nucleotide differences of an SLV that were introduced by recombination.

Locus	Allele lengths (bp)	p	Expected Number Mut.	Relative Rate of Rec	95% CI	% Differences due to Rec
*asp*A	477	0.09	0.11	8.91	(5.98,16.06)	99
*gln*A	477	0.09	0.11	8.86	(5.96,15.91)	98
*glt*A	402	0.10	0.12	7.84	(5.43,13.12)	97
*gly*A	507	0.21	0.32	2.97	(2.40,3.77)	97
pgm	498	0.10	0.15	7.14	(5.06,11.41)	98
tkt	459	0.11	0.15	6.81	(4.87,10.66)	99
uncA	489	0.12	0.16	6.21	(4.53,9.37)	99
Average	472.71	0.12	0.16	6.96	(6.08,8.09)	98

Our analysis has similarities to that of Schouls et al. [325], who used the approach described by Feil et al. [105] to estimate the relative rate of recombination and mutation for *C. jejuni*. The original idea of Feil et al.'s method [105] is put forward by [149]. However, their method overestimates the ratio of recombination to mutation, compared to ours. They also analysed SLVs, though restricted to pairs of SLVs within the same clonal complex. Furthermore, rather than the model-based approach we consider, they used a simple rule to classify which SLVs had been caused by mutation as opposed to recombination. The rule was that if a pair of SLVs varies by a single nucleotide difference and one of the MLST alleles at the locus was unique, it is due to a mutation, whereas all other pairs of SLVs are caused by recombination. This means that, under this algorithm, SLVs that differ by two nucleotide differences could not have arisen by two independent mutation events, and recombination events that mask mutation events are not considered. Both assumptions may lead to an underestimate of mutation. The analysis of Schouls et al. [325] estimate that recombination is approximately eight times more likely to change an allele than mutation. This is larger than our estimate, which is likely to be due to these biases in the method used by Schouls et al. [325]. According to Feil et al.'s method [325], Schouls et al. [325] estimated a recombination size about 3.3kb. We implemented a simplified version of Feil et al.'s method [105] (details in the Supplementary Material), the results show under the 3kb recombination size, the ratio is overestimated.

Our estimates suggest a more important role for recombination in producing new diversity into *C. jejuni* than more recent studies which have analysed samples of *C. jejuni* isolates from different source populations. Fearnhead et al. [99] estimate

that recombination rates are if anything less than mutation rates. While Vos and Didelot [388] and Wilson et al. [403] all give estimates of the proportion of nucleotide differences introduced by recombination as opposed to mutation which are much smaller than the ones we obtain. Both studies concluded that the number of nucleotide differences introduced by recombination are only approximately twice as many as those introduced by point mutation: 2.2 for Vos and Didelot [388], 2.67 (95% CI 1.39, 4.95) for Wilson et al. [403].

The difference between our study and these is that we analyse only SLVs, which means we are looking at closely related STs for which there has been less time for selection to act. Intuitively, selection is likely to be strongest against recombination events that introduce large differences, although it is possible that some recombination events may introduce a section of DNA from an organism that is highly adapted and 'successful' in the given environment. Therefore, although we estimate that recombination is introducing more differences than previously thought in our closely related, recently evolved STs, many of these differences may be subsequently purged from the population due to weak purifying selection. This is consistent with the effects of purifying selection described in Wilson et al.'s paper [403].

Whole genome analysis may provide a greater insight into the genome-wide evolution of *Campylobacter* and provide further explanations for the apparent differences between previous estimates of recombination and mutation. Recently, Biggs et al. (2011) analysed the genomes of two closely related *Campylobacter* ST-474 isolates that also had identical *fla*A SVR regions and compared them to available *C. jejuni* reference strains. They estimated that around 97% of the nucleotide differences between these two closely related isolates were caused by recombination. This estimate is similar to ours, and suggests that the importance of recombination for driving changes in *C. jejuni* is not just confined to the MLST housekeeping genes we have studied.

The aim of this study was to increase our understanding of the evolution of *C. jejuni* and *C. coli* by investigating the generation of SLVs. The availability of the large database of *C. jejuni* and *C. coli* isolates provides a good opportunity to investigate the evolution of *C. jejuni* and *C. coli* using SLVs. Using seven independent housekeeping loci we used the method proposed in this chapter to estimate that recombination contributes roughly seventimes as much as mutation to the generation of SLVs for *C. jejuni*, and equal for *C. coli*, which provides further evidence that recombination plays a more important role in the evolution of *C. jejuni* and *C. coli* than mutation.

Our results also point to important differences in terms of the forces driving evolution for *C. jejuni* and *C. coli*; and suggest that the relative role of recombination to

mutation may differ between genes, and these differences themselves may be different for *C. jejuni* and *C. coli*. Understanding what is causing these differences will be important for fully understanding how these bacteria may evolve in the future. However the fact that we observed differences in recombination between *C. jejuni* and *C. coli* is consistent with the introgression hypothesis of Sheppard et al.'s paper [336], which implies that patterns of genetic exchange have changed over time. The research on SLVs described in this paper could be extended either by considering more genes, such as flagellin genes (*fla*A and *fla*B) [249], and *por*A, the gene encoding the major outer membrane proteins (MOMPs) [49, 417], or by considering other species of *Campylobacter*.

Acknowledgements

We acknowledge the Marsden Fund project 08-MAU-099 (Cows, starlings and *Campylobacter* in New Zealand: unifying phylogeny, genealogy and epidemiology to gain insight into pathogen evolution) for funding this project. During this study we used the MLST website (http://pubmlst.org/campylobacter) developed by Keith Jolley and Man-Suen Chan and sited at the University of Oxford (Jolley et al. 2004 [189], BMC Bioinformatics, 5:86). The development of this site has been funded by the Wellcome Trust. BRH acknowledges the Australian Research Council (grant FT100100031).

Supplementary material

Given two STs that form an SLV, we assume a random variable, H, represents the number of base pairs that differ between the two STs. We are interested in the probability that an SLV at locus i with h differences is due to mutation only, that is, $Pr(h|SLV_i; M)$, where the event M={differences due only to mutation}. M^c is the complement of M, i.e. the event of recombination(s) as well as mutations.

The model for differences due to mutation only

Firstly, the $Pr(h|SLV_i; M)$ was estimated according to Bayesian theory,

$$Pr(h|SLV_i; M) \propto Pr(h|M) \times Pr(SLV_i|h; M), \tag{3.1}$$

where $Pr(SLV_i|h; M)$ represents the likelihood function and $Pr(h|M)$ is the prior distribution of the number of nucleotide differences for the mutation only model.

Table 3.5: Comparison of the effect of different prior parameters for the mutation only model on the probability ratio of recombination vs. mutation to generate a new allele for *C. jejuni*. Only part of the comparison results are shown here.

λ	0.5	0.25	0.01
aspA	8.91	9.86	11.50
glnA	8.86	9.75	10.11
gltA	7.84	8.72	10.11
glyA	2.97	3.78	5.67
pgm	7.14	9.32	13.29
tkt	6.81	7.82	9.00
uncA	6.21	7.06	8.09
Average	6.96	8.04	9.68

Assume we have a mutation rate θ_i and recombination rate ρ_i at locus i. Denote $\theta = \sum_{i=1}^{7} \theta_i$ and $\rho = \sum_{i=1}^{7} \rho_i$. We set $w_i{=}\theta_i/\theta$, so w_i is the probability that if a mutation occurs it occurs at locus i. We estimate the w_is through estimating θ_is by the average number of base-pair differences between all alleles at locus i. Then we model the probability of an SLV at locus i given h base-pair differences, and that the SLV is caused only by mutation, as

$$Pr(SLV_i|h; M) \quad = \quad (w_i)^h. \tag{3.2}$$

This comes from the need that given M, there have been h mutations, and for it to be an SLV all mutations must occur at the same locus (locus i).

Finally, from coalescent theory we model that the probability of h mutations, given that there have only been mutations prior to the common ancestor of the pair of isolates, is geometric with parameter $\lambda = \theta/(1 + \rho + \theta)$. Thus we model

$$Pr(h|M) \quad \propto \quad Geometric(1 - \lambda), \tag{3.3}$$

and make the simplifying assumption that $\theta \approx \rho$ and $\theta, \rho >> 1$, thus, we have $\lambda \approx 0.5$. We use this value of λ in our analysis, but also considered how robust the results were to varying $\lambda < 0.5$ (as it appears that if anything $\rho > \theta$). Tables 3.5 and 3.6 show that the choice of λ does not have a large effect on the results, with different choices of λ giving larger estimates for the relative rate of recombination to mutation.

The recombination related model

The first step is to draw two alleles in that locus randomly based on the frequency of these alleles in one locus in PubMLST. The second step is to compare this pair

Table 3.6: Comparison of the effect of different prior parameters for the mutation only model on the probability ratio of recombination vs. mutation to generate a new allele for *C. coli*. Only part of the comparison results are shown here.

λ	0.5	0.25	0.01
aspA	2.23	0.32	$9.87e^{-15}$
glnA	0.23	0.78	0.49
gltA	0.72	0.55	0.53
glyA	0.54	0.59	0.37
pgm	1.04	0.42	0.35
tkt	0.85	0.47	0.21
uncA	1.44	0.40	0.38
Average	1.01	0.50	0.33

of alleles and record the number of differences. This step was repeated for 1,000,000 iterations to obtain a stable empirical probability distribution for observing h differences due to recombination for this locus: $Pr(h|M^c)$.

A naive approach to estimating the probability of observing h nucleotide differences being introduced by events that include recombination would be:

$$Pr(h|M^c) = \frac{n_h}{n_d} \tag{3.4}$$

in which h represents the number of nucleotide differences; n_h represents the item count of h differences (how many times h differences appears), n_d represents the number of all differences $n_d = \sum_{h=1}^{a} n_h$, where a is the maximum observed number of differences between any pair of alleles for the locus under consideration. However this is not robust, and the reason is that there are some values (say 30 to 45) of h for which $n_h = 0$, i.e. pairs of alleles with 30 to 45 differences are never observed in the sample. Using Equation 3.4 would then estimate the probability of recombination producing such a number of differences as 0; and if we observe h differences in our SLVs data, our model would have to assign this to mutations. A simple way around this is to introduce a Dirichlet prior on $Pr(h|M^c)$, which gives the posterior estimates:

$$Pr(h|M^c) \quad = \quad \frac{1 + n_h}{a + n_d}. \tag{3.5}$$

Mixture model

For this part, we will consider a fixed locus i; and estimate the probability that an SLV was the result of mutation only for that locus (p_i). It was assumed that

there are data h_1, h_2, h_3..., $h_{n_{data}}$, where n_{data} is the number of pairs of distinct STs with SLVs in PubMLST dataset, and h_j $(j = 1, 2, ..., n_{data})$ represents the number of nucleotide differences of the j^{th} SLV.

The distribution for h for an SLV is

$$f(H) = \begin{cases} Pr(H|SLV_i, M), & if\ z = 1 \\ Pr(H|SLV_i, M^c), & if\ z = 0 \end{cases} \tag{3.6}$$

in which $Pr(H|SLV_i, M)$ is the model for solely mutation, and $Pr(H|SLV_i, M^c)$ is the recombination related model. The latent variable Z is introduced as the indicator to tell whether the data h_j comes from either of these two models. For example, when $z_j = 1$, h_j comes from the mutation model; whereas when $z_j = 0$, h_j comes from the recombination model. Thus the probability of $z_j = 1$ is the proportion of SLVs caused solely by mutation (p_i), and

$$f(h_j,\, z_j|SLV_i) = p_i \times Pr(h_j|SLV_i, M) + (1 - p_i) \times Pr(h_j|SLV_i, M^c). \tag{3.7}$$

We can estimate p_i under this model by maximum likelihood using the EM Algorithm. The expectation-maximization (EM) algorithm is an approach for finding maximum-likelihood estimates by iterative computation, when the statistical model depends on unobserved latent variable [65]. It includes two steps: expectation step (E-step) and maximization step (M-step). In the E-step, the expectation of log likelihood was calculated, in the M-step, the expectation was maximized. 100 iterations were run to get stable parameter estimates, although the parameters converged after 50 iterations.

From the EM algorithm, we get:

$$\hat{p}_{i,new} = \frac{\mathrm{E}(M|p_{i,old})}{n}, \tag{3.8}$$

in which

$$\mathrm{E}(M|p_{i,old}) = \sum_{j=1}^{n} Pr(z_j = 1|p_{i,old}). \tag{3.9}$$

The model for an event being mutation rather than recombination

SLVs can be caused by multiple events. Let K be the number of events separating the two branches in the evolutionary tree of each locus. Coalescent theory gives that the number of events (mutation/recombination) between two randomly chosen isolates follows a geometric distribution with parameter $1/(1+\rho+\theta)$:

$$Pr(k) = \left(\frac{\rho+\theta}{1+\rho+\theta}\right)^k \left(\frac{1}{1+\rho+\theta}\right) \tag{3.10}$$

The probability of an SLV at locus i given $K = k$ is

$$Pr(SLV_i|k) = \left(\frac{\rho_i+\theta_i}{\rho+\theta}\right)^k$$

Thus,

$$\begin{aligned} Pr(k|SLV_i) &\propto Pr(SLV_i|k) \times Pr(k) && (3.11)\\ &\propto \left(\frac{\rho_i+\theta_i}{\rho+\theta}\right)^k \left(\frac{\rho+\theta}{1+\rho+\theta}\right)^k && (3.12)\\ &\propto \left(\frac{\rho_i+\theta_i}{1+\rho+\theta}\right)^k, && (3.13) \end{aligned}$$

in which, $Pr(SLV_i|k)$ means the probability that SLVs at locus i are caused by k events. Assuming $\theta+\rho >> 1$ and ρ_i is roughly proportional to θ_i we then have the approximation

$$Pr(K = k|SLV_i) \propto \omega_i^k. \tag{3.14}$$

As $Pr(k|SLV_i)$ is a probability mass function we obtain

$$Pr(k|SLV_i) = (1-\omega_i) \times (\omega_i)^{k-1}, \text{for } i = 1, 2, \ldots. \tag{3.15}$$

Now, p_i has been defined as the probability of an SLV involves only mutation and no recombination at locus i, while x_i is defined as the actual probability that an event for generating new alleles that led to SLVs is mutation rather than recombination at locus i. The relationship between p_i and x_i is thus

$$p_i = \sum_{k=1}^{\infty} (1-\omega_i) \times (\omega_i)^{k-1} x_i^k, \tag{3.16}$$

where the sum is over the number of events, and we need all events to be mutations. Thus we get that $x_i = p_i/(1-\omega_i + p_i \times \omega_i)$. Then $(1-x_i)/x_i$ represents the relative rate of recombination to mutation.

Using the same model we obtain an estimate of the number of mutation events at an SLV at locus i. From Equation 3.15 we have the expected number of events is $\mathrm{E}(K|SLV_i) = 1/(1-\omega_i)$, and a proportion x_i of all such events are mutations. Thus the expected number of mutation events is $x_i/(1-\omega_i)$. The proportion of nucleotide differences of an SLV due to recombination is calculated by $1 - x_i/(d - d \times \omega_i)$, d is the average number of differences in all SLVs at locus i.

The comparison with Feil et al.'s method

A simplified version of Feil et al.'s [105] method that assumed that all differences of one nucleotide were caused by mutation, but larger nucleotide differences were due to recombination was also applied.

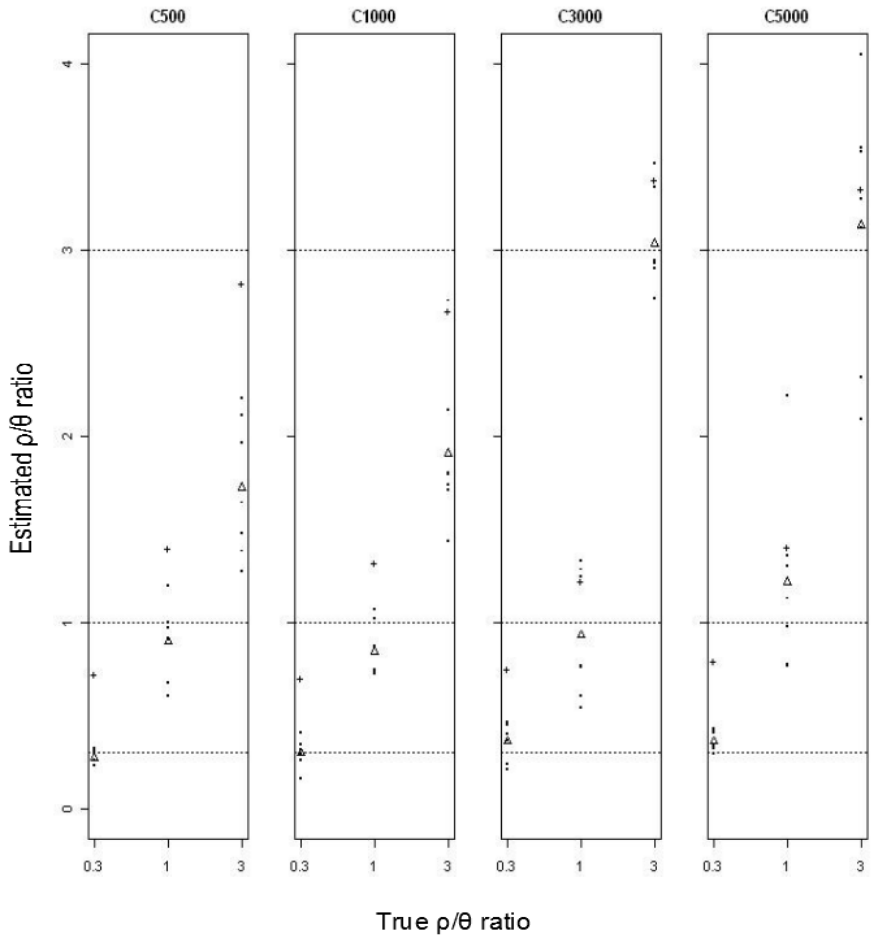
Results (Supplementary Figure 1) show, for the simplified version of Feil et al.'s method, the ratio of recombination to mutation were overestimated, compared to our results. As Feil et al.'s full method has the potential to underestimate mutation even more, their estimation of the ratio of recombination to mutation will be apparently higher than our estimates.

Supplementary Table 1: Comparison of the effect of different prior parameters for the mutation only model on the probability ratio of recombination vs. mutation to generate a new allele for *C. jejuni*. Only part of the comparison results are shown here.

λ	0.5	0.25	0.01
*asp*A	8.91	9.86	11.50
*gln*A	8.86	9.75	10.11
*glt*A	7.84	8.72	10.11
*gly*A	2.97	3.78	5.67
pgm	7.14	9.32	13.29
tkt	6.81	7.82	9.00
*unc*A	6.21	7.06	8.09
Average	6.96	8.04	9.68

Supplementary Table 2: Comparison of the effect of different prior parameters for the mutation only model on the probability ratio of recombination vs. mutation to generate a new allele for *C. coli*. Only part of the comparison results are shown here.

λ	0.5	0.25	0.01
*asp*A	2.23	0.32	$9.87e^{-15}$
*gln*A	0.23	0.78	0.49
*glt*A	0.72	0.55	0.53
*gly*A	0.54	0.59	0.37
pgm	1.04	0.42	0.35
tkt	0.85	0.47	0.21
*unc*A	1.44	0.40	0.38
Average	1.01	0.50	0.33



Supplementary Figure 1: Simulation work under constant population size models. X-axes represent the simulation work under different given ratios of recombination to mutation; y-axes represent the ratio of recombination to mutation (ρ/θ). The four plots represent the results under constant population size models. True values are shown as dotted horizontal broken lines, dots are our estimations, triangles are the average for the seven estimates from loci, and plus signs were calculated from Feil et al.'s method.

Chapter 4

The relative roles of recombination and point mutation to the generation of single locus variants in a range of bacterial pathogens

4.1 Summary

Within a bacterial species, a clonal complex is a cluster of closely related bacterial strains that group around a founder (or ancestral) strain. Estimates of the rate of DNA sequence evolution at the clonal complex level have been found to be faster than those obtained using more distantly related isolates [71], perhaps due to the lack of time for purifying selection to act. By making use of DNA sequence data gathered as part of multilocus sequence typing (MLST) schemes, research can focus on pairs of strains within clonal complexes that share a very recent common ancestor. These pairs of closely related strains are known as single locus variants (SLVs) as they differ at only one gene of the seven to eight genes used in the MLST scheme. This study used an expectation-maximization (EM) algorithm to fit a model that accounts for differences between pairs of strains caused by both point mutations (changes at single sites in the sequence) and recombination events (which affect blocks of nucleotides). The estimated ratios of recombination events to point mutations were larger for SLVs than the ratios for more distantly related strains for the majority of the tested bacterial species. This indicates that the purifying selection may act more stringently on recombination events than on point mutations. Results in this chapter are predominantly consistent with previous research. There has been some debate in the last decade about the relative importance of point mutations versus

recombination for the species *Staphylococcus aureus*; results in this chapter support that recombination is more important than point mutation.

4.2 Introduction

From Chapter 3, it can been been seen that recombination has been more important than mutation in producing genetic diversity in both *C. jejuni* and *C. coli*. This finding indicates that purifying selection plays an important role in the evolution of *Campylobacter*.

In molecular epidemiology, multilocus sequence typing (MLST) is a widely used technique for typing bacteria based on the nucleotide sequences of multiple housekeeping genes [237]. This technique usually types seven housekeeping genes by the analysis of 400 to 500 bp. For each locus, a distinct allele number is assigned to a unique gene sequence. For each distinct combination of seven numbers, a distinct sequence type (ST) number is given. A single locus variant (SLV) is a pair of sequence types (STs) that differ at one and only one out of (usually) the seven alleles that make up the MLST profile [103]. SLVs are pairs of STs that are closely related in the evolution of clonal complexes. The evolution of STs in clonality can be quite different from the evolution of distantly related STs [71, 104, 149, 388], especially when selection pressure plays a part in the evolution [71]. A clonal complex is a group of STs that have most likely evolved from a recent common ancestor. The pairs of STs that share a very recent common ancestor are more likely to be SLVs. Therefore, comparing the evolutionary estimates, such as the ratio of recombination to mutation, to that of distantly related STs not only helps our understanding of the evolution and molecular epidemiology of pathogens, but also reveals the role played by selection pressure.

SLVs can be caused by recombination and/or point mutation(s). Recombination has been proved to be a major driving force of evolution in bacteria [71, 99, 357], and can occur through three mechanisms: conjugation, transduction and transformation [71, 368]. The relative roles of recombination and point mutation in generating genetic diversity has also been estimated in previous studies [101, 102, 104, 106, 320]. Although it has been concluded that recombination occurs and contributes much more than mutation to genetic diversity [71], it is still unclear about the relative number of events and nucleotide differences that may be attributable to recombination and mutation processes in closely related STs, such as SLVs. At present, MLST profiles have been characterized and stored in large accessible databases. The isolates are stored in collections. The availability of these data provides a good opportunity to study SLVs in detail. In this chapter, a statistical method (put forward

in Chapter 3) is applied to estimate the relative roles of recombination and point mutation in generating SLVs in a range of bacterial pathogens.

In order to estimate the relative roles of recombination and point mutation in generating SLVs, the ratio of the occurrence rates of recombination and mutation (ρ/θ [253]) in generating SLVs is used. This evolutionary-based measurement has been estimated in many previous studies [95, 99, 103, 104]. Another commonly used measurement is the ratio of rates at which one nucleotide is changed by recombination and point mutation, r/m. The main difference between these two ratio estimates is that r/m measures the per-site ratio of a nucleotide being substituted by recombination or mutation, whereas ρ/θ is the per-event ratio. For example,if $\rho/\theta = 3$, then the occurrence of recombination events is three times as frequent as point mutation, and if one recombination event introduced six nucleotide differences, then r/m equals 18.

The r/m ratio can be estimated by several well-known software packages, such as eBURST [103] and ClonalFrame [95], whereas previous research has failed to consider estimating ρ/θ using a mathematical model [101, 102, 104, 105]. Although ClonalFrame [95] has an option for calculating ρ and θ, little research has been done to estimate this ratio. One possible reason for this is the gap of estimating the ρ/θ ratio is that this ratio alone does not give the impact of recombination on the evolutionary process of the population, and the estimation is not easy to calculate when only comparing sequences by manual inspection. In addition, ClonalFrame assumes recombination comes from outside the population, which could underestimate the recombination effect [69]. In this chapter, a method is proposed to estimate the ρ/θ ratio of STs with SLVs, and to also note the impact of recombination on the evolution of the population.

This research is based on several bacterial pathogens (selection criteria for analysed bacteria will be explained in the Methods section): *Bacillus cereus* (*B. cereus*); *Enterococcus faecium* (*E. faecium*); *Haemophilus influenzae* (*H. influenzae*); *Klebsiella pneumoniae* (*K. pneumoniae*); *Streptococcus uberis* (*S. uberis*); *Streptococcus zooepidemicus* (*S. zooepidemicus*); *Staphylococcus aureus* (*S. aureus*); *Neisseria lactamica*; *Neisseria gonorrhoeae*; and *Neisseria meningitidis*. All the STs are downloaded from publicly accessible databases: *B. cereus, S. uberis*, *S. zooepidemicus*, and *Neisseria* spp. from the PubMLST database (http://pubmlst.org); *E. faecium*, *H. influenzae*, *S. aureus*, from the Multi Locus Sequence Typing database (http://www.mlst.net); and *K. pneumoniae* from the Institut Pasteur MLST Databases (http://www.pasteur.fr). In order to understand whether there are differences in the mechanisms that produce SLVs across the bacterial genomes under analysis, SLVs were divided into groups of multi-locus genotypes, depending on the locus at

which the STs differ. The distribution of nucleotide differences within SLVs was explored. The nucleotide differences between two STs that form an SLV can be generated by two different kinds of events: recombination or mutation. Intuitively, SLVs that comprise two STs which differ at many nucleotide positions are more likely to be due to recombination, whereas those that differ at only a few nucleotide positions may be the result of point mutations. In this study, an expectation-maximization (EM) algorithm (described in Chapter 3) was applied to allocate SLVs into either a recombination model or a point mutation-only model. Two key parameters were estimated: the proportion of SLVs that arose due to point mutation(s) only, and the probability that an event which led to a new allele was a point mutation rather than a recombination.

Firstly, a direct comparison among these tested bacteria is made, then these estimates are compared to the results from previous studies on distantly related STs (non-SLVs) [388]. These also are compared to the results from previous studies on the closely related isolates, but on a limited number of isolates ($<$200) [101, 104, 105]. Then the relationship is tested between the occurrence of SLVs and clonal complexes for a given species.

4.2.1 A brief introduction into the selected bacteria

Bacillus cereus is a gram-positive, rod-shaped, facultative anaerobic, spore-forming bacterium that can cause food poisoning [31]. Typical symptoms include diarrhoea, severe nausea, or vomiting [211]. This bacteria can be found in protein rich edible material, cooked rice dishes, or improperly cooked or stored food [378]. Most patients will recover from illness within one day [84].

Enterococcus faecium (*E. faecium*) is a non-mobile, gram-positive, spherical bacterium that colonizes in pairs or chains [56, 313]. It can be found in many human organs, including the gastrointestinal tract and skin, and on some inanimate objects. It can cause several illnesses in humans, including nosocomial infections, surgical wound infections, and urinary tract infections. Some strains can be highly drug resistant, especially to vancomycin, penicillin, and gentamicin [241].

Haemophilus influenzae is a gram-negative, non motile, rod-shaped bacterium [216]. These bacteria can be found as normal flora in the human nose, throat, and the upper respiratory tract, but they can also cause life threatening diseases, such as meningitis and pneumonia, when hosts are infected by other factors, including viral infections, or reduced immune function [115, 119, 408]. The disease is spread through person-to-person contact via nasal discharges and other body fluids contaminated with the bacteria [50, 182, 377]. In developing countries, where the vaccine is not commonly

used, half a million children under five years old die due to the infections caused by *Haemophilus influenzae* [24, 263]. In developed countries, such as the US, where the *Haemophilus* vaccine is widely used for children, the incidence of *Haemophilus* infections has decreased largely [5, 27, 289].

Klebsiella pneumoniae is a gram-negative, non-motile, rod shaped bacterium [313]. It can be found in the mouth, skin, and intestines of human as commensal flora. When it exist in the in the upper respiratory tract or in blood it can cause infections. The disease caused by *Klebsiella pneumoniae* can be spread by through person-to-person contact. A range of diseases can be caused by exposure to the bacteria, including pneumonia, upper respiratory tract infection, urinary tract infection, wound infection, even meningitis, and septicemia. If untreated the death rate of pneumonia, caused by *Klebsiella pneumoniae*, is high (ranging from 50% to 70%) [16, 382]. These can cause infections in hospitalized patients [276].

Streptococcus uberis (*S. uberis*) is one of the major causes of bovine mastitis all over the world [242]. This mastitis pathogen can be found in environmental reservoirs, mammary reservoir, manure, and bedding. The economic loss caused by clinical mastitis is huge [53, 210, 396].

Streptococcus zooepidemicus is a non-motile, gram-positive, and coccoid bacterium [129]. It can be found in a wide range of animal species, such as cattle, sheep, and horses [219, 370]. It has been recognized as a commensal organism of horses, but it can cause infection in the upper respiratory tract of horses [229] and mastitis in cattle [354]. It may also cause foodborne infections in humans [28, 219].

Staphylococcus aureus is a nonmotile, gram-positive, facultatively anaerobic, and spherical bacterium, which can appear as single, pairs, or grape-like clusters under a microscope. It can frequently be found in the nose and on the skin as normal skin flora [206]. About a quarter of the healthy human population are carriers of *S. aureus*, without any active infection [183, 206, 235]. *S. aureus* can cause diseases both in humans and domestic animals. In human, it can cause minor skin infections, such as abscesses and carbuncles, as well as life-threatening diseases, such as meningitis, bacteremia, and septicemia. Furthermore, *Staphylococcus aureus* is one of the leading causes of nosocomial infections, especially after surgery [206]. It can cause mastitis in dairy cows [41] and bumblefoot in poultry [185].

In general, *Neisseria* spp. are non-pathogenic except for *Neisseria gonorrhoeae* and *Neisseria meningitidis* [23, 184]. However, sometimes some normal commensal species like *Neisseria lactamica* can still cause life threatening diseases like pneumonia and septicaemia in immunocompromised patients [324]. The general population may be carriers of *Neisseria meningitidis*, but *Neisseria gonorrhoeae* has only been found after sexual contact with infected individuals.

Neisseria lactamica is a gram-negative, oxidase-positive, diplococcus bacterium [30, 413]. It is a harmless human commensal species found in the upper respiratory tracts of most children under five years old [22, 140]. The colonisation by *Neisseria lactamica* has an inverse relationship with *Neisseria meningitidis* [22, 52, 140]. The existence of *Neisseria lactamica* can actually reduce the chances of being infected by *Neisseria meningitidis*.

Neisseria gonorrhoeae is a gram-negative, aerobic, diplococcus bacterium [9, 297, 314]. These bacteria can invade mucous membranes of the mouth, throat, eyes, and anus of males and females [350, 391]. It can adhere to the epithelial cells and penetrates and reaches into the subepithelial space to cause the symptoms of the disease [167, 141]. These bacteria can spread through sexual contact or the infected mother to child during delivery. Millions of people are infected by *Neisseria gonorrhoeae* annually [23, 394, 412].

Neisseria meningitidis can be spread through direct and prolonged general contact with infected persons, like kissing, coughing, or sneezing over someone [312]. Because these bacteria only can get iron from human sources, they only infect humans and not animals.

4.3 Material and Methods

4.3.1 Isolates

There are 553 distinct STs with 281 SLVs for *B. cereus*; 336 distinct STs with 481 SLVs for *E. faecium*; 795 distinct STs with 977 SLVs for *H. influenzae*; 650 distinct STs with 404 SLVs for *K. pneumoniae*; 475 distinct STs with 356 SLVs for *S. uberis*; 272 distinct STs with 148 SLVs for *S. zooepidemicus*; 1997 distinct STs with 7982 SLVs for *S. aureus*; 340 distinct STs with 282 SLVs for *N. lactamica*; 206 distinct STs with 418 SLVs for *N. gonorrhoeae*; and 8673 distinct STs with 32000 SLVs for *N. meningitidis* (Table 4.1). These 10 examples were chosen as they meet our criteria of coming from publicly available MLST datasets and each have more than 250 STs.

4.3.2 Modelling procedure

SLVs can be generated by recombination(s) or mutation(s)in the process of evolution. In this chapter, a recombination event refers to a contiguous segment of DNA being exchanged between two STs, while a point mutation means a single nucleotide change. An event refers to either a recombination or a mutation, and an SLV can

Table 4.1: Number of STs, Number of SLVs and ratio of SLVs to STs for several bacteria

Locus	Number of STs	Number of SLVs	Ratio of SLVs to STs
B. cereus	553	281	0.51
E. faecium	336	481	1.43
H. influenzae	795	977	1.23
K. pneumoniae	650	404	0.62
S. uberis	475	356	0.75
S. zooepidemicus	272	148	0.54
S. aureus	1997	7982	4.00
N. lactamica	340	282	0.83
N. gonorrhoeae	206	418	2.03
N. meningitidis	8673	32000	3.69

be generated by one or more events. In reality, recombination event(s) are quite common in the evolution of bacteria [71], and the effect of recombination tends to mask mutation event(s) because one recombination event can introduce multiple nucleotide differences. In order to simplify the situation, a probability model for the number of nucleotide differences within one SLV being introduced solely by mutation was put forward, and this model also estimates the ρ/θ ratio (the relative occurrence frequency of recombination to point mutation). More details on building the model are described in Chapter 3.

There were two major steps in preparing for carrying out this analysis. Firstly, data were collected from publicly accessible MLST websites. Secondly, SLVs were recorded. This step produced a table which contained, for each pair of STs that form an SLV, the locus at which the pair of STs differed and the number of nucleotide differences for that given pair of STs.

In order to estimate the ρ/θ ratio, the method described in the published paper [414] (Chapter 3) was applied. Two models (more details about mutation and recombination models in section 4.6) were put forward: one calculating the probability of observing $h\,(h = 1; 2; 3...)$ nucleotide differences introduced solely by point mutation event(s); the other was the recombination-related event model.

4.4 Results

4.4.1 The distribution of nucleotide differences within SLV for each bacterium

Each SLV relates to one pair of STs, and the plots (Figure 4.1) show the nucleotide differences that occurred within those pairs of STs at each MLST locus for all tested bacteria. For an SLV, the number of nucleotide differences can be counted. These plots show the frequencies for each nucleotide difference within the pairs of STs with SLVs.

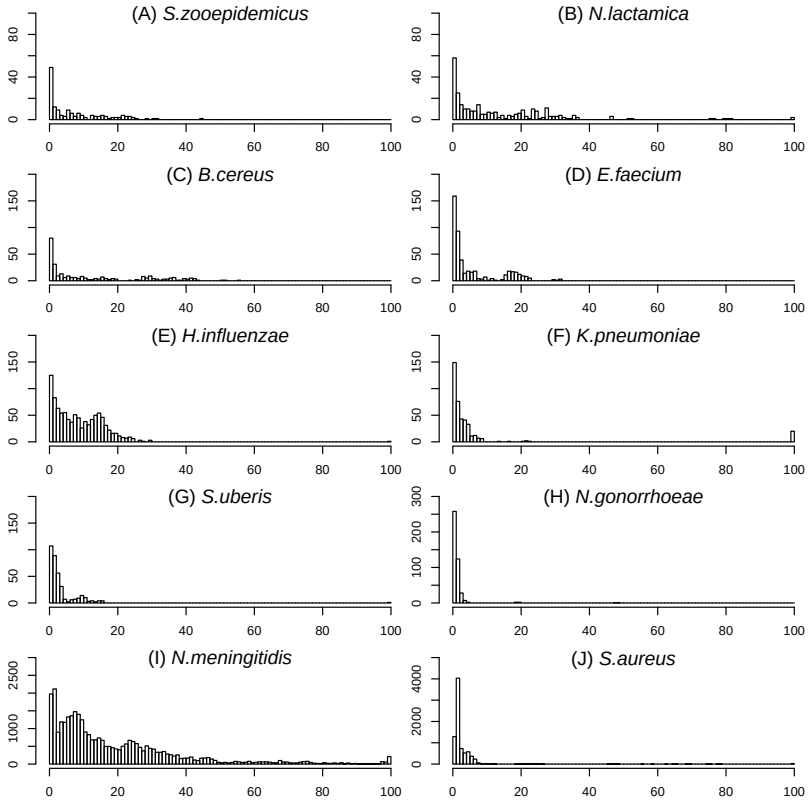


Figure 4.1: Number of nucleotide differences in SLVs for several separate bacteria. The x-axes represent the number of nucleotide differences between STs that make up an SLV; y-axes represent the number of recorded events, and different scales are used, due to wildly different values on the y-axis. Nucleotide differences larger than 100 have been plotted at 100 on x-axes scales. In order to make it easier to compare, these plots are ordered by the scale of the y-axis.

These plots show that SLVs with a large number of nucleotide differences (larger than 45) occurred in several tested bacteria, such as *B. cereus*, *K. pneumoniae*, *S. uberis*, and *N. meningitidis*. The pairs of STs with a large number of nucleotide differences (50–80) are almost certainly due to recombination, as it is highly unlikely that more than 50 independent point mutations would occur at a single locus while the other six loci remained unchanged.

Table 4.2 shows the differences between the estimation of the ρ/θ ratio across various bacteria. The average ratio of recombination events to mutation events from the seven loci is 1.97 for *B. cereus*, 3.09 for *E. faecium*, 17.13 for *H. influenzae*, 1.75 for *K. pneumoniae*, 7.17 for *S. uberis*, 2.54 for *S. zooepidemicus*, 3044.21 for *S. aureus*, 2.54 for *N. lactamica*, 0.07 for *N. gonorrhoeae*, and 19.39 for *N. meningitidis* (Table 4.2). This table demonstrates that recombination contributed more to the generation of SLVs than did mutation for all of the analyzed bacteria, except for *N. gonorrhoeae*.

Table 4.3 compares the differences between the median of the ρ/θ estimations across various bacteria. Compared to Table 4.2, these medians are not affected by the extreme values. For *S. aureus*, the median is 8.1, whereas the average ratio of recombination events to mutation events is 3044.21. Similarly, for *H. influenzae*, unlike the mean value of ρ/θ in 4.2 (17.13), the median of ρ/θ is 5.71.

For all bacteria, the majority of SLVs have a small number of nucleotide differences (less than 20). Regardless of the number of distinct STs, some bacteria, such as *K. pneumonia*, *S. uberis*, and *S. aureus*, have a very large number of nucleotide differences in some rare cases (larger than 200). For *Neisseria*, there are 31059 SLVs, but the largest number of nucleotide differences is 137.

Table 4.4 shows the comparison between estimates (ρ/θ) in this chapter and the estimates from the most recent analysis[388] which estimating the relative of recombination to mutation for a range of bacteria. From 4.4, it can be seen that ρ/θ and r/m have the similar order of magnitude. By definition, ρ/θ should be larger than r/m, because ρ/θ incorporate the length of recombination, whereas r/m does not. The similar order of magnitude ρ/θ and r/m is important because it demonstrates estimates in this chapter are much larger than previous analysis [388].

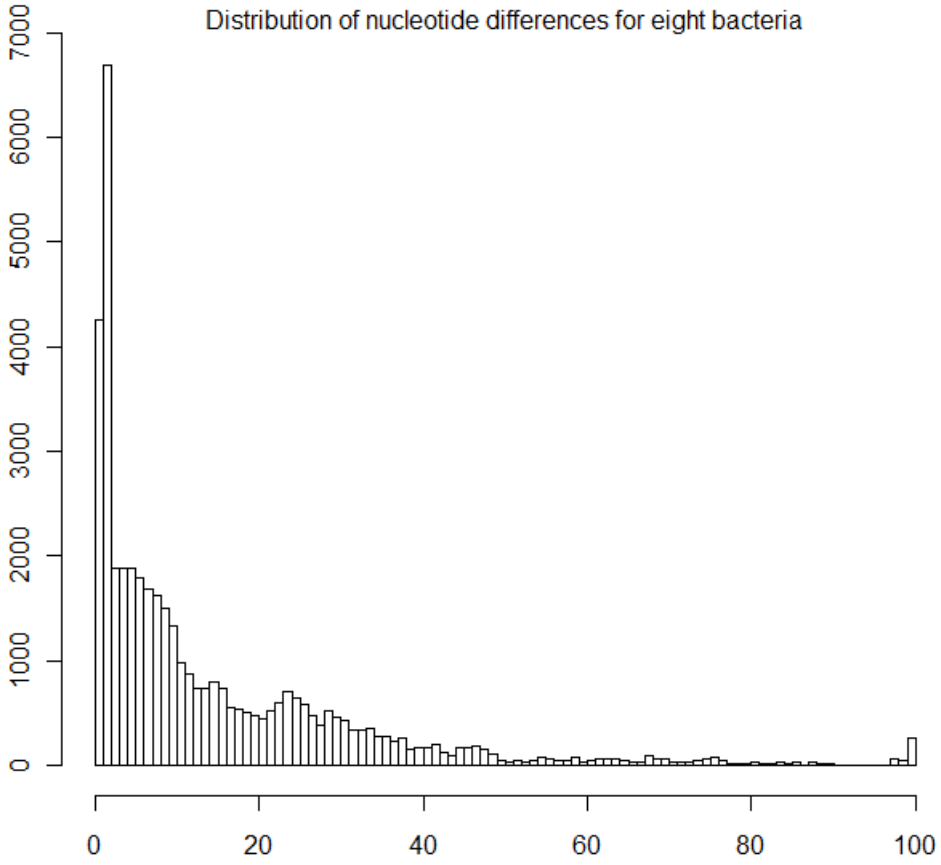


Figure 4.2: Number of nucleotide differences in SLVs for all tested bacteria. The x-axis represents the number of nucleotide differences between STs that make up an SLV; the y-axis represents the number of recorded events. Nucleotide differences larger than 100 have been plotted at 100 on the x-axis scale.

Table 4.2: Estimates of several bacteria for each housekeeping locus in MLST profile of the probability of an SLV being caused by mutation only (p); the expected number of mutations for an SLV; the relative rate of recombination to mutation; and the % of nucleotide differences of an SLV that were introduced by recombination.

Locus	p	Expected Number Mut.	Relative Rate of Rec (Mean)	% Differences due to Rec
B. cereus	0.38	0.47	1.97	90
E. faecium	0.32	0.39	3.09	84
H. influenzae	0.14	0.18	17.13	98
K. pneumoniae	0.47	0.57	1.75	78
S. uberis	0.19	0.25	7.17	89
S. zooepidemicus	0.36	0.46	2.54	90
S. aureus	0.15	0.31	3044.21	90
N. lactamica	0.36	0.46	2.54	90
N. gonorrhoeae	0.93	1.32	0.07	27
N. meningitidis	0.05	0.10	19.39	100

Table 4.3: Median estimates of several bacteria for each housekeeping locus in MLST profile of the probability of an SLV being caused by mutation only (p); the expected number of mutations for an SLV; the relative rate of recombination to mutation; and the % of nucleotide differences of an SLV that were introduced by recombination.

Locus	p	Expected Number Mut.	Relative Rate of Rec (Median)	% Differences due to Rec
B. cereus	0.29	0.38	2.09	95
E. faecium	0.37	0.48	1.43	87
H. influenzae	0.14	0.17	5.73	98
K. pneumoniae	0.41	0.47	1.31	88
S. uberis	0.1	0.14	7.51	93
S. zooepidemicus	0.34	0.45	1.59	94
S. aureus	0.1	0.12	8.1	96
N. lactamica	0.24	0.3	2.33	98
N. gonorrhoeae	0.98	1.04	0.02	27
N. meningitidis	0.05	0.06	17.37	100

Table 4.4: Estimates of several bacteria for comparison between ρ/θ from the method in this chapter and r/m from previous analysis. NA represents data not available in Vos and Didelot' s paper [388].

Locus	Average Relative Rate of Rec (ρ/θ)	Median Relative Rate of Rec (ρ/θ)	Estimates from Vos and Didelot' s paper[388] (r/m)
B. cereus	1.97	2.09	0.7
E. faecium	3.09	1.43	6.2
H. influenzae	17.13	5.73	3.7
K. pneumoniae	1.75	1.31	0.3
S. uberis	7.17	7.51	NA
S. zooepidemicus	2.54	1.59	NA
S. aureus	3044.21	8.1	NA
N. lactamica	2.54	2.33	6.2
N. gonorrhoeae	0.07	0.02	NA
N. meningitidis	19.39	17.37	7.1

4.4.2 Estimates for several bacteria by loci

The per-event ratio of a recombination to a point mutation ρ/θ varies for different bacteria (shown in Tables 4.5 to 4.14 in section 4.6). The range of per event ratio varies by bacteria, and even by loci. Together, these tables show for each given bacterium, the estimation of the ρ/θ ratio varies across different genes.

For *B. cereus* (Table 4.5), the relative role of recombination is larger than mutation for loci *glp*, *ilv*, *pur*, and *pyc*, but smaller than mutation for *gmk*, *pta*, and *tpi*. Similarly, for *E. faecium* (Table 4.6), the relative role of recombination is larger than mutation for loci *atpA*, *ddl*, *gdh*, *purK*, and *pstS*, but smaller than mutation for *gyd*, and *adk*. Compared to other loci, the relative role of recombination is larger for *atpA* and *ddl* for *E. faecium*. For *H. influenzae* (Table 4.7), the relative role of recombination is larger than mutation for all seven loci, with the maximum value for locus *fucK*. The average value is heavily affected by this extreme value. For *K. pneumoniae* (Table 4.8), the relative role of recombination is larger than mutation for loci *gapA*, *pgi*, *phoE*, and *tonB*, but smaller than mutation for *infB*, *mdh*, and *rpoB*. For *S. uberis*, the relative role of recombination is larger than mutation for all loci except for *tpi*. For *S. zooepidemicus*, the relative role of recombination is larger than mutation for all loci except for *tdk*. For *S. aureus*, the relative role of recombination is larger than mutation for all loci except for *aroE*. For *N. lactamica* (Table 4.12), the relative role of recombination is larger than mutation for all loci except for *aroE* and *pdhC*. There are many *N. meningitidis* MLST data available in

PubMLST (Table 4.14) , and large numbers of SLVs occurred in all seven loci. For all the MLST loci, the relative rates of recombination to mutation for *N. meningitidis* are all larger than 10, except for locus *aroE*. Five loci (*abcZ*, *adk*, *fumC*, *pdhC*, and *pgm*) have been estimated as 100% of nucleotide differences due to recombinations. In contrast, for *N. gonorrhoeae* (Table 4.13), the relative role of recombination is smaller than mutation for all loci.

Compared to other species, the differences among different loci is not very large for *B. cereus* and *K. pneumoniae*. In contrast, for both *H. influenzae* and *S. aureus*, the average value is heavily affected by the largest value. Compared to other species, the differences among different loci is very large for *S. aureus*. Table 4.11 shows the average value (3044.21) is heavily affected by the extreme value of 21245.94 for locus *glpF*. The average value of the relative role of recombination is 12.65, if the highest and lowest values are removed. For *N. gonorrhoeae* (Table 4.13), the relative role of recombination is smaller than mutation for all loci. This is very different to other species. In addition, the number of SLVs varies largely across the seven loci, which is different from other species as well.

For all of these tables, the first two columns relate to mutations, and the third and fourth to recombinations. However, the greater probability of an SLV being caused by mutation only does not lead to the smaller relative rate of recombination to mutation.

4.5 Discussion

We firstly make a direct comparison among these tested bacteria. Then we compare these estimates to the results from previous research [69, 68, 102, 371, 388] on distantly related STs (non-SLVs), and also to the results from previous research on the STs with SLVs. The direct comparison of several bacteria shows that estimates of ρ/θ vary (listed by order from small to large): *N. gonorrhoeae* (0.07); *K. pneumoniae* (1.75); *B. cereus* (1.97); *S. zooepidemicus* (2.54); *N. lactamica* (2.54); *E. faecium* (3.09); *S. uberis* (7.17); *H. influenzae* (17.13); *N. meningitidis* (19.39); and *S. aureus* (3044.21). Based on previous studies [69, 68, 102, 105, 371, 388], the per-site r/m ratio for different bacteria varies, and is listed from low to high: *S. aureus* (0.1); *K. pneumoniae* (0.3 [388]); *B. cereus* (0.7 [68, 388]; 1.3 to 2.8 [69]); *E. faecium* (1.1 [388]); *H. influenzae* (3.7 [388]); *Neisseria meningitidis* (5 [69], 7.1 [388], 80 to 100 [102, 105]); *N. lactamica* (6.2 [388]); and *S. uberis* (226 [371]). This comparison shows the difference between r/m and ρ/θ estimates, and one estimate cannot be used to infer the other. For each bacteria, if the distribution of the number of differences is similar to each other (or similar to the total distribution), the ρ/θ ratio

should have the same order of magnitude. The different mechanisms or evolutionary dynamics among species produce different distributions of the number of differences for SLVs.

In these tables (Tables 4.5 to 4.14 in section 4.6), the range is from 1.75 to 19.39, except for *S. aureus* (3000+) and *N. gonorrhoeae* (0.07). The extreme values for these two species may be due to several reasons. The main reason could be the different ability to recombine across different bacteria. *N. gonorrhoeae* can only be found in an infected person, or persons who have had sexual contact with infected persons. The reason for the limited ability of *N. gonorrhoeae* to adapt a new environment could be lack of recombination. In addition, the great ability of *S. aureus* to exist widely could be due to a higher ratio of recombination to mutation.

The differences in the third column in Table 4.2 demonstrate the differences between the estimation of the ρ/θ ratio from the average recombination and mutation rates for each bacterium. This variation may demonstrate different evolutionary dynamics in different genes and species. Different bacteria have the different capacities for genetic exchange [101, 106, 225, 388], some bacteria can uptake of DNA from the extracellular environment and integrate the free DNA into their genomes, but some bacteria do not have this capacity [234].

Table 4.4 shows estimates in this chapter are much larger than the more recent analysis [388] on estimating the relative recombination to mutation. One explanation is that the estimates in this chapter have been calculated from closely related STs (SLVs), whereas Vos and Dedilot's analysis [388] works on distantly related ST. This comparison demonstrates that the purifying selection plays a role in the evolution of all the compared species: *B. cereus*, *E. faecium*, *H. influenzae*, *K. pneumoniae*, *N. lactamica*, and *Neisseria meningitidis*.

A range of previous research on ρ/θ estimates show that the ratio for overall *Neisseria* is 0.7 to 1.2 [69] or 3.6 to 5 [102, 100]. The latter range is calculated on SLVs, whereas the former is calculated on more distantly related strains. This difference already shows that the closely related strains in *Neisseria* have a higher ρ/θ ratio [100, 102] than that of distantly related strains [69]. Similarly, the ρ/θ ratio for *B. cereus* is 0.125 to 0.25 [68] or 0.2 to 0.5 [69]. Unlike estimate in this chapter for *B. cereus* (1.97), their smaller estimates have been calculated based on distantly related STs. This comparison shows that the purifying selection plays a role in the evolution of both *Neisseria* and *B. cereus*.

For analyzed bacteria, the estimates based on SLV strains are much larger than those based on the more distant strains, although slightly different estimates were obtained using different methods. The larger estimates for clonal strains indicate purifying selection plays a role for the tested species. Most of the previous research

on SLV strains are based on Feil's methods [101, 104, 105], which were originally put forward by Guttman and Dykhuizen (1994) [149]. As described in their research [101, 104, 105], Feil's analysis provides a lower bound of the ρ/θ estimates. With the increasing amount of MLST data, the per-event ratio of recombination to mutation can be calculated and refined.

Feil's method [101, 104, 105] also works on SLV data from MLST. Compared with Feil's method [101, 104, 105], in this chapter's analysis more STs are available and analysed, so the results are less biased by the sampling methods. In addition, the new method put forward in this chapter is less labour intensive compared with Feil's and other studies based on manual inspection. In addition, the estimates are calculated for each locus, rather than only one value used to represent seven genes across the genome.

The estimate in this chapter for *S. aureus* is consistent with some previous publications [62, 102, 343]. However, the larger recombination effect than mutation for *S. aureus* is different from a previous study [100]. That study [100] indicates that the mutation effects are larger than the recombination, and also identified the error in one previous publication [62]. Feil et al. [100] claim that some errors were found in the data used in Day's analysis [62]; therefore, they used the revised data to calculate the recombination and mutation effect for *S. aureus*. Both studies [62, 100] applied Feil's methods. One limitation of Feil's method is that the analysis is based on a limited number of STs and SLVs (there are 35 SLVs out of 75 unique STs). Recent research on the comparison of recombination to mutation in *S. aureus* [18] indicates that the role played by recombination may be quite large, but the result in that paper is not conclusive.

Combined with the results for *Campylobacter*, recombination occurs more frequently than mutation for the majority of tested bacteria, except for *N. gonorrhoeae*. The smaller ratio of recombination rate to mutation rate for *N. gonorrhoeae* may be due to the limited sample size. The overall results show that recombination can occur more frequently than mutation for a range of bacteria. Furthermore, purifying selection may act more stringently on recombination events than on point mutations for a range of bacteria.

As mentioned in a previous review [71], the evolution of clonality can be different from that of STs that are distantly related in their evolutionary history. The evolutionary rates within clonality for several bacteria have been calculated, and the results show the evolutionary history within clonal complexes. These kind of STs have less chance of experiencing natural selection, but more chance of reflecting the real evolutionary dynamics than the distantly related STs. Future work on the mechanism of recombination and selection pressure is needed.

4.6 Supplementary material

4.6.1 Recombination and mutation models

For the recombination-related model, two assumptions are made to make the situation mathematically solvable: (1) recombination changes the locus as an entity rather than as a part; and (2) the extra mutation events along with recombination are ignored. If the first assumption is invalid, the ρ/θ ratio will possibly be underestimated. This means recombination only introduces a small part of the DNA sequences from another allele in that locus, and this fragment of DNA sequences only contains a few visible nucleotide differences from the original allele. Our model, based on these two assumptions, tends to overestimate the proportion of SLVs due to mutation only.

For the mutation model, the coalescent theory for the prior distribution (a geometric distribution) [158] is applied to model only the mutation event(s) that occur to generate SLVs. We need all mutation events to happen at the same locus; therefore, the likelihood function is the probability that an SLV occurs based on the given number of nucleotide differences. Then, according to Bayesian theory, the required conditional distribution for only point mutation event(s) is obtained. The obtained conditional distribution of the mutation model lies mainly around small numbers of nucleotide differences. The robustness of the choice of different prior parameters is also tested.

4.7 Estimates for tested bacteria by loci (tables)

Table 4.5: Estimates for each housekeeping locus for *B. cereus* of (i) the probability of an SLV being caused by mutation only (p); (ii) the expected number of mutations for an SLV; (iii) the relative rate of recombination to mutation; and (iv) the % of nucleotide differences of an SLV that were introduced by recombination.

Locus	p	Expected Number Mut.	Relative Rate of Rec	% Differences due to Rec	The number of SLVs
glp	0.27	0.31	2.49	95	38
gmk	0.70	0.86	0.37	88	25
ilv	0.16	0.28	3.83	99	58
pta	0.54	0.61	0.80	81	33
pur	0.29	0.38	2.09	96	54
pyc	0.18	0.30	3.42	98	53
tpi	0.54	0.57	0.83	75	20
Average	0.38	0.47	1.97	90	40.14
Median	0.29	0.38	2.09	95	38

Table 4.6: Estimates for each housekeeping locus for *E. faecium* of (i) the probability of an SLV being caused by mutation only (p); (ii) the expected number of mutations for an SLV; (iii) the relative rate of recombination to mutation; and (iv) the % of nucleotide differences of an SLV that were introduced by recombination.

Locus	p	Expected Number Mut.	Relative Rate of Rec	% Differences due to Rec	The number of SLVs
atpA	0.09	0.28	5.60	97	170
ddl	0.10	0.11	8.59	97	82
gdh	0.37	0.48	1.43	83	49
purK	0.23	0.30	2.86	87	52
gyd	0.50	0.52	0.98	66	23
pstS	0.39	0.49	1.34	91	91
adk	0.54	0.55	0.85	63	14
Average	0.32	0.39	3.09	84	68.71
Median	0.37	0.48	1.43	87	52

Table 4.7: Estimates for each housekeeping locus for *H. influenzae* of (i) the probability of an SLV being caused by mutation only (p); (ii) the expected number of mutations for an SLV; (iii) the relative rate of recombination to mutation; and (iv) the % of nucleotide differences of an SLV that were introduced by recombination.

Locus	p	Expected Number Mut.	Relative Rate of Rec	% Differences due to Rec	The number of SLVs
adk	0.12	0.17	6.14	98	125
atpG	0.23	0.29	2.95	96	67
frdB	0.14	0.16	5.73	98	128
fucK	0.01	0.01	90.20	100	287
mdh	0.11	0.15	7.06	99	208
pgi	0.19	0.30	3.35	98	99
recA	0.16	0.22	4.47	97	63
Average	0.14	0.18	17.13	98	139.57
Median	0.14	0.17	5.73	98	125

Table 4.8: Estimates for each housekeeping locus for *K. pneumoniae* of (i) the probability of an SLV being caused by mutation only (p); (ii) the expected number of mutations for an SLV; (iii) the relative rate of recombination to mutation; and (iv) the % of nucleotide differences of an SLV that were introduced by recombination.

Locus	p	Expected Number Mut.	Relative Rate of Rec	% Differences due to Rec	The number of SLVs
gapA	0.40	0.42	1.46	72	33
infB	0.61	0.69	0.58	56	38
mdh	0.87	0.99	0.14	88	28
pgi	0.41	0.47	1.31	88	37
phoE	0.15	0.19	5.04	98	73
rpoB	0.69	0.75	0.42	43	33
tonB	0.13	0.47	3.29	98	162
Average	0.47	0.57	1.75	78	57.71
Median	0.41	0.47	1.31	88	37

Table 4.9: Estimates for each housekeeping locus for *S. uberis* of (i) the probability of an SLV being caused by mutation only (p); (ii) the expected number of mutations for an SLV; (iii) the relative rate of recombination to mutation; and (iv) the % of nucleotide differences of an SLV that were introduced by recombination.

Locus	p	Expected Number Mut.	Relative Rate of Rec	% Differences due to Rec	The number of SLVs
arcC	0.09	0.10	9.63	96	72
ddl	0.10	0.14	7.51	95	30
gki	0.17	0.20	4.43	91	54
recP	0.10	0.11	8.37	93	39
tdk	0.04	0.07	17.74	99	88
tpi	0.50	0.56	0.94	54	27
yqiL	0.32	0.54	1.56	93	46
Average	0.19	0.25	7.17	89	50.86
Median	0.1	0.14	7.51	93	46

Table 4.10: Estimates for each housekeeping locus for *S. zooepidemicus* of (i) the probability of an SLV being caused by mutation only (p); (ii) the expected number of mutations for an SLV; (iii) the relative rate of recombination to mutation; and (iv) the % of nucleotide differences of an SLV that were introduced by recombination.

Locus	p	Expected Number Mut.	Relative Rate of Rec	% Differences due to Rec	The number of SLVs
arcC	0.34	0.49	1.57	89	12
nrdE	0.26	0.31	2.56	98	20
proS	0.09	0.12	8.56	99	38
spi	0.37	0.41	1.59	94	23
tdk	0.73	0.92	0.31	64	15
tpi	0.25	0.45	2.10	95	10
yqiL	0.47	0.52	1.06	90	30
Average	0.36	0.46	2.54	90	21.14
Median	0.34	0.45	1.59	94	20

Table 4.11: Estimates for each housekeeping locus for *S. aureus* of (i) the probability of an SLV being caused by mutation only (p); (ii) the expected number of mutations for an SLV; (iii) the relative rate of recombination to mutation; and (iv) the % of nucleotide differences of an SLV that were introduced by recombination.

Locus	p	Expected Number Mut.	Relative Rate of Rec	% Differences due to Rec	The number of SLVs
arcC	0.05	0.06	17.59	99	1168
aroE	0.65	1.72	0.25	45	1658
glpF	0.00	0.00	21245.94	100	1113
gmk	0.12	0.13	7.04	96	697
pta	0.04	0.04	23.35	98	996
tpi	0.10	0.12	8.10	96	1035
yqiL	0.11	0.14	7.17	95	1315
Average	0.15	0.31	3044.21	90	1140.29
Median	0.1	0.12	8.1	96	1113

Table 4.12: Estimates for each housekeeping locus for *N. lactamica* of (i) the probability of an SLV being caused by mutation only (p); (ii) the expected number of mutations for an SLV; (iii) the relative rate of recombination to mutation; and (iv) the % of nucleotide differences of an SLV that were introduced by recombination.

Locus	p	Expected Number Mut.	Relative Rate of Rec	% Differences due to Rec	The number of SLVs
abcZ	0.27	0.31	2.21	97	28
adk	0.12	0.14	6.24	99	41
aroE	0.47	0.51	0.96	97	39
fumC	0.04	0.05	21.13	99	45
gdh	0.17	0.20	4.04	99	59
pdhC	0.87	0.87	0.14	52	17
pgm	0.24	0.30	2.33	98	53
Average	0.36	0.46	2.54	92	40.29
Median	0.24	0.3	2.33	98	41

Table 4.13: Estimates for each housekeeping locus for *N. gonorrhoeae* of (i) the probability of an SLV being caused by mutation only (p); (ii) the expected number of mutations for an SLV; (iii) the relative rate of recombination to mutation; and (iv) the % of nucleotide differences of an SLV that were introduced by recombination.

Locus	p	Expected Number Mut.	Relative Rate of Rec	% Differences due to Rec	The number of SLVs
abcZ	0.98	1.04	0.02	31	76
adk	1.00	1.06	0.00	0	5
aroE	0.93	3.31	0.02	27	27
fumC	0.74	0.79	0.33	52	116
gdh	0.88	0.93	0.12	59	120
pdhC	1.00	1.03	0.00	15	46
pgm	1.00	1.06	0.00	10	35
Average	0.93	1.32	0.07	27	60.71
Median	0.98	1.04	0.02	27	46

Table 4.14: Estimates for each housekeeping locus for *N. meningitidis* of (i) the probability of an SLV being caused by mutation only (p); (ii) the expected number of mutations for an SLV; (iii) the relative rate of recombination to mutation; and (iv) the % of nucleotide differences of an SLV that were introduced by recombination.

Locus	p	Expected Number Mut.	Relative Rate of Rec	% Differences due to Rec	The number of SLVs
abcZ	0.05	0.06	17.37	100	4360
adk	0.02	0.03	40.50	100	1970
aroE	0.08	0.37	3.61	99	4827
fumC	0.03	0.04	27.50	100	6053
gdh	0.07	0.08	12.67	99	4484
pdhC	0.05	0.05	20.45	100	4946
pgm	0.06	0.08	13.65	100	5360
Average	0.05	0.10	19.39	100	4571.43
Median	0.05	0.06	17.37	100	4827

Chapter 5

Investigating the impact of geographical isolation on the evolution of *Campylobacter* by comparing New Zealand and United Kingdom datasets

5.1 Summary

Host association, geographical isolation, and agricultural activities may all play a role in the evolution of *Campylobacter*. Phylogenetic and population genetic tools were applied to investigate the effect of geographical isolation on the evolution of *Campylobacter* by comparing the datasets from one historically more isolated country, New Zealand (NZ), to a well-connected country, the United Kingdom (UK). The study is based on 947 *Campylobacter* isolates from the Manawatu, NZ and 1815 *Campylobacter* isolates from Lancashire, UK; both samples were collected over the same time period (2006-2007). The NZ and UK isolates were all sequenced as part of a multilocus sequence typing scheme. There is evidence that geographical isolation affects the evolution and diversity of *Campylobacter* genotypes over short time-scales but that this effect diminishes over longer time-scales. This can be seen by analysing sequence data at different levels of resolution, from the nucleotide level, to the allelic profile level, to the sequence type level. The findings in this chapter also support previous research that suggests that host association and the onset of agricultural activities have played a role in the evolution of *Campylobacter*. Although geographical effects appear to be short-lived, there is evidence of that some

NZ specific and NZ-associated lineages of *Campylobacter*, and that these have distinct evolutionary histories. Our data indicate that some strain types existed in NZ before Polynesian settlement and the introduction of livestock, whereas some strain types diverged after their arrival, and a few of them, such as ST-2381 and ST-474, spread widely in NZ as recently as a few hundred years ago.

5.2 Introduction

The genus *Campylobacter* is a major cause of human gastroenteritis worldwide [265, 399]. Much research has been done on the molecular epidemiology and evolution of *Campylobacter* in different countries, including New Zealand [123, 246, 260, 262]; the United Kingdom [122, 239, 333, 403]; Africa [199, 200]; Finland [193]; Norway [163, 318]; Switzerland [209]; Denmark [232]; and the United States [255, 367]. Geographical isolation can provide a barrier that limits the opportunity for genetic exchange and restricts the diversity of material available for recombination. In this chapter, the impact of geographical isolation means the effect of physical barriers brought about by relative geographical isolation on the evolution of *Campylobacter*. Comparing the evolutionary processes and diversity patterns of *Campylobacter* in a geographically isolated country and a non-geographically isolated country can help to understand the factors that drive the evolution of *Campylobacter*. So far, only limited research has been done in this area [332]. Although the multilocus sequence typing (MLST) *Campylobacter* dataset (PubMLST: http://pubmlst.org/) is widely available and contains isolates from many countries worldwide, studies on the effect of geographical isolation are still difficult to perform due to the problems in accessing equivalent datasets in terms of sampling time, host species, area, and culture methods.

As a multihost species, *Campylobacter* has shown apparent host association [122, 243, 332]. Sheppard [332] found that host association was a more important determinant of variation in genotypes than geographical distance. They focused on the comparison of the effect of host association versus the effect of geographical location, but this viewpoint does not address the role played by geographical isolation in the evolution of *Campylobacter*. The focus of this chapter is to explore the similarities and differences in genotype distributions isolated from multiple hosts in two geographically separated populations. All isolates in this study were sequenced using MLST [74]. The NZ and UK datasets were sampled within the same time period (2006 to 2007), and have equivalent areas, and an equivalent mix of urban and rural populations. This choice of datasets helps to overcome some possible confounding factors, including differences in sampling time and space.

New Zealand has a particularly high incidence of human campylobacteriosis [10, 269], although it is not known why this is so. In New Zealand, the *Campylobacter jejuni* strain type 474 (ST-474) [25, 259] was, until recently, responsible for more than a quarter of the notified human campylobacteriosis cases, and is widely distributed [246]. However, outside NZ, ST-474 has only been found sporadically and infrequently in other countries, such as one isolate reported from a poultry sample in the Czech Republic in 1999 (recorded in the *Campylobacter* PubMLST website) [188] and a human sample in France in 2003 [49]. Further, there are some strain types which have only been recorded in NZ [259, 261].

Compared to the UK, New Zealand is a country that is geographically isolated from much of the rest of the world. NZ is located in the southwestern Pacific Ocean, it has a long history of isolation from other continental landmasses and has no native land mammals except two species of bat. About 80 million years ago, New Zealand began to separate from the ancient supercontinent of Gondwana [57, 389]. The separation continued until about 60 million years ago, when the formation of the Tasman Sea separated New Zealand from Australia [298]. Thousands of years ago, most of NZ was covered by rainforest. The surrounding expanses of seas isolated New Zealand into a distinctive ecological system, which included unique fauna and flora [57, 90]. The geographical isolation of NZ created an opportunity for endemic flightless species of birds, such as the takahe and kiwi to evolve [67, 342]. New Zealand was the last large landmass to be colonized by humans and domestic animals. Approximately 700 to 1000 years ago, Polynesians settled and brought domestic animals with them [40, 164, 244, 245] and, from the 17th century, Europeans arrived and brought some wild and domestic animals [247, 337]. A few hundred years ago, pukeko arrived and colonized NZ [20, 373, 374]. Many animals were imported in 19th century, but relatively few were recently. As to poultry, NZ does not import or export poultry. The uniqueness of the geographical location of NZ and the distinctive history of the introduction of European wildlife and livestock to the country also provide a special opportunity to carry out analysis of the effect caused by geographical isolation on the evolution of *Campylobacter*. The current population status of *Campylobacter* in NZ may also be affected by complex demographic factors, including the domination of endemic wild bird populations.

The similarities and differences between the frequency distributions of sequence types (ST) from different host sources from NZ and the UK were examined. Population differentiation and genetic distance of the isolates were also measured from the different countries and a wide range of hosts by estimating a fixation index (Fst) and constructing phylogenetic networks based on Fst distances for different host sources from both NZ and the UK. The analysis of the sequence data at different levels of

resolution, from the nucleotide level, to the allelic profile level to the ST level, was used to reflect the distribution of *Campylobacter* genotypes at different time scales [98]. The allelic profile consists of seven numbers, and can be considered to be an intermediate stage between the nucleotide and ST levels. Analysis at the ST level places more weight on the most recent genetic changes. This is because a single nucleotide difference will change the ST; that is, it has the same effect as changing many nucleotide bases. Changes at the nucleotide level accumulate through time, and this accumulation can only be reflected at the allelic profile level or the nucleotide level, rather than at the ST level. A rarefaction analysis was applied in order to assess the STs richness for different host sources between the two countries. A Chao1-bc estimator was calculated to show and compare the genetic diversity between the two countries. Lastly, the population structure of isolates from the different geographical locations (NZ or the UK) and different host species were investigated and a unique phylogeny of *Campylobacter* from the NZ dataset was constructed.

5.3 Material and Methods

5.3.1 Isolates

There were 2762 isolates of *Campylobacter jejuni* (*C. jejuni*) and *Campylobacter coli* (*C. coli*) isolated during the same time period (from 2006 to 2007) from both countries available for analyses. The study areas were Manawatu, NZ, and Lancashire, UK [146, 403] (Figure 5.1). A total of 947 NZ isolates were taken from a range of hosts (humans, ruminants, poultry [259, 261], and environmental water). A total of 1815 UK isolates were obtained from human, poultry, and ruminant (cattle and sheep) host sources. All the isolates were genotyped using MLST [74, 237]. Some analyses applied in this chapter only used a subset of these datasets, and are specified in the relevant section.

Several isolates that have, to date, only been recorded from NZ include a *C. jejuni* ST (ST-2381) that has mainly been found in NZ rivers and water rails [39]. It is predominantly associated with native water rails, including the flightless takahe and the pukeko, which was believed to have become established in NZ following its introduction [376, 374]. ST-474 is a strain type commonly found in NZ human cases, but rarely found in other countries in the world. ST-3609 [259] was only identified in 2005 in samples from a NZ poultry supplier [259]. The ruminant-associated strains, ST-3795 and ST-3798, have both, so far, only been found in NZ. Two strain types under investigation are from a proposed new *Campylobacter* species, provisionally named *Campylobacter species nova* (*C.* sp. nov.) [38].

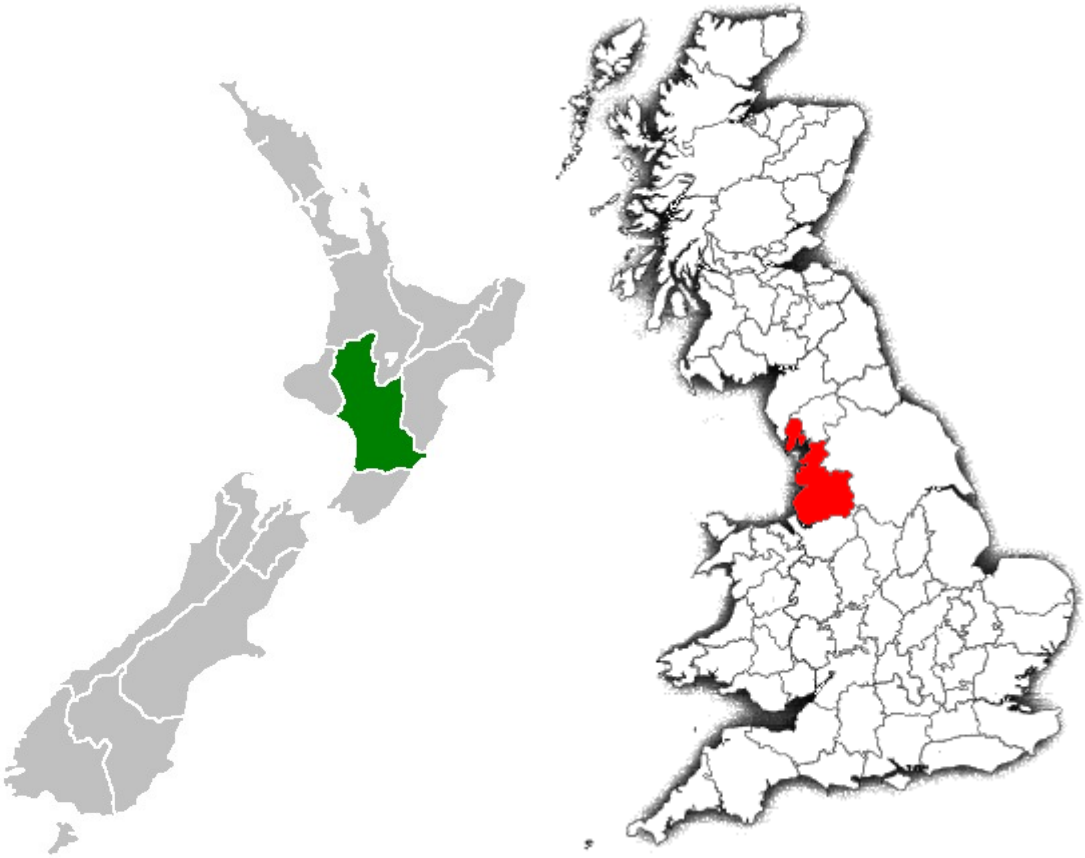


Figure 5.1: Map of NZ with the Manawatu region highlighted (left). Map of the UK with the Lancashire area highlighted (right).

5.3.2 Analysis overview

Population genetic tools, phylogenetic network methods, and molecular evolution tools were applied to investigate the similarity and diversity of *C. jejuni* and *C. coli* from both countries. A χ^2 test was applied to the ST frequency of the NZ and UK datasets to investigate their similarity. The proportional similarity index (PSI) [107, 207, 309, 398] and corresponding confidence intervals were calculated to measure the variation between the frequency distributions of STs among different hosts and among different countries. The relationship between isolates from different geographical locations (NZ or UK) and host species were also investigated by Fst [305, 341]; analysis of molecular variation (AMOVA) [94, 168]; Tajima's D test [361]; the rarefaction technique [157]; and the Chao1-bc estimator [43, 46, 331]. Fst and AMOVA were calculated at three different levels of sequence resolution: the nucleotide level, the allelic profile level, and the ST level [98]. The population structures of the isolates related to geographical isolation and host association were also investigated. A phylogeny of *C. jejuni* and *C. coli* from the NZ dataset was constructed using BEAST [80], and the findings compared with those in a previously published report on a UK dataset [403].

5.3.3 Population genetics tools and network methods

In order to test for similarities in the frequency of STs for NZ and UK datasets, a χ^2 test was applied to test whether or not the underlying distributions of the STs from the two populations are the same. Another measurement of the similarities between the frequency distributions of STs is the PSI, also known as the Bray-Curtis Index or Czekanowski's Quantitative Index [107, 207, 309, 398]. This index is defined by the formula:

$$PSI = 1 - 0.5(\sum_i |p_i - q_i|) = \sum_i min(p_i, q_i), \tag{5.1}$$

in which p_i and q_i are the proportions of ST_i in each population.

The PSI provides a simple estimate of the similarity between the frequency distributions of sequence types from different sources [309], such as the NZ and UK and data sets. If the frequencies in the two populations are identical, PSI equals 1; if these two populations have no STs in common, the PSI equals 0 [238, 281]. This method was used to test the level of similarity between *C. jejuni* and *C. coli* STs from both countries and also for different hosts. A bootstrap method using the Monte Carlo algorithm for case resampling was used to produce confidence intervals for this measure [126, 260]. A population distance matrix was calculated based on

the pairwise 1-PSI values. This was represented as a phylogenetic network using the Neighbor-net method [15, 34].

In population genetics, Fst is widely used as a simple descriptive statistic to measure population differentiation [168]. It is a measurement of genetic diversity (focusing on allele frequencies) within and between populations. Fst values were calculated in Arlequin version 3.5 [93], using concatenated nucleotide sequences for seven loci from the *Campylobacter* PubMLST database (http://pubmlst.org/campylobacter/). Neighbor-joining trees for both Fst and 1-PSI were drawn with SplitsTree4 software [177, 179, 180].

AMOVA was also performed using Arlequin version 3.5 [93]. AMOVA is a method that investigates the genetic structure of populations using the variance of gene frequencies and the estimated number of changes between haplotype sequences [51, 233, 395]. AMOVA is a hierarchical analysis of variance which was used to partition the total variance into a geographically-related genetic structure, a host-associated genetic structure within each country, and the variance related to sequences within host-associated populations. Then a similar analysis was applied to partition the total variance into a host-associated genetic structure, a geographically-related genetic structure within each host source, and the variance related to sequences within each country. The input files were prepared using DnaSP (DNA sequence polymorphism), version 5 [228]. Tajima's D is a test to identify DNA sequences that do not evolve neutrally under the assumption of the neutral theory model of molecular evolution [198, 361], such as under selection pressure and/or experiencing demographic changes. Tajima's D tests were applied using Arlequin version 3.5 [93] to the *Campylobacter* isolates from different host sources in both countries.

Species richness was standardised and compared by the rarefaction technique [157]. Rarefaction curves overcome the problem caused by unequal sample size by resampling given datasets. This technique calculates the species richness for a given number of sampled individuals and includes two steps: 1) random resampling of the pool of N samples multiple times; and 2) plotting of the average number of species found in each sample $(1, 2, ..., N)$ [143]. The rarefaction curve shows the number of unique STs as a function of the number of individuals sampled.

Species diversity was compared using the Chao1-bc estimator [46, 331], as implemented in SPADE (Species Prediction And Diversity Estimation) [45]. Chao1-bc is a bias-corrected version for the Chao1 [44]. Chao1 estimates the number of missing strains from the numbers of strains with low frequency counts, such as singletons and doubletons [43, 46, 331]. For Chao1 estimator, a higher value means larger diversity. These statistics are calculated for eight sets of data for *C. jejuni* and *C.*

coli isolates from a range of host sources: NZ human (NZH); UK human (UKH); NZ poultry (NZP); UK poultry (UKP); NZ ruminant (NZR); and UK ruminant (UKR) over the same time period (2006-2007). The NZ dataset combines all the data from NZ, including NZ human (NZH), NZ poultry (NZP), and NZ ruminant (NZR), while the UK dataset combines all the data from the UK, including UK human (UKH), UK poultry (UKP), and UK ruminant (UKR).

5.3.4 Bayesian Phylogenetic analysis

A BEAST analysis [80] was applied to construct the *Campylobacter* phylogeny, focusing on some specific strains of *C. jejuni* and *C. coli* (ST-2678, ST-45, ST-403, ST-3798, ST-3795, ST-2026, ST-48, ST-3309, ST-474, ST-21, ST-2381, ST-1132, ST-854, ST-3323, and ST-3310). Other species (*C. fetus* ST-4, *C. helveticus* ST-2, *C. insulaenigrae* ST-12, *C. lari* ST-6, and *C. upsaliensis* ST-25) from the *Campylobacter* genus and a newly proposed species, *C.* sp. nov. [38], were also analysed. These species have been characterised by MLST schemes [74, 256, 381]. Seven loci were used for each MLST scheme, and there were four shared loci for all these eight species tested in the *Campylobacter* genus: *gln*A, *gly*A, *tkt*, and *unc*A (also known as *atp*A).

A Sawyer's runs test [78, 323] was applied to test that the alleles are all non recombinants within a set of allele alignments by loci using START2 [187]. Nielsen and Yang's codon substitution model [274] was then applied in BEAST [80]. The code for BEAST analysis was modified from Dr. Daniel Wilson's BEAST analysis [334, 403] by adding more sequences. Speciation events were modelled by the Yule process [415], with a Jeffreys prior for the speciation rate. This prior choice is consistent with the work of Dr. Daniel Wilson et al. [403], the rationale for which is that the Yule process and a Jeffreys prior represent the most simple model. This choice will generate a random branching process [195]. The Markov Chain Monte Carlo (MCMC) chain was run for 10,000,000 iterations, and parameters were stored every 1000 iterations. Two chains were run, and convergence was compared and visually checked. The effective sample size for posterior distribution was 407; the edge lengths in the BEAST trees were scaled according to time.

5.4 Results

The similarity of strains in two distant countries, NZ and the UK, and several different hosts (human, poultry and ruminant) were examined. There is a significant difference between the ST frequencies in NZ and the UK (χ^2 test, $p<0.01$). In order

to further examine similarity in the distribution of STs in the two countries and different hosts, the Proportional Similarity Index (PSI) was calculated. The matrix of 1-PSI therefore demonstrates the dissimilarity of the ST frequencies observed in both different host sources and/or geographical locations.

A neighbor-net based on the 1-PSI distances (Figure 5.2) shows evidence of variation between the two countries. There is also evidence of variation between different host sources: isolates from poultry are grouped together, as are isolates from ruminants. Figure 5.2 shows that for both countries, gene flow within the ruminant *Campylobacter* population (sheep and cattle) is larger than that between isolates from ruminant host sources and other host sources. In addition, for both countries, isolates from poultry always have a greater similarity with isolates identified in human cases than isolates from other groups. For both countries, isolates from ruminant (sheep and cattle) always share a greater similarity than isolates from other groups (PSI value of isolates identified between in sheep and cattle: 0.50, 95% CI [0.36, 0.55] for NZ; 0.34, 95% CI [0.27, 0.39] for the UK). In addition, isolates identified in sheep share greater similarities than those from cattle for the two countries.

5.4.1 Fst and AMOVA at different levels

When Fst and AMOVA analyses are performed at the ST level, they emphasize recent events for closely related strains. For example, a single event (recombination or mutation) will change the ST [414]. In contrast, analyses at the nucleotide level can reflect genetic changes further back in time. Figure 5.3 shows Neighbor-net plots for Fst at different levels, from the ST level, to allelic profile, to the nucleotide level. It can be seen that there is some evidence of clustering by geographical region at the ST and allelic profile levels (Figure 5.3 top and middle), however at the nucleotide level there is relatively stronger clustering by host association (Figure 5.3, bottom).

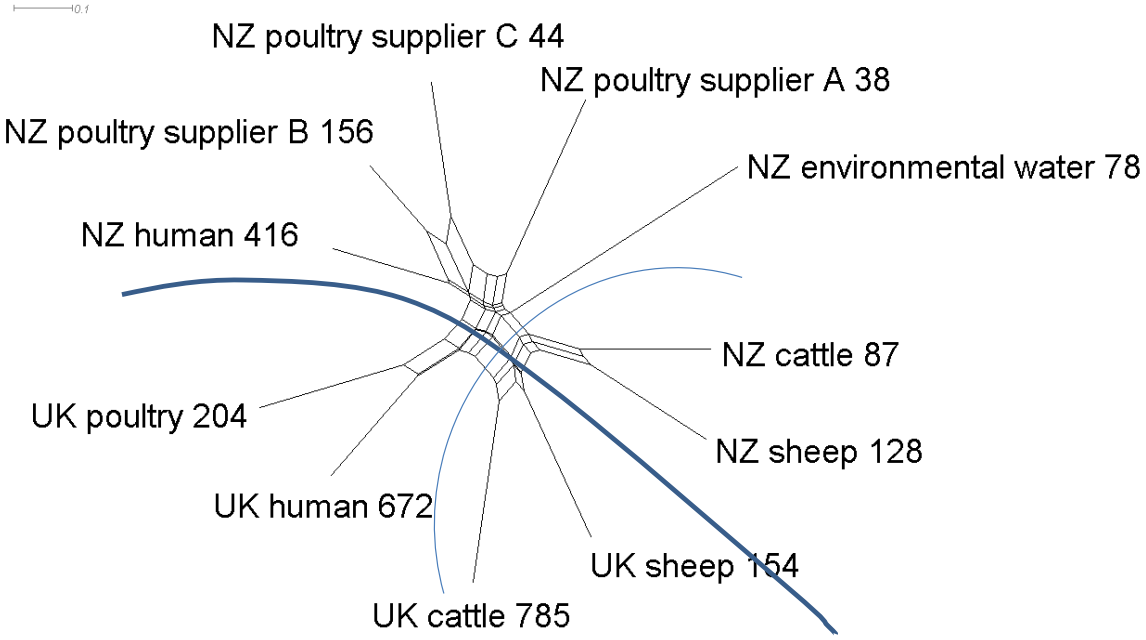


Figure 5.2: Neighbor Net of 1-PSI matrix for ST type frequency. These STs are grouped by host species and sampling countries. The number after each host source represents the sample size. The thick line shows the split between the two geographical locations, and the thin line separates the *Campylobacter* isolates from ruminant hosts and the isolates from human and poultry host sources.

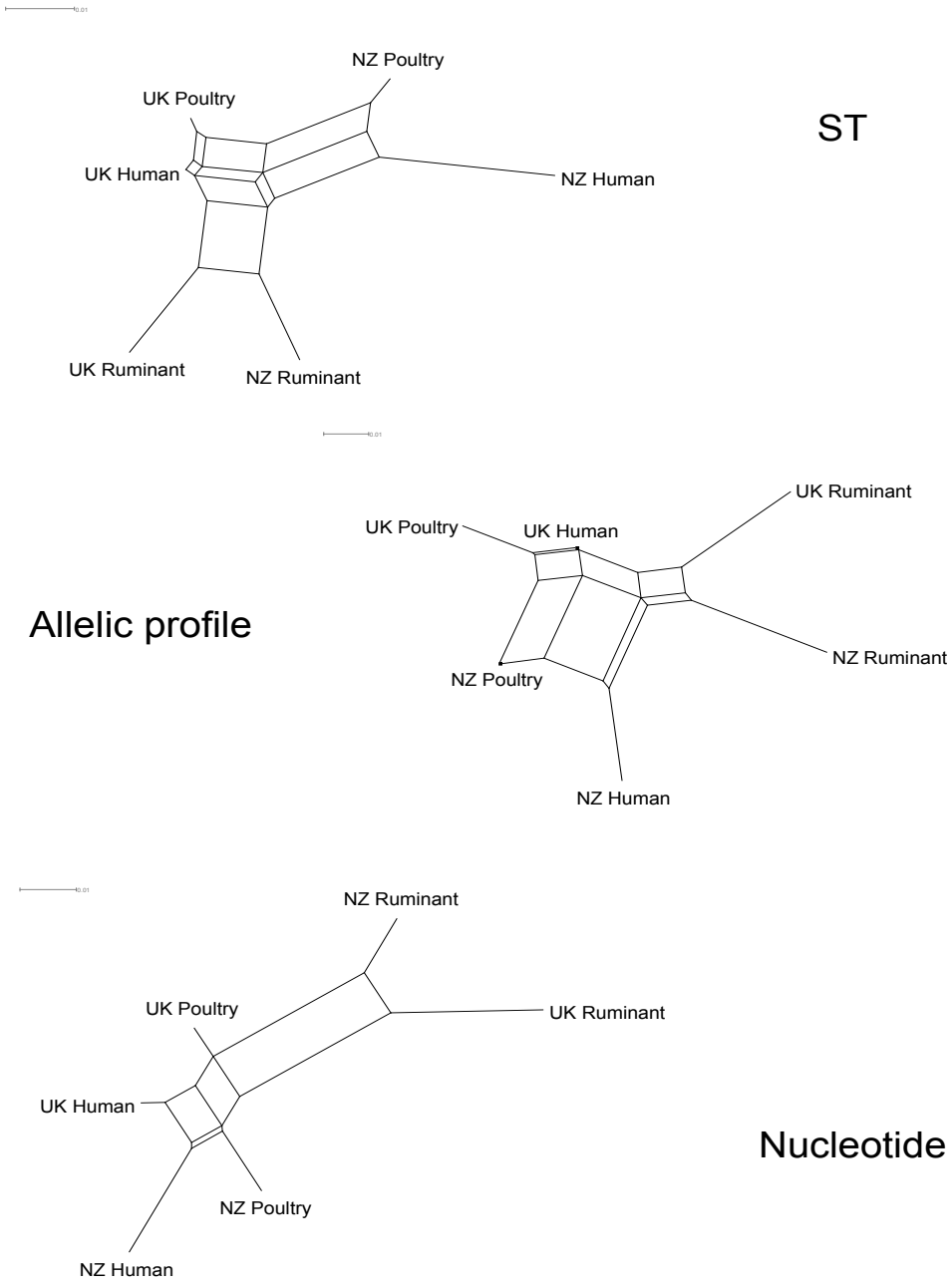


Figure 5.3: Neighbor-Net plot of pairwise Fst values at different levels from ST level, to allelic profile level, to nucleotide level showing the pattern of gene flow between a variety of host sources and geographical locations.

For the AMOVA analyses, isolates from NZ and the UK were separated into two groups and, within each, the isolates were divided by different host sources: ruminant, poultry, and human. Table 5.1 shows the results when AMOVA is used with countries defined as the higher grouping. At the ST level, 1.77% of the variation of genetic structuring in haplotype sequence is attributed to the country level (AC), 3.81% of the variation of genetic structuring in haplotype sequence is attributed to host species within countries (among host species within countries (AHC)), and the rest of the variation (94.42%) is assigned to within-population (within hosts within countries (WH)). However, at the nucleotide level, the AMOVA results show that most variation (about 93%) of genetic structuring is attributed to variation within hosts within countries, less (about 7%) was attributed to host species, and none attributed to geographical isolation. Results of AMOVA at the allelic profile level are located between the results from ST and the nucleotide levels. The results show that 1.77% of variation due to countries (AC) found at the ST level was greater than that at allelic level (1.39%), and at the nucleotide level no variation as assigned to country. This is important because it demonstrates the effect of geographical isolation diminishes from the ST level to the nucleotide level. The results also show that variation due to different host sources within countries (AHC) increased from the ST level (3.81%), to the allelic level (6.02%), and to the nucleotide level (7.23%). This demonstrates the effect of host association increased from the ST level to the nucleotide level.

Table 5.1 contains the AMOVA analyses when the host grouping was set at the higher level. Table 5.1 shows there are 4.76%, 5.96% and 3.68% of the variation of genetic structuring attributed to the country level at all three levels. This shows that geographical isolation plays a role in the variation of distribution of genotypes of *Campylobacter* population. From the ST level to nucleotide level, the variation assigned to host level is increased from 0.11% to 3.72%. This is the same with the results (AHC) in Table 5.1, which means that host association increased from the ST level to the nucleotide level.

Table 5.1: AMOVA with country defined as higher grouping: from the nucleotide level, to the allelic profile level, to the sequence type level (AC means among countries, AHC represents among hosts within countries, WH represents within hosts). Negative values mean there is no variance contributed [93].

Source of variation	d.f.	Sum of squares	Variance components	Percentage of variation
ST level				
AC	1	19.39	0.01	1.77
AHC	4	32.17	0.02	3.81
WH	2678	1248.88	0.47	94.42
Total	2683	1300.43	0.49	
Allelic level				
AC	1	133.54	0.04	1.39
AHC	4	299.13	0.18	6.02
WH	2678	7372.03	2.75	92.59
Total	2683	7804.71	2.93	
Nucleotide level				
AC	1	713.32	-0.24	-0.84
AHC	4	3514.27	2.12	7.23
WH	2678	73340.26	27.39	93.6
Total	2683	77567.85	29.26	

Table 5.2: AMOVA with host as higher grouping: from the nucleotide level, to the allelic profile level, to the sequence type level. (AH means among hosts, ACH represents among countries within hosts, WC represents within countries.)

Source of variation	d.f.	Sum of squares	Variance components	Percentage of variation
ST level				
AH	2	24.86	0	0.11
ACH	3	26.7	0.02	4.76
WC	2678	1248.88	0.47	95.13
Total	2683	1300.43	0.49	
Allelic level				
AH	2	233.21	0.03	1.08
ACH	3	199.47	0.18	5.96
WC	2678	7372.03	2.75	92.96
Total	2683	7804.71	2.96	
Nucleotide level				
AH	2	2967.37	1.10	3.72
ACH	3	1260.22	1.09	3.68
WC	2678	73340.26	27.39	92.6
Total	2683	77567.85	29.57	

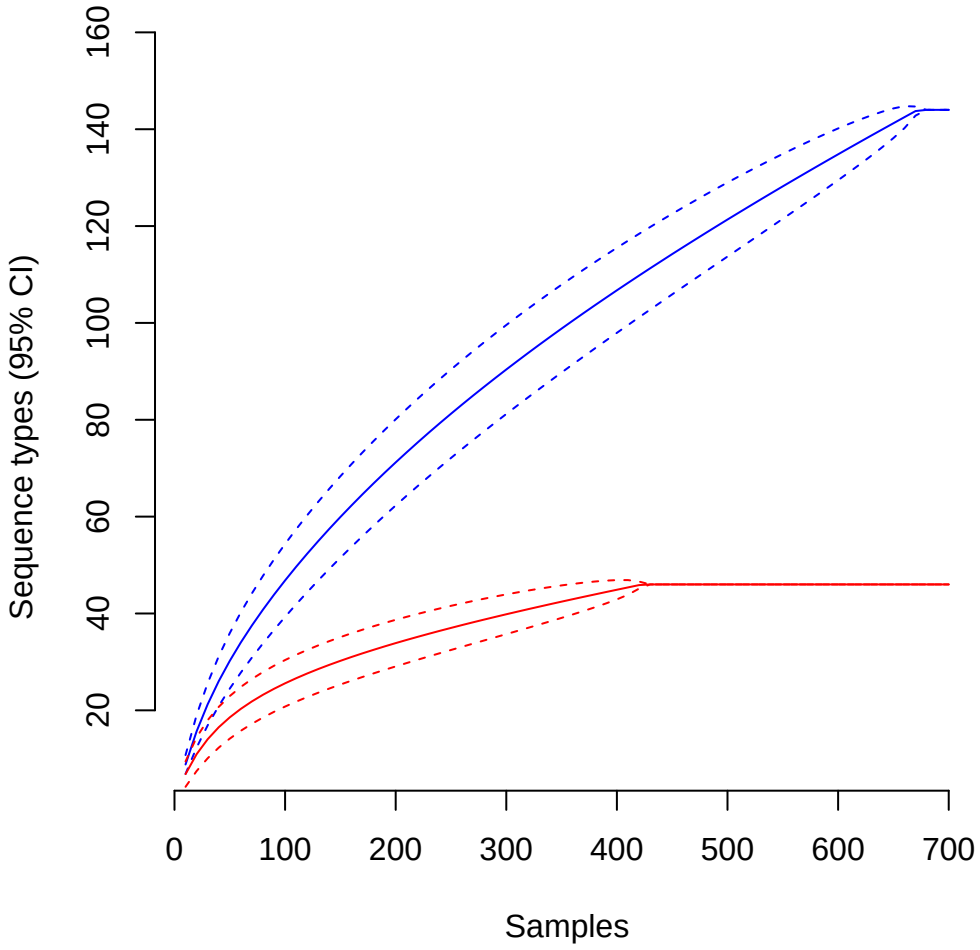


Figure 5.4: Rarefaction plot for UK and NZ data on human host source (2006–2007). The lower line represents NZ human data from 2006–2007; 416 isolates in total, in which the number of distinct STs is 42. The upper line represents UK human data over the same time period. The UK dataset has 672 isolates, of which the number of distinct STs is 144. If a vertical line was drawn on the 200th sample (on the x-axis), there will be over 30 distinct STs for the NZ dataset and over 70 distinct STs for the UK dataset.

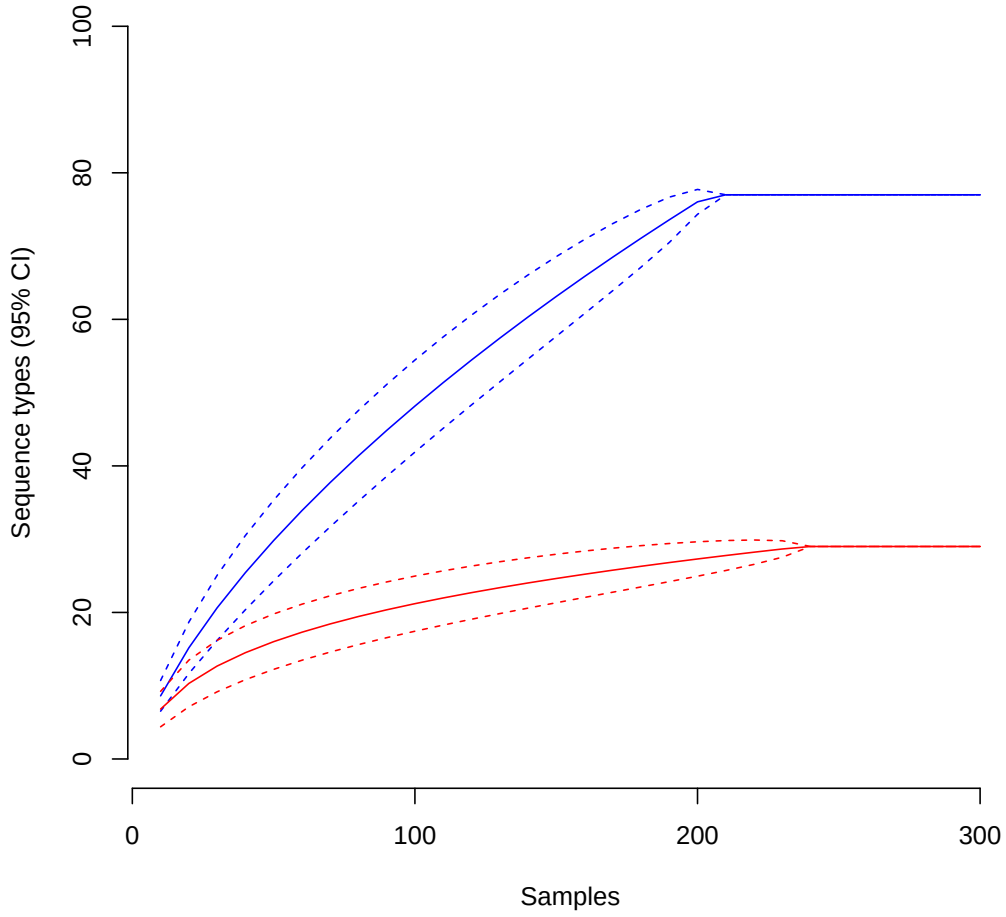


Figure 5.5: Rarefaction plot for UK and NZ data on poultry host source (2006–2007). The lower line represents NZ poultry *Campylobacter* data from 2006–2007; 238 isolates in total, in which the number of distinct STs is 29. The upper line represents UK poultry *Campylobacter* data over the same time period. The UK dataset has 204 isolates, of which the number of distinct STs is 77. If a vertical line was drawn on the 200th sample (on the x-axis), there will be over 20 distinct STs for the NZ dataset and over 70 distinct STs for the UK dataset.

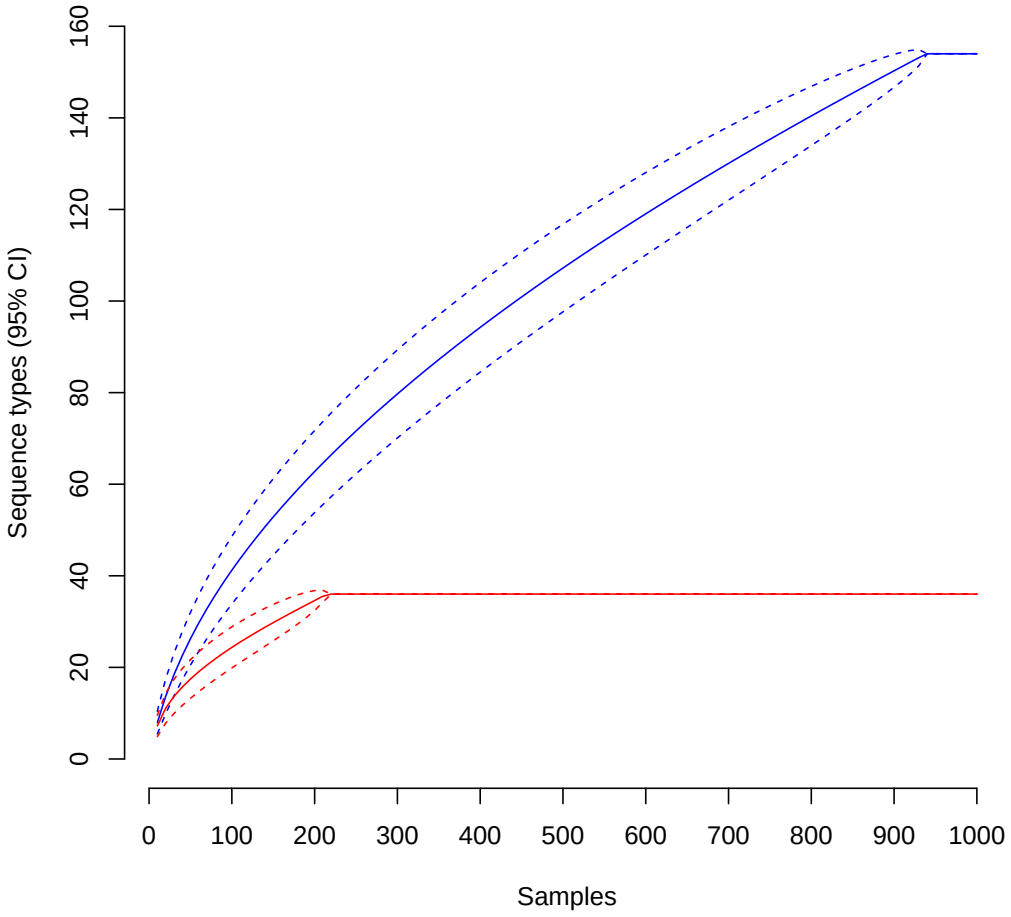


Figure 5.6: Rarefaction plot for UK and NZ data on ruminant host source (2006–2007). The lower line represents NZ ruminant *Campylobacter* data from 2006–2007; 215 isolates in total, in which the number of distinct STs is 36. The upper line represents UK ruminant *Campylobacter* data over the same time period. The UK dataset has 939 isolates, of which the number of distinct STs is 154. If a vertical line was drawn on the 200th sample (on the x-axis), there will be over 30 distinct STs for the NZ dataset and over 50 distinct STs for the UK dataset.

Tajima's D test only shows significant population expansion in size and/or directed selection in *Campylobacter* isolates (-1.34, p-value 0.04) from UK human cases. Rarefaction analysis was applied to evaluate the ST richness of *Campylobacter* that cause human infection. Figure 5.4 reveals the higher diversity of UK human infection isolates than that of NZ isolates. Similarly, Figure 5.5 and Figure 5.6 reveal the higher diversity of UK poultry and ruminant *Campylobacter* population (sheep and cattle) than that of NZ isolates.

Diversity (as measured by Chao1-bc statistics) for the UK dataset, and three host-specific UK datasets (UK human, UK poultry, and UK ruminant) are all larger than the comparable datasets from NZ. For the combined UK dataset, the Chao1-bc is larger than that of the combined NZ dataset. For isolates from human host sources (UK: 883.2, 95% CI: 673.3, 1212.3; NZ: 235.5, 95% CI: 129.6, 526.5), Chao1-bc for isolates from the UK human host is larger than that from the NZ human host (UK: 361.9, 95% CI: 259.2, 555.9; NZ: 72.0, 95% CI: 50.7, 145.2). For isolates from poultry host sources, Chao1-bc for isolates from the UK poultry host is larger than that of the NZ poultry host (UK: 167.5, 95% CI: 119.0, 271.8; NZ: 38.0, 95% CI: 31.1, 67.0). For isolates from ruminant host sources, Chao1-bc for isolates from the UK ruminant host is larger than that from the NZ ruminant host (UK: 371.6, 95% CI: 271.9, 555.3; NZ: 131.0, 95% CI: 64.1, 357.1). Only the confidence intervals for isolates from the ruminant host sources overlap.

5.4.2 BEAST analysis

The models used by BEAST assume that sequences have not undergone recombination. The Sawyer's runs test did not provide evidence to reject the null hypothesis of no recombination among alleles by the loci under investigation (*gln*A, *gly*A, *tkt*, and *unc*A). From the BEAST analysis (Table 5.3 and Figure 5.7), the estimated time of divergence of *C. jejuni* and *C. coli* is consistent with previous research [334, 403]; around 6000 years ago (5837.8, 95% HPD: 2730.1, 9719.9). HPD stands for highest posterior density and is similar to the credible interval in Bayesian theory. It can be interpreted as the most compact interval in parameter space that contains 95% of the posterior probability [80]. The BEAST results in Table 5.3 and Figure 5.7 show that the proposed new species *Campylobacter species nova* (*C.* sp. nov.) separated before the estimated time of divergence between *C. jejuni* and *C. coli*. The estimated time of divergence for *C.* sp. nov. is 6693 years ago (95% HPD: 3084.2 to 11020.3).

C. jejuni strains (ST-2768, ST-45, ST-403, ST-3798, ST-2026, ST-3795, ST-48, ST-3609, ST-474, ST-21, and ST-2381) are grouped into two different branches of

Table 5.3: BEAST results of the mean of split time [334]

Splits Name	Time (unit year)	95% HPD lower, upper range
fetus - *jejuni*	32616.9	15494.5, 55097.6
lari - *jejuni*	19587.7	9766.0, 33755.9
helveticus - *jejuni*	14716.2	6805.0, 24582.6
insulaenigrae - *lari*	7266.7	3344.8, 12257.2
sp. nov. - *jejuni*	6693	3084.2, 11020.3
jejuni - *coli*	5837.8	2730.1, 9719.9
helveticus - *upsaliensis*	4563.5	2072.0, 7596.1
coli splits	1698.8	830.2, 2837.7
coli ST3323-ST3310	1698.8	830.2, 2837.7
coli ST3323-ST854	1168.1	520.1, 1835.2
sp. nov. splits	1168.1	586.3, 1984.0
jejuni splits	538.4	251.0, 904.4
jejuni ST2381-ST2768	533.1	251.0, 903.4
jejuni ST48-ST2381	507.5	206.1, 843.9
jejuni ST2678-ST45	456.1	178.8, 783.9
jejuni ST474-ST21	243.5	108.1, 422.7
jejuni ST3795-ST45	239.6	107.0, 422.7
jejuni ST48-ST474	118.5	45.5, 220.7
jejuni ST3798-ST3795	65.5	20.9, 123.7
coli ST1132-ST854	38.2	7.9, 75.2
jejuni ST3795-ST2026	12.9	0.2, 33.5
jejuni ST3798-ST403	10.5	0.2, 26.4
jejuni *S*T48-ST3609	8.1	0.2, 21.5

the topology (Figure 5.7). The estimated divergence of the two groups is 533.1 years ago (95% HPD: 251.0 to 903.4). One contains ST-2768, ST-45, ST-403, ST-3798, ST-2026, and ST-3795, the other contains ST-48, ST-3609, ST-474, ST-21, and ST-2381. All analysed strains which belong to clonal complex ST-403 (CC 403), including ST-403, ST-2026, ST-3798, and ST-3795, are grouped together. All analysed strains which belong to CC 48, including ST-48, ST-474, and ST-3609, are grouped together.

There are some strains, such as ST-3798, ST-3795, ST-2381, and ST-474, which are highly prevalent in NZ and rarely found elsewhere in the world. ST-3798 and ST-3795 diverged from the ST-45 branch 239.6 years ago (95% HPD: 107.0 to 422.7). The estimated time of divergence of NZ isolates, such as *C. jejuni* ST-2381 and *C. jejuni* ST-21 is 507.5 years ago (95% HPD: 206.1 to 843.9). For some more recent examples, ST-474 also diverged from the ST-21 branch 243.5 years ago (95% HPD: 108.1 to 422.7). The estimated divergence of *C. jejuni* ST-474 and *C. jejuni* ST-48 is 118.5 years ago (95% HPD: 45.5 to 220.7), and *C. jejuni* ST-3609 and *C. jejuni* ST-48 is estimated to be 8.1 years ago (95% HPD: 0.2 to 21.5).

5.5 Discussion

When clustering strains based on distances that give high weight to slight differences (Figure 5.3, top) there is some evidence of clustering by geographical region. There is a clear separation between STs from NZ and the UK. However for distance measures that take progressively more detailed sequence information into account (Figure 5.3 bottom) there is no apparent clustering by geographical region and clustering by host appears to be apparent. Fearnhead [2007] demonstrates that this may reflect structure at different time-scales which suggests that for the *Campylobacter* data the effect of geographical structure may be short-lived. The effect brought about by geographical isolation then plays a short-lived role in the evolution and diversity of *Campylobacter* genotypes. This observation indicates that geographical isolation reduces the chance of genetic exchange between *Campylobacter* strains, but that it has not built up a biological barrier for them.

For the nucleotide level analysis, host association has a more apparent effect than country association which can be seen from the different grouping patterns and branch lengths for geographical isolation and host association in Figure 5.3. Furthermore, at the nucleotide level, the AMOVA result (Tables 5.1 and 5.1) shows there is no variance attributed to the country level. This again supports the conclusion that geographical isolation has little effect on the long time-scale evolution of *Campylobacter* strains, a result consistent with Sheppard et al. [2010a]. An alternative explanation of the fact that host association has more impact than geography may be that *Campylobacter* adapts to its host, and this produces the unequal distribution of STs among hosts that we observe. However, there is some evidence of the effect of country geographical separation at the ST level. As was shown in Figure 5.3, at the ST level, the geographical effect is much stronger than at the other two levels, and is also supported by the observation that the branch lengths for the host association are shorter than those at the nucleotide and allelic profile levels. Both observations are consistent with Fearnhead [2007] which indicates that because spatial structure has more effect on recent coalescent events, it is often more noticeable at the ST level or allele level than the nucleotide level which is affected most by the timings of older coalescent events.

For the short time period, the Neighbor-net plot for Fst shows evidence for geographical clusters. Moreover, the PSI results show that the geographical isolation effect is more important (or of equivalent importance) than host sources for the evolution at the ST level. PSI and 1-PSI were calculated and 1-PSI was plotted by the Neighbor-net method (Figure 5.2) to measure similarities/differences between the distribution of sample STs from different host sources and countries. Higher PSI values were observed within each country and also between the two countries, a

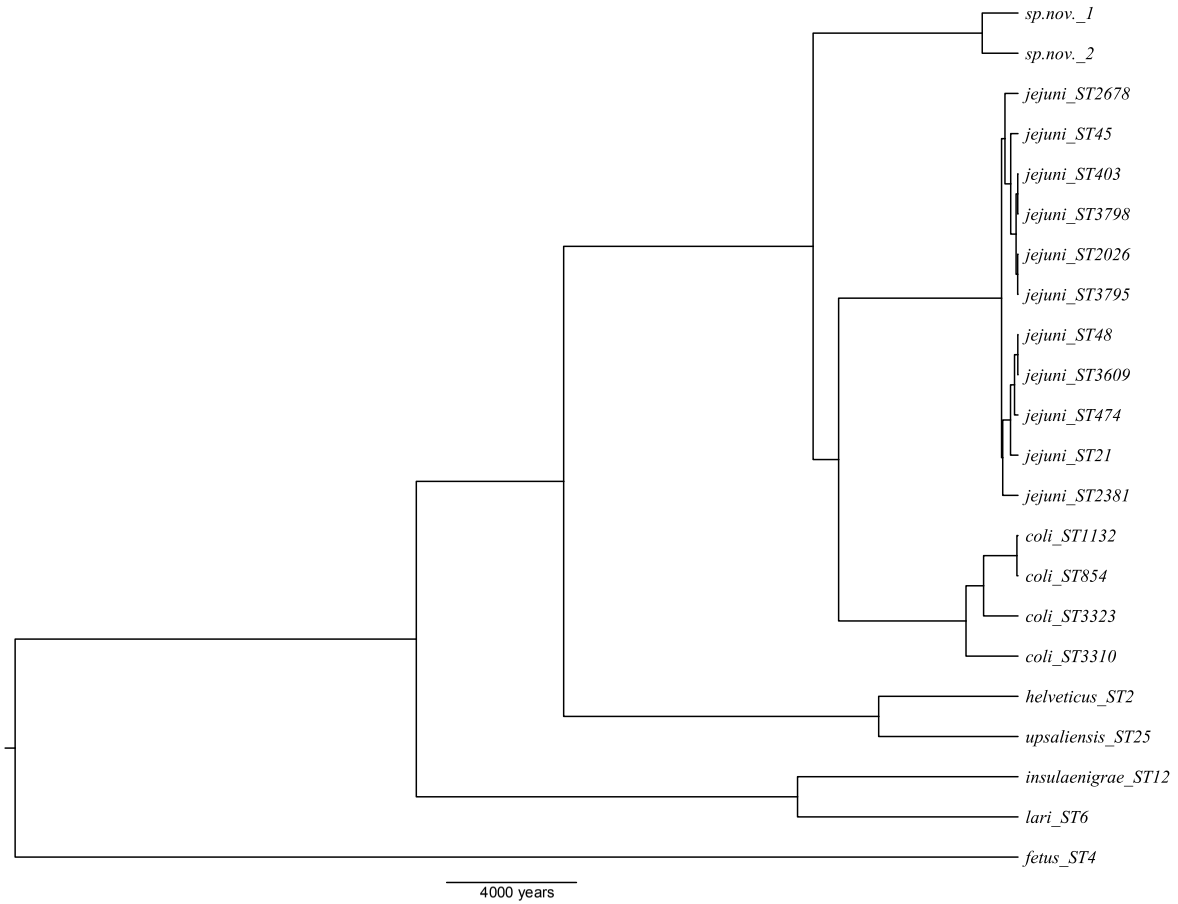


Figure 5.7: Reconstruction of the phylogeny of some NZ specific strains. For the NZ-specific isolates, ST-2381 is mainly found in NZ rivers and water rails; it is mainly attributed to the flightless native birds, including the rare takahe and the common pukeko. ST-474 is a strain type commonly found in NZ human cases, but rarely found in other countries in the world. ST-3609, which shares five loci with ST-474, has not been found since 2005. ST-3795 and ST-3798 are found only in NZ ruminant host sources. Both, however, are from a lineage associated with isolates from UK ruminant hosts. The scale bar represents 4000 years.

finding that is supported by the observation of the Fst cluster between different host sources and countries (Figure 5.2). This result is consistent with the existing effect of geographical isolation on the evolution and diversity of *Campylobacter* genotypes.

Different results from Tajima's D tests for isolates sampled from the same host species but different countries imply different evolutionary histories of population size (or selection pressure). For example, for *Campylobacter* isolates from human infection cases, in the UK, Tajima's D values supported the hypothesis that the population size of *Campylobacter* increased. In contrast, no significant increase in population size was observed in NZ. Rarefaction curves also demonstrate apparent differences between the species richness in NZ and the UK. The richer diversity of host animals in the UK may explain the larger species richness of *Campylobacter* in the UK compared to NZ. The effect caused by geographical isolation is also supported by the rarefaction analyses (Figures 5.4, 5.5, and 5.6) which show there is more diversity in the *Campylobacter* genotype from the UK dataset than the NZ dataset. Chao1-bc estimators also show that, in general, there is more diversity in the genotypes of *Campylobacter* from the UK dataset than from the NZ dataset, for different host sources (human, poultry, and ruminant) and also for the combined datasets. This diversity in the UK dataset may indicate the results of more frequent gene exchange than the geographically isolated country, NZ. This diversity in the UK dataset also demonstrates that the human population in UK are exposed to a more diverse range of *Campylobacter* genotypes. Furthermore, the lower diversity of NZ datasets is likely to be the results of geographical isolation prevents some STs from reaching NZ for at least short time periods (i.e. only a subset of the global population of STs have been introduced into NZ, and these formed 'founder' for subsequence evolution).

Reconstruction of the phylogeny of NZ-specific *Campylobacter* STs reveals that different lineages of *Campylobacter*. Some lineages, such as ST-2381 and *Campylobacter* sp. nov. strains, pre-existed in NZ because they diverged earlier than other STs sampled from other countries in the world. Other lineages, such as ST-3798 and ST-2026, may have evolved in NZ (after being introduced into NZ) because they diverged after several commonly found STs. Furthermore, some strains, such as ST-3798, ST-3795, ST-3609 and ST-474, diverged later than the globally distributed STs and so far with the exception of ST-474, have only reported in NZ not elsewhere in the world. Different evolutionary patterns for different lineages demonstrate different time-scale for NZ strains because these share many STs with the global gene pool. However, NZ also has its own gene pool, which includes strains that diverged before and after the introduction of common STs.

The time when NZ unique strain types/isolates *C.* sp. nov. diverged from the

common ancestors of *C. jejuni* and *C. coli* is estimated to be 6700 (95% HPD: 3080 to 11020) years ago. This estimated time is long before the time when common livestock associated STs were likely to have been introduced into NZ. The estimated time for divergence of ST-2381 and ST-21 is around 508 (95% HPD: 206 to 844) years ago. This is consistent with the arrival time of pukeko from Australia (about a few hundred years ago) [254, 373, 375]. These results show that some NZ isolates, such as *C.* sp. nov. and ST-2381, were likely to have diverged earlier than the globally distributed STs (like ST-21, ST-45, and ST-48). The host species for both *C.* sp. nov. and ST-2381 are only found in NZ, and the closely related purple swamphen found in Australia. This indicate that both geographical isolation and host association played a role in the distribution of *Campylobacter* genotypes.

The estimated time for divergence of ST-474 and ST-48 is around 119 (95% HPD: 46 to 221) years. In NZ, ST-474 has been responsible for 25% of the notified human campylobacteriosis cases, and is widely distributed around NZ [246]. However, ST-474 has rarely been reported elsewhere around the world. The highly specific geographical distribution of these recent diverged strains clearly support the role played by geographical isolation. The estimated time of divergence for both ST-2381 and ST-474 occurred after the introduction of the common STs (ST-21 and ST-48). This is also consistent with the arrival time of pukeko from Australia.

In contrast, the estimated time of divergence of ST-3609 from a worldwide ST-48 is estimated to be less than 10 years. This recent evolution is consistent with the results from laboratory work: ST-3609 was identified in one of the NZ poultry suppliers only at the beginning of the present project (in 2005); it has not been found since then. ST-3609 is a single locus variant of ST-48. During 2006 to 2008 in NZ, several voluntary and regulatory interventions were introduced to control foodborne pathways of campylobacteriosis, particularly in poultry industry [13, 260, 328]. Following that, there was a marked decline in the incidence of foodborne campylobacteriosis [328]. These observations show that human activity can affect the diversity of *Campylobacter*.

The phylogeny of *Campylobacter* estimated in this study is consistent with Wilson et al.'s [2009] research using BEAST. The estimated time for divergence of *C. jejuni* and *C. coli* (about 6000 years ago) is consistent with the most recent dating [334], and suggests that human agricultural activity had an effect on the evolution of *Campylobacter*. The estimated time of divergence of some NZ-associated isolates, such as *C. jejuni* ST-474 and *C. coli* ST-3310 is about 5840 (95% HPD: 2730 to 9720) years ago, which is long before the beginning of Polynesian settlement (around 1000 years ago) [245] and the introduction of livestock (200 years ago) [54]. This again could be because geographical isolation only plays a short-lived role in the evolution

of *Campylobacter*. For long time periods, evolution can reduce the signal caused by geographical isolation. The similarity of the distribution of some common strains, such as ST-45, ST-48, and ST-21, supports the view that globalisation (tourism or international imports and exports) also reduces the effects of geographical factors on the evolution of *Campylobacter*, especially over long timescales.

From Figure 5.7, we can also see host association within different lineages. The proposed species (*C.* sp. nov.) is mainly found in NZ water and water rails and appears to be genetically distinct from the common ancestors of *C. jejuni* and *C. coli*. ST-45 and ST-21 are largely found in a wide range of host sources, and both of them are considered as founder strains and diverged earlier than other strains in the two groups [55, 122, 193, 287, 345] (Figure 5.7). The ST-45 group includes ST-45, ST-403, ST-3795, and ST-3798. ST-45 can be found in a wide range of animal sources, including human and ruminant [220, 221]; ST-403 also has a wide range of host sources, but mainly excludes poultry [188, 221, 333]. At the time of writing, ST-3798 and ST-3795 have only been isolated from a NZ ruminant host source. So far, for the ST-21 group, ST-474 has predominantly been found in human and poultry samples, and ST-3609 in a poultry host source.

Whole genome analysis on both worldwide and NZ specific strains can provide more information on the effect of geographical isolation on the evolution of *Campylobacter*. Three possible mechanisms for the genotype patterns of *Campylobacter* have been identified [334]: physical barriers, biological mechanisms, and selection pressure. Globalisation tends to break the physical geographical barriers for gene exchange, but is a more recent activity, compared with the evolutionary timescale for *Campylobacter*. Because NZ's unique geographically isolated location has provided at least a short-lived physical barrier for the exchange of gene flow over the centuries, a few (if any) specific genetic markers can be identified on the whole *Campylobacter* genome of distinct NZ strains. Based on the seven housekeeping genes, a certain pattern for geographical isolation and host association has been observed. Increased globalisation, tourism and international trade are likely to introduce new strains of *Campylobacter* bacteria and other bacteria to NZ. Whole genome analysis can help us better understand geographical factors, such as whether or not some of the current variation in the diversity distribution of certain *Campylobacter* STs arises from geographical isolation.

This is one of the first attempts to investigate the effect of geographical isolation on the evolution of a specific bacteria, and to enhance our understanding of the effect of ecological barriers on the evolution of *Campylobacter*. There are two main findings in this study: a short-lived role of geographical location in the evolution of *Campylobacter* and the existence of some unique strain types in NZ. These findings

provide a better understanding of the physical barrier for gene exchanges caused by geographical isolation, which may be helpful in disease control and intervention of campylobacteriosis and even other infectious disease, and also useful in determining the origins of *Campylobacter* species.

Acknowledgements

I acknowledge Paul Fearnhead, Barbara R. Holland, Patrick Biggs, and Nigel French for all the guidance and suggestions and Dai Grove-White and Phil Carter for the datasets. We acknowledge the Marsden Fund project 08-MAU-099 (Cows, starlings and *Campylobacter* in New Zealand: unifying phylogeny, genealogy and epidemiology to gain insight into pathogen evolution) for funding this project. BRH acknowledges the Australian Research Council (grant FT100100031).

5.6 Additional structure analysis

5.6.1 Bayesian cluster analysis

In order to investigate the relationship between *Campylobacter jejuni* isolates from different geographical locations (NZ or UK) and different host species, a model-based clustering method was performed to infer population structure for MLST data using Structure 2.3.3 [96, 97, 169, 303]. There are four sets of data for *Campylobacter* isolates host sources: NZ poultry (NZP), NZ ruminant (NZR), UK poultry (UKP), and UK ruminants (UKR). Host sources and geographical locations were pre-specified for three sets of individuals, then the origin for the rest of the individuals was estimated. USEPOPINFO model, PopFlag, and PopData options were applied in Structure 2.3.3 [169, 303]. The no-admixture model was employed with 10,000 iterations after 5000 burn-ins. The convergence was checked using visual plots and the comparison of two chains.

This part of the analysis is looking at *Campylobacter jejuni* isolates from NZ and UK over the same time period (2006-2007). The analysis was performed to allocate individual isolates to different combinations of known geographical locations and host species based on the different genotype frequencies of each combination. The analyses were carried out at both the nucleotide level and the allelic profile level, whereas previous research [332] was only carried out at the allelic profile level. This Bayesian model-based cluster analysis was carried out to to assign NZ *Campylobacter* isolates from poultry host sources into NZ *Campylobacter* isolates from

ruminant host sources and UK *Campylobacter* isolates from poultry and UK ruminant host sources. Then a similar analysis was applied to assign UK *Campylobacter* isolates from poultry host sources into UK *Campylobacter* isolates from ruminant host sources, NZ *Campylobacter* isolates from poultry host sources and NZ ruminant host sources.

5.6.2 Structure analysis results

In the Bayesian clustering analysis, the sampling location information has been incorporated into the modelling to allocate individuals of unknown origin into several pre-defined groups, which included the *Campylobacter* population with known host sources. At the nucleotide level, NZ poultry isolates were assigned into the other three pre-defined sources (NZ ruminant, UK poultry and UK ruminant), with probabilities of 76.8% for UK poultry, 14.2% for UK ruminant, and 9% for NZ ruminant. At the allelic profile level, 60.2% of NZ poultry isolates were assigned into UK poultry, 34.4% of NZ poultry isolates were assigned into UK ruminant, and 5.4% were assigned into NZ ruminant. Assignment of the UK poultry isolates to the other three predefined sources at the nucleotide level with use of Structure 2.3.3 [169, 303] produces probabilities of 19.9% for NZ poultry, 65.1% for UK ruminant, and 15% for NZ ruminant. At the allelic profile level, the assignment of UK poultry isolates produces probabilities of 55% for UK ruminant, 31% for NZ poultry, and 14% for NZ ruminant.

Figure 5.8 shows the analyses carried out at both the nucleotide and the allelic profile levels. The analysis at the nucleotide level provides similar information about the host and geographical effect as the allelic profile level. For NZ isolates from poultry host sources, the host association of *Campylobacter* strains transcends geographical factors because NZ isolates from the poultry industry are more associated with UK poultry than with UK or NZ ruminants. The results also indicate that isolates from UK poultry are more associated with UK ruminant, less associated with NZ poultry, and least associated with NZ ruminant. For UK *C. jejuni* isolates from poultry host sources, geographical factors appear to be more important than host factors.

5.6.3 Discussion about structure analysis

Both observations from Structure 2.3.3 analysis for NZ and UK datasets revealed that geographical isolation plays a role in the evolution and genotype diversity of *Campylobacter*. For the NZ dataset, Structure 2.3.3 results suggest *Campylobacter* genotypes have more association with host species than geographical locations,

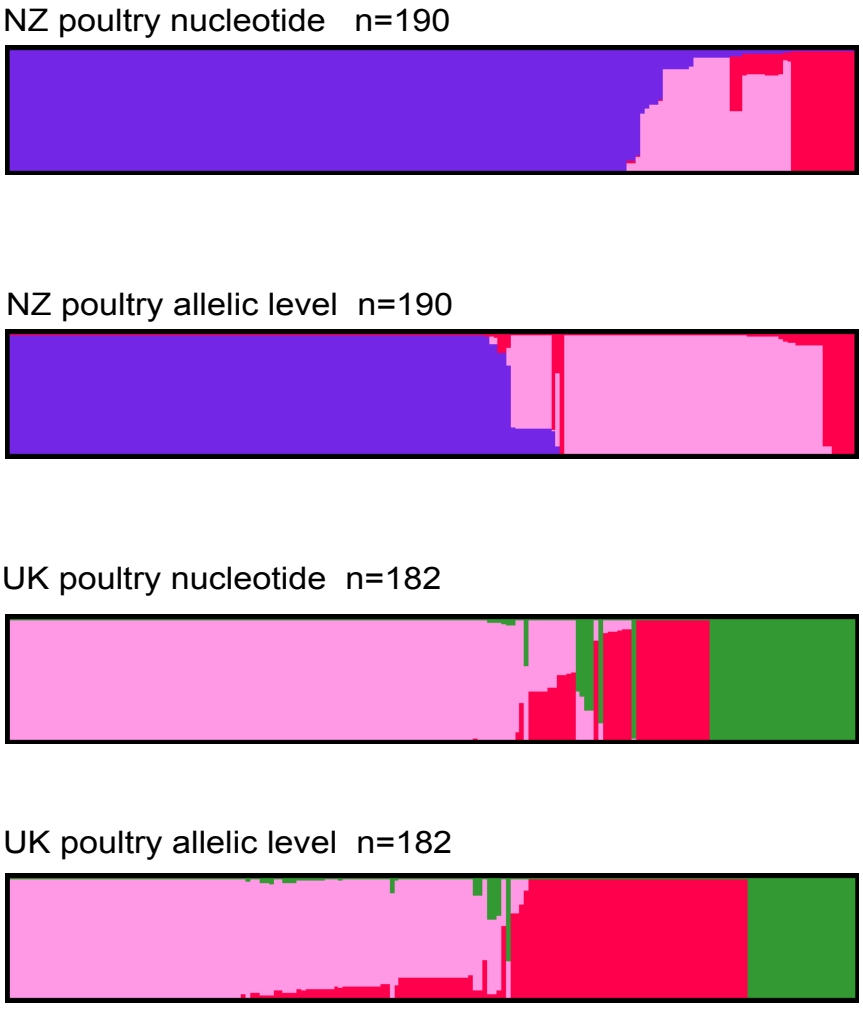


Figure 5.8: The first two plots are produced by Structure 2.3.3 to assign 190 *Campylobacter* isolates from NZ poultry (NZP) into NZ ruminants (NZR), and UK poultry (UKP) into UK ruminants (UKR) host sources. The last two plots assign 182 *Campylobacter* isolates from NZ poultry into UK ruminant, UK poultry, and NZ ruminant host sources. The sample sizes for NZ poultry, NZ ruminants, UK poultry, and UK ruminants are 190, 126, 182, and 716, respectively. In the plot, each isolate is represented by a coloured vertical line. Isolates are grouped by the potential attributed source. For each vertical line (isolates), the estimated probability of the origin is shown by different colours: green represents *Campylobacter* isolates from NZP; red represents *Campylobacter* isolates from NZR; purple represents isolates from UKP; and pink represents isolates from UKR.

which is consistent with previous research [243, 332, 334]. Similar results are also supported by AMOVA analysis. However, these results do not show the same pattern for the UK poultry dataset. To find out why this is so, different patterns from the UK poultry dataset from Structure 2.3.3 analysis were observed, and compared to previous research [243, 332]. One possible reason is the distinctive geographical location of NZ, as geographical isolation can create a unique environment for some specific strain types to survive. Due to the existence of some unique alleles in NZ datasets, it could be difficult to assign UK alleles into them. Previous research [243, 332] has focused on European countries or those in the Northern Hemisphere, therefore, apparent geographical isolation will not affect the assignment of ST as seen in those studies. Structure 2.3.3 analysis is based on the allele frequency, and this is a snapshot of the evolution. Because of the long time period need to allow enough time to exchange or communicate the genetic material, the uniqueness of the NZ database caused by geographical isolation is not strong enough to transcend the NZ and UK commonalities, such as similar host reservoir. This is supported by further detailed comparison of STs between these two countries, which indicates that NZ datasets contain a subset of the UK dataset, and also has its uniqueness in the dataset.

Chapter 6

Estimating the clonal genealogy for ST-474, a commonly found New Zealand *Campylobacter* sequence type

6.1 Introduction

In New Zealand (NZ), until recently, *Campylobacter jejuni* strain type 474 (ST-474) [25, 259] was responsible for more than a quarter of the notified human campylobacteriosis cases, and was widely distributed [246]. Outside NZ, ST-474 has been found infrequently, such as one isolate reported from a poultry sample in the Czech Republic in 1999 (recorded in the *Campylobacter* PubMLST website) [188] and a human sample in France in 2003 [49]. ST-474 shares five out of seven loci with ST-48, and belongs to the clonal complex 48 (CC-48). Compared to the globally distributed ST-48, ST-474 is regarded as endemic to NZ. This raises questions about the mechanisms by which *Campylobacter* generates diversity.

An accurate estimate of the clonal genealogy of ST-474 and related strains would help in understanding how *Campylobacter* populations and their hosts interact. Knowing the clonal relationships of a given group of strain types is also helpful for estimating the age of their most recent common ancestor; the positions of nodes on a given tree can represent the divergence times in the evolutionary process. With the recent rapid development of phylogenetics and bioinformatics, there are now a wide range of methods which can be applied to infer the clonal genealogy of a set of sequences. In this study, the evolutionary relatedness of 59 isolates, including seven ST-474 isolates, can be used to indicate the origin of ST-474 strains.

Phylogeny can be used to reflect the relationships among a group of strain types. On a phylogenetic tree, leaves and nodes are used to represent extant samples and common ancestors.

A simple clonal genealogy can be represented by a tree. However, the occurrence of recombination can disrupt the tree-like relationship for a given group. The occurrence of recombination will make the inference complicated, because when recombination occurs in the evolutionary history of the given sequences, two branches of genealogical tree join together to form a cycle. After several recombination events, it is difficult to represent a clonal genealogy as a tree-shaped relationship. Thus, under these circumstance it is not appropriate to reconstruct a single phylogenetic tree and make inference based on this single tree [404]. Moreover, when recombination occurs, different sets of genes will have different phylogenies, and the current analyses will have limited power to estimate those trees accurately [301, 404].

Most traditional methods, such as Neighbor Joining [316], Maximum Parsimony (MP), and Maximum Likelihood (ML), make the unrealistic assumption that no recombination occurs in the evolutionary history of a set of sequences. In contrast, ClonalFrame version 1.1 [69] is one of the few software programs which currently take into account recombination.

The construction of phylogenetic trees provides important information about the clonal relationship of sequences and is crucial for making inferences about the evolutionary process. Several analytical tools are available for phylogenetic inference using whole genome data. There are two broad types of phylogenetic tree reconstruction methods: methods based on the distance matrix, such as neighbor joining method [316]; and methods based on characters, including maximum parsimony (MP), maximum likelihood (ML), and Bayesian methods [411]. Methods based on the distance matrix are calculated by comparing every possible pair of sequences, then reconstructing the phylogenetic tree based on the genetic distances. In contrast, methods based on characters compare a score calculated by different methods [411]. The MP score for a particular tree is the least number of nucleotide changes on the tree to reconstruct the observed sequence data. For ML, the compared score is a log-likelihood value of each tree [411]. ML and Bayesian methods are all model-based methods [411], while MP is a non-parametric method. Bayesian methods calculate a posterior probability for a given tree. ClonalFrame applies a Bayesian inference method to infer the clonal relationships of a given dataset and also takes the location of recombination into account [69].

This chapter applies a range of phylogenetic methods on 59 isolates of *Campylobacter* to infer clonal genealogy. First, a case study was carried out to estimate and compare the evolutionary clonal genealogy for whole genome length data using a range of

mathematical tools: UPGMA [344], NJ [316], MP, ML, and Bayesian methods [109, 201, 215]. Second, these phylogenetic methods were applied on both multilocus sequence typing (MLST) and the targeted gene reference set (TGRS) datasets and the results compared. Third, based on the first two steps, a phylogeny was chosen to map events on to seven ST-474 isolates. Fourth, the compatibility plot for the 78 specific genes of ST-474, located in the third step, was analysed.

6.2 Phylogenetic analysis and methods

6.2.1 Data

Two datasets were analysed in this section: one simulated dataset and one real *Campylobacter* dataset. The simulated dataset, containing 60 isolates using SimMLST [70], was generated under the constant growth model, recombination rate (100), mutation rate (100), and recombination tract length (500). SimMLST [70] simulates both sequence data and the underlying clonal genealogy that gave rise to these sequence data using a coalescent method. The sequence lengths are 100,000 base pairs. SimMLST can produce four kinds of output for one simulation: the sequences in XMFA format, the clonal genealogy in Newick format, all the trees in Newick format, and the full description of the graph representing the ancestry. In this chapter, the first two outputs were mainly used: the sequences in XMFA format and the clonal genealogy in Newick format.

The real dataset comprised 59 whole genome-sequenced isolates, of which 44 were *C. jejuni* and 14 were *C. coli* isolates, and one was *C.* sp. nov.. They were sequenced using the Allan Wilson Centre Genome Service's Illumina GAII sequencer. These isolates were chosen to ensure they were reasonably representative of the *Campylobacter* population in New Zealand. The 59 isolates were sequenced using indexing technology, and then mapped to *C. jejuni* NCTC 11168 reference genome using the mapper BWA [226] to generate initial consensus sequences. As this approach is limited to showing only sequences that map to the reference with only a few mismatches per read, it ignores any sequence not found in the reference genome. For this reason, the short read de novo assembler Velvet [416] was used to generate de novo contigs for the genome. Gene prediction was performed on the de novo contigs using Glimmer [63, 317].

The targeted gene reference set (TGRS) is defined to include: 1) the seven full length MLST genes for *Campylobacter*, including the whole *gln*A gene sequence, and three regions between the six MLST genes (*asp*A-*atp*A inclusive, *gly*A-*pgm* inclusive, and *tkt*-*glt*A inclusive); and 2) three "hypervariable" regions, including flagellar genes,

lipo-oligosaccharide biosynthesis, and capsular polysaccharide biosynthesis. The MLST profile for a particular strain can be determined by its TGRS profile by restricting to the 7 gene fragments used in MLST. The TGRS data contain thousands of loci, whereas ST is defined by MLST profile; therefore, several isolates can share one ST. There were 34 unique STs in the TGRS dataset; most STs were represented by one to three isolates, whereas there were seven ST-474 isolates.

6.2.2 Phylogenetic analysis and methods comparison

The clonal genealogies were estimated by UPGMA, NJ, MP, ML, and Bayesian methods. UPGMA was applied using Geneious Pro v5.6.5 [79], NJ and MP were applied using PAUP*, version 4.0 [359]; and ML was calculated by RAxML web server [346], with the best model of DNA substitution tested in ModelTest, version 3.7 [300]. The GAMMA model of rate heterogeneity is selected using the Akaike Information Criterion (AIC) 32, 299 as the criterion. The Bayesian methods were applied using ClonalFrame (CF), version 1.1 [69]. Among these methods, Clonal Frame is the only one that accounts for recombination as well as mutation. The estimated clonal genealogy for the simulated dataset by different methods was compared to the true genealogy using the program PAUP*, version 4.0 [359]. For the two given trees, the symmetric difference, also known as the Robinson-Foulds distance, is calculated by counting the number of splits that only appear in one tree but not in the other [308, 293]. For the output of several runs in ClonalFrame, Gelman and Rubin convergence test 132 was applied to check the convergence, using ClonalFrame (CF), version 1.1 [69].

6.2.3 Mapping events on the ST-474 branch

In order to identify the genes which are compatible or incompatible with the given topology, eight isolates (seven ST-474 isolates and one ST-48 isolate) were compared using the online tools on the BIGSdb website (http://pubmlst.org/) and mapped to the reference *C. jejuni* strain NCTC11168. All eight isolates belong to the ST-48 complex, and they were isolated from samples in New Zealand.

6.2.4 Compatibility

The compatibility plot has been drawn for the informative loci. The informative locus requires to have two different alleles appear at least twice on that locus for a set of isolates. The uninformative loci refer to the loci that can always fit exactly

on any tree, e.g. the variant loci on pendant edges. The compatibility plot is based on the number of extra events that would be required over and above the minimum for them to fit on the same tree (this minimum is just the number of unique alleles - value 1 for each loci). The 0s indicate compatible characters, 1s and 2s represent incompatible characters. UPGMA is used on the matrix of excess distances (i.e. the compatibility matrix) to find the cliques. A group of pairwise compatible loci and to which no other loci can be added without conflict is termed a clique [91, 401].

6.3 Results

6.3.1 The phylogeny of the simulated dataset

Figures 6.1-6.6 show the true clonal genealogy produced by SimMLST and estimated clonal genealogies from UPGMA, NJ, MP, and ML using simulated data. Table 6.1 shows the symmetric-difference distances among trees obtained from the generated trees, UPGMA (Figure 6.2), NJ (Figure 6.3), MP (Figure 6.4), and ML (Figure 6.5). The symmetric-difference distances are defined as a set which contains only one of two sets of splits, and excludes the shared splits. All of these splits belong to the same sets of sequences. ClonalFrame (CF), version 1.1 [69] does not use the concatenated simulated data, instead it requires an input file which separates every locus by equals signs. The six main clusters were reconstructed correctly by applied methods (UPGMA, NJ, MP, ML, and CF) and one outlier (sequence type 19), but the true phylogeny has not been identified clearly. The largest number of incorrect splits was obtained using CF [69].

The true clonal genealogy (Figure 6.1) demonstrates six major phylogenetic groups with one outlier (sequence type 19), designated by different colours. The true clonal genealogy was produced by SimMLST, and colors represent different groups. The estimated clonal genealogies from UPGMA (Figure 6.2) reconstruct the whole clonal genealogy, although the outlier sequence 19 is not identified. The estimated clonal genealogies from NJ (Figure 6.3) also reconstruct the whole clonal genealogy. However, the outlier sequence 19 is not identified, and the divergence of the red cluster has been over estimated. The estimated clonal genealogies from MP (Figure 6.4) is the strict consensus tree for 12 clonal genealogies, which are based on the MP method. This method produces 12 trees with the same score of the estimated clonal genealogies. It also reconstructs six major clusters for the whole clonal genealogy, although the divergence of the red cluster has been over estimated and the outlier 19 is not clearly identified. The estimated clonal genealogies from ML (Figure 6.5) also reconstructs the whole clonal genealogy. Again, the outlier 19 is not identified, and

Table 6.1: Symmetric-difference distances between trees for ML, UPGMA and NJ, with all trees unrooted. The tree was calculated for the simulated dataset.

	CF	NJ	UPGMA	ML	MP	TRUE
CF	0					
NJ	98	0				
UPGMA	98	22	0			
ML	98	32	22	0		
MP	81	25	21	17	0	
TRUE	98	32	20	22	19	0

the divergence of red and purple clusters has been over estimated. The estimated clonal genealogies from CF (Figure 6.6) also reconstructs the whole clonal genealogy, but the divergence between the red, green, and blue, and the yellow, black, and purple clusters has been over estimated.

All of the applied phylogenetic methods can consistently identify the major clusters; however, none of the applied methods can reconstruct the true clonal genealogy within clusters accurately. Some detailed divergence has not been constructed by any of the applied methods. It can be seen from Table 6.1, even if the convergence has been achieved for CF, the results are largely different from the true tree.

6.3.2 Results for the targeted gene reference set

The aim of this analysis was to apply the existing methods to make phylogenetic inferences on the Targeted Gene Reference Set (TGRS) dataset, which is currently the most comprehensive dataset on NZ isolates available. In all the applied phylogenetic methods (Figures 6.7-6.11), the green colour block represents *C. coli* and the yellow represents *C. jejuni* ST-474. Although the phylogenies for all isolates inferred by the various methods are different (Figures 6.7-6.11), at the tips of all the inferred phylogenies, all the isolates which have known STs are grouped together according to their STs, such as ST-50, ST-53, ST-190, ST-520, ST-1324, ST-1342, and ST-2536. These listed STs relate to two or more isolates.

The results of UPGMA (Figure 6.7), NJ (Figure 6.8), MP (Figure 6.9), ML (Figure 6.10), and two results from ClonalFrame show there are two apparent clusters: one for *C. jejuni* and one for *C. coli*. Two chains were run by CF (Figure 6.11, and Figure 6.12). Figure 6.11 shows 10000 iterations after the burn-in, 10000 burn-in iterations, and 10 iterations performed between recording the parameter values in the posterior sample (also called the thinning interval). By comparing the results from CF (Figure 6.11 and Figure 6.12), it can be seen from Gelman and Rubin convergence test (values > 1.2) that convergence has not been achieved, after 300+

hours desktop computation for each run. The main difference is the divergence time of *C.* sp. nov.. This occurs because it is difficult for CF to converge. From Section 6.3.1, it can be seen that even if convergence has been achieved for CF, the results are less useful for making phylogenetic inference.

6.3.3 Results for the MLST dataset

Figures 6.13-6.17 show estimated clonal genealogies from UPGMA (Figure 6.13), NJ (Figure 6.14), MP (Figure 6.15), ML (Figure 6.16), and CF (Figure 6.17) using MLST data. In order to compare the phylogeny inferred by MLST and TGRS, the known sequence types (STs), and their MLST scheme genes have been extracted from PubMLST databases. In total, there are 34 different STs from *C. jejuni* and *C. coli* that have been analysed. Three of them are *C. coli* (ST-3302, ST-3232, and ST-3072), and the others are *C. jejuni*.

Although the phylogenies for all isolates inferred by various methods are different (Figures 6.13- 6.17), ST-48 and ST-474 are always grouped together across all the applied phylogenetic methods. Furthermore, three *C. coli* STs (ST-3302, ST-3232, and ST-3072) are always grouped together across all the applied phylogenetic methods.

The results of UPGMA (Figure 6.13), NJ (Figure 6.14), MP (Figure 6.15), ML (Figure 6.16), and two results from ClonalFrame show there are two apparent clusters: one for *C. jejuni* and one for *C. coli*. They also show ST-48 and ST-474 are grouped together.

For the UPGMA method, the phylogeny inferred by MLST (Figure 6.13) is quite different from the phylogeny inferred by TGRS (Figure 6.7). The differences can be seen from the divergence of ST-520: in the phylogeny inferred by MLST (Figure 6.13), ST-520 diverged from the branch containing ST-50 and ST-21, but in the figure that relates to TGRS (Figure 6.7), ST-520 (isolates P28a and 28127) diverged from the ST-48 clonal complex branch (ST-474 and ST-48).

For the NJ method, the phylogeny inferred by MLST (Figure 6.14) is also dissimilar to the phylogeny inferred by TGRS (Figure 6.8). The variations can also be seen from the divergence of ST-520. In the phylogeny inferred by MLST (Figure 6.14), ST-520 diverged from the branch containing ST-50 and ST-21, but in the phylogeny inferred by TGRS (Figure 6.8), ST-520 (isolates P28a and 28127) diverged from the ST-21, ST-474 and ST-48 branch.

For the MP method, unlike TGRS (Figure 6.9), the phylogeny inferred by MLST (Figure 6.15) grouped ST-2341 with ST-3711 and at the deeper branch. The ST-2341 (isolates M880a and S263a) is also at the shallow part of the phylogeny.

For the ML method, the phylogeny inferred by MLST (Figure 6.16) also contrasts with the phylogeny inferred by TGRS (Figure 6.10) for the divergence of ST-2341.

Based on the above four analyses, the comparison of results from MLST and TGRS show that different evolutionary histories are inferred by the same method between these two datasets. This is important because it shows that different subset of genes produce different tree topologies.

Differences can also be seen in the phylogeny reconstruction s using CF. CF methods reached convergence for the MLST dataset (Gelman and Rubin convergence test value < 1.2 69, 132), after five hours desktop computation for each run. However, it did not reach convergence for TGRS datasets (Gelman and Rubin convergence test value > 1.2), after 300+ hours desktop computation for each run. For TGRS datasets, the *C. coli* clade can be identified clearly. Similarly to other results, ST-48 and ST-474 are grouped together. In Figure 6.17, ST-48 and ST-474 can be seen grouped together at the shallow part of the branch, however, in Figure 6.11, ST-48 and ST-474 are grouped together. Compared to the result for the simulated dataset, it can be seen that CF can capture the main structure of phylogeny.

6.3.4 Mapping events on ST-474 related phylogeny

There were 1667 loci available in the online database for the seven ST-474 and one ST-48 isolates with full genome data. Within these loci, 274 loci contain variant allelic profiles. Figure 6.18 is a parsimony tree produced by a heuristic search using PAUP*, version 4.0 [359]. It uses 849 nucleotide changes on the complete data set of 274 loci. The parsimony informative sites are defined as the loci that have two alleles that appear at least twice each. Out of these 274 loci, 83 informative loci were extracted and compared, the others (191 loci) can always fit exactly on any tree so they are not informative about topology in a parsimony setting. For the seven ST-474 isolates, 677 mutations are used on the complete data set of 274 loci, and 78 informative loci were located.

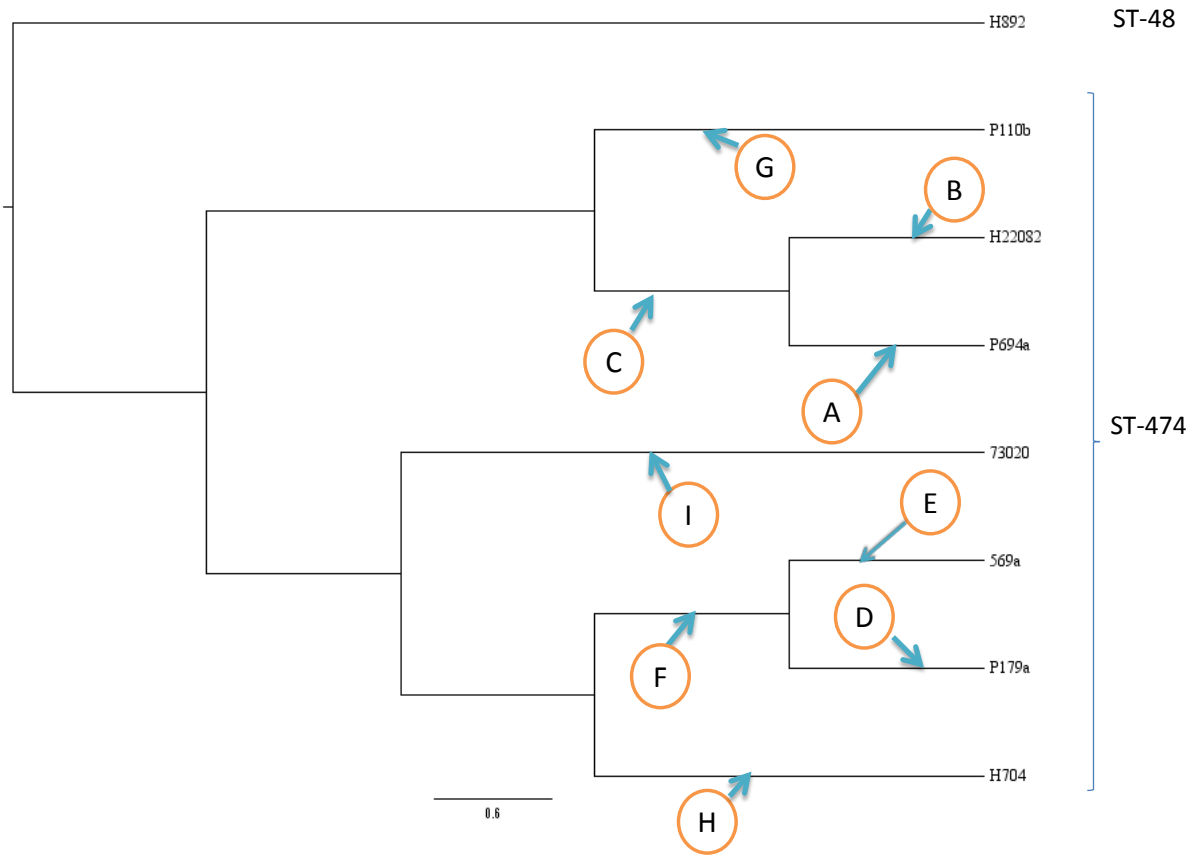


Figure 6.18: Mapping events on the phylogeny of ST-474 produced using PAUP*, version 4.0 [359].

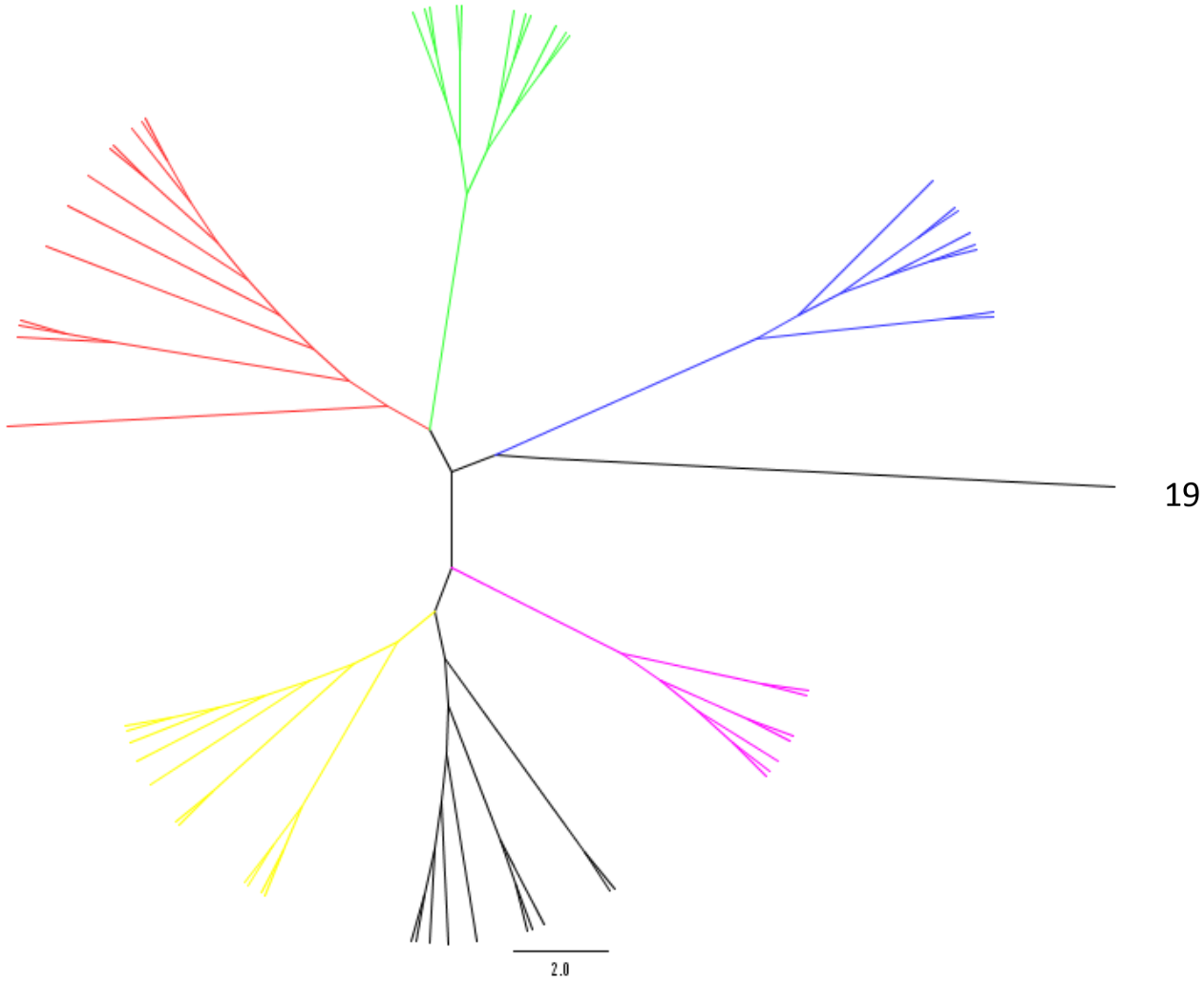


Figure 6.1: The clonal genealogy generated by SimMLST [70] under given recombination rate (100) and mutation rate (100), with recombination tract length 500. Different colours represent different clonal lineages. Six main clusters and one sequence type, number 19, are observed. This represents the true clonal genealogy generated by SimMLST for all given parameters.

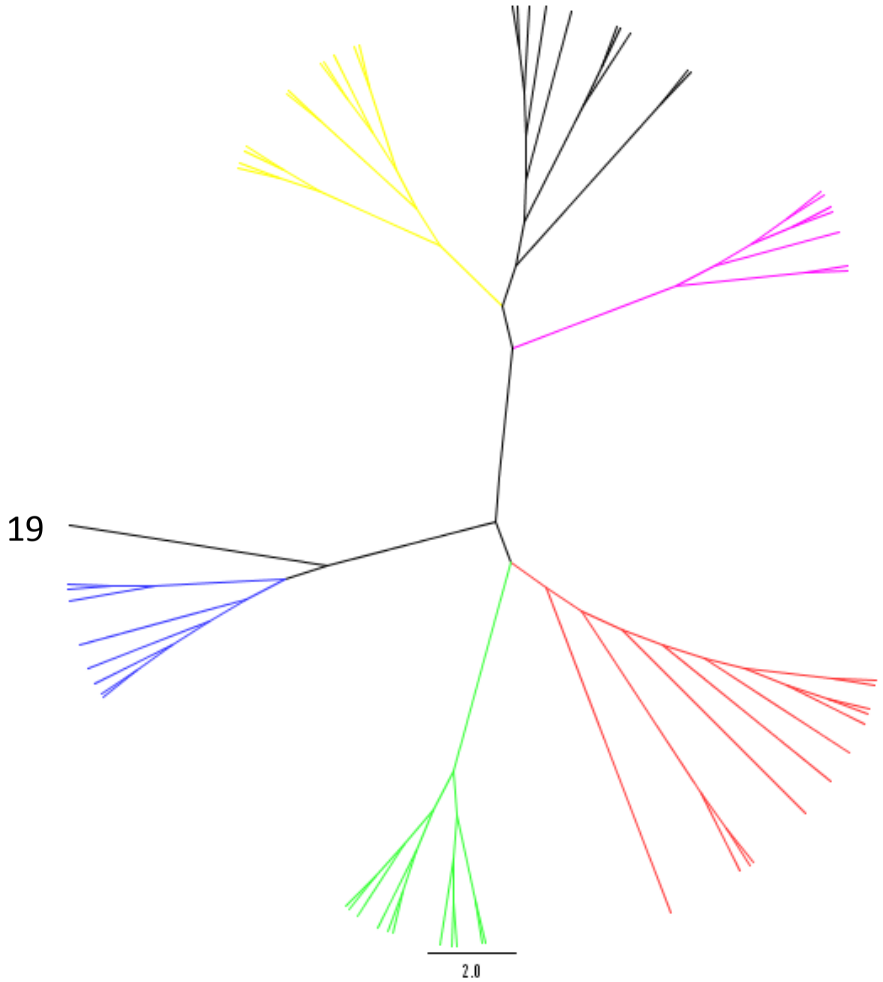


Figure 6.2: The UPGMA tree was produced by Geneious Pro v5.6.5 [79]. The tree was calculated for the simulated dataset. Different colours represent different clonal genealogies/clusters. Six main clusters are consistent with the true genealogy, although sequence type number 19 is misclassified into the blue cluster.

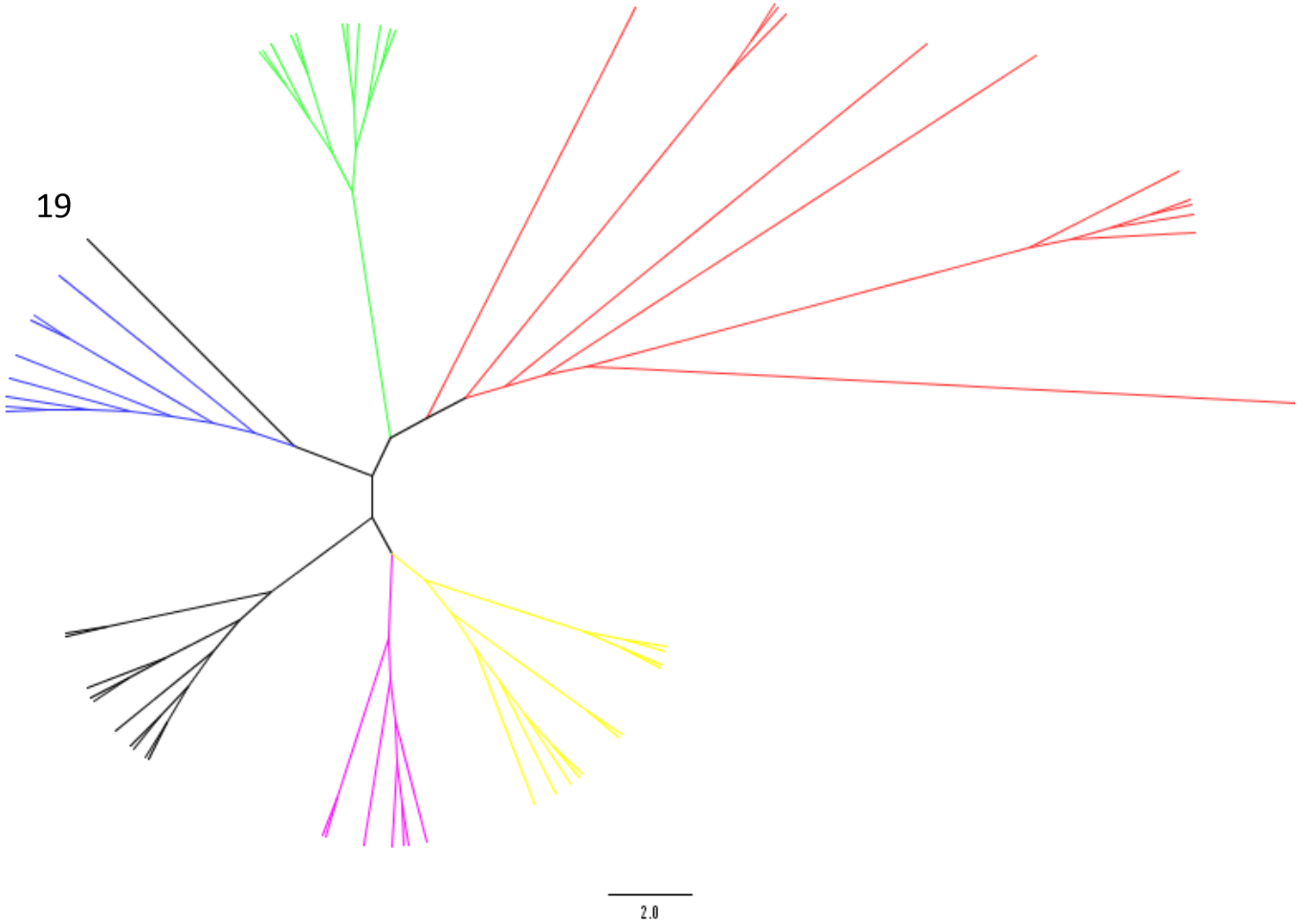


Figure 6.3: The NJ tree was produced by PAUP*, version 4.0 [359]. The tree was calculated for the simulated dataset. Different colours represent different clonal genealogies/clusters. Six main clusters are consistent with the true genealogy, although the diversity for the red clusters is overestimated compared to the true genealogy, and sequence type number 19 is misclassified into the blue cluster.

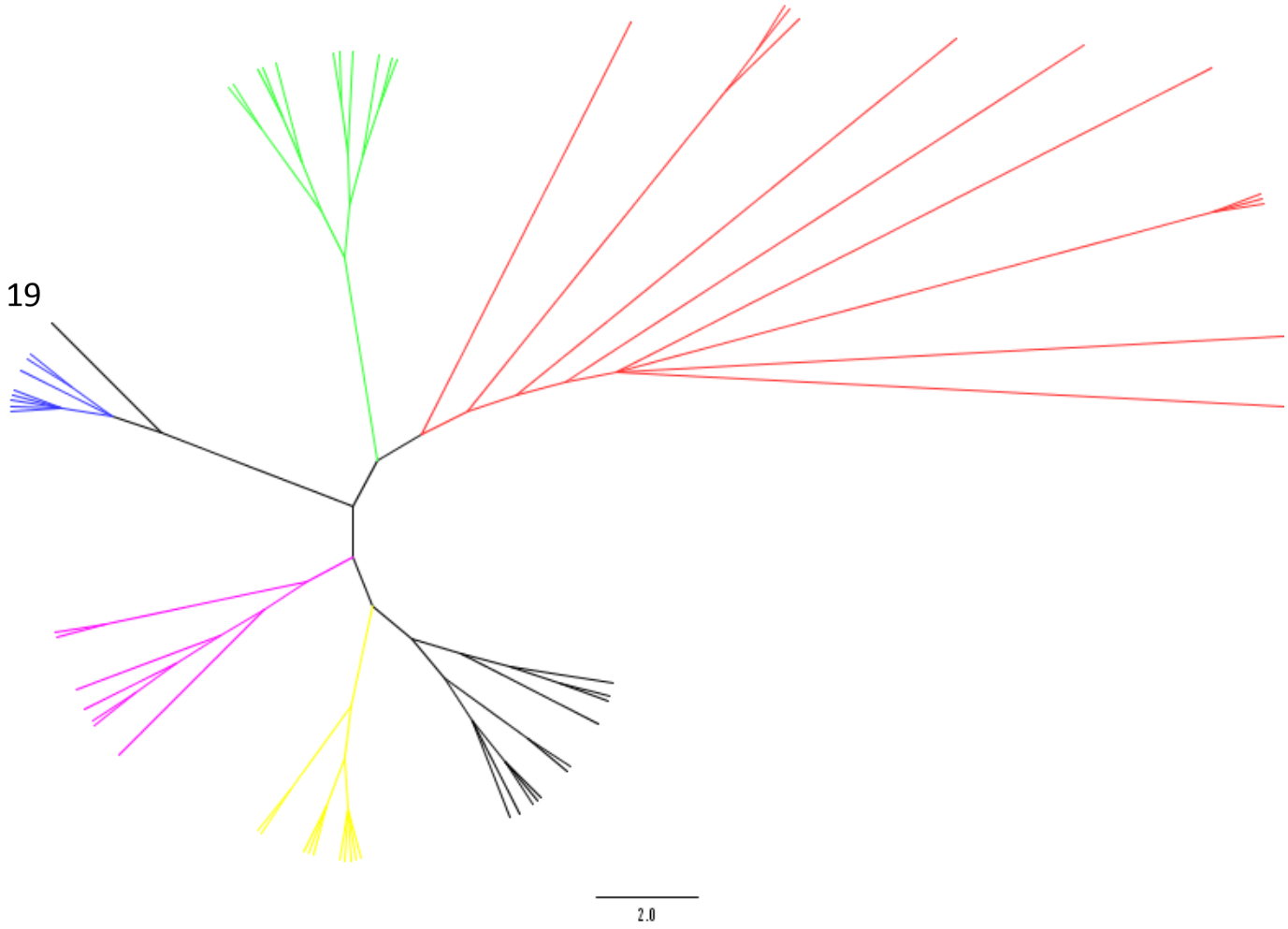


Figure 6.4: The strict consensus tree of the set of 12 maximum parsimony trees. The tree was calculated for the simulated dataset by maximum parsimony using PAUP*, version 4.0 [359]. Different colours represent different clonal genealogies/clusters. Six main clusters are consistent with the true genealogy, although the diversity for the red clusters is overestimated compared to the true genealogy, and sequence type number 19 is misclassified into the blue cluster.

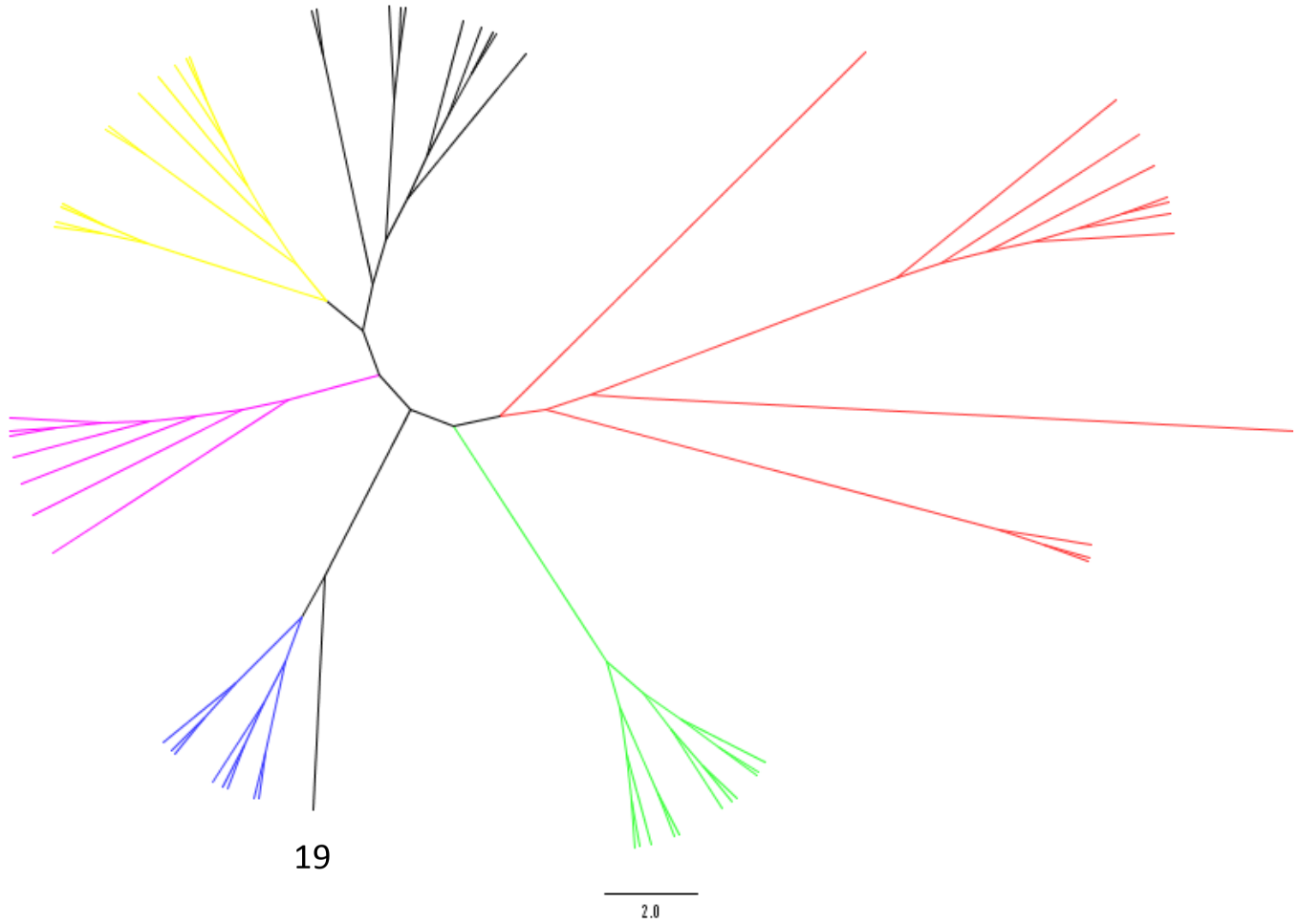


Figure 6.5: The ML tree was produced by RAxML web sever [346]. The tree was calculated for the simulated dataset. Different colours represent different clonal genealogies/clusters. Six main clusters are consistent with the true genealogy, although the diversity for the red clusters is overestimated compared to the true genealogy.

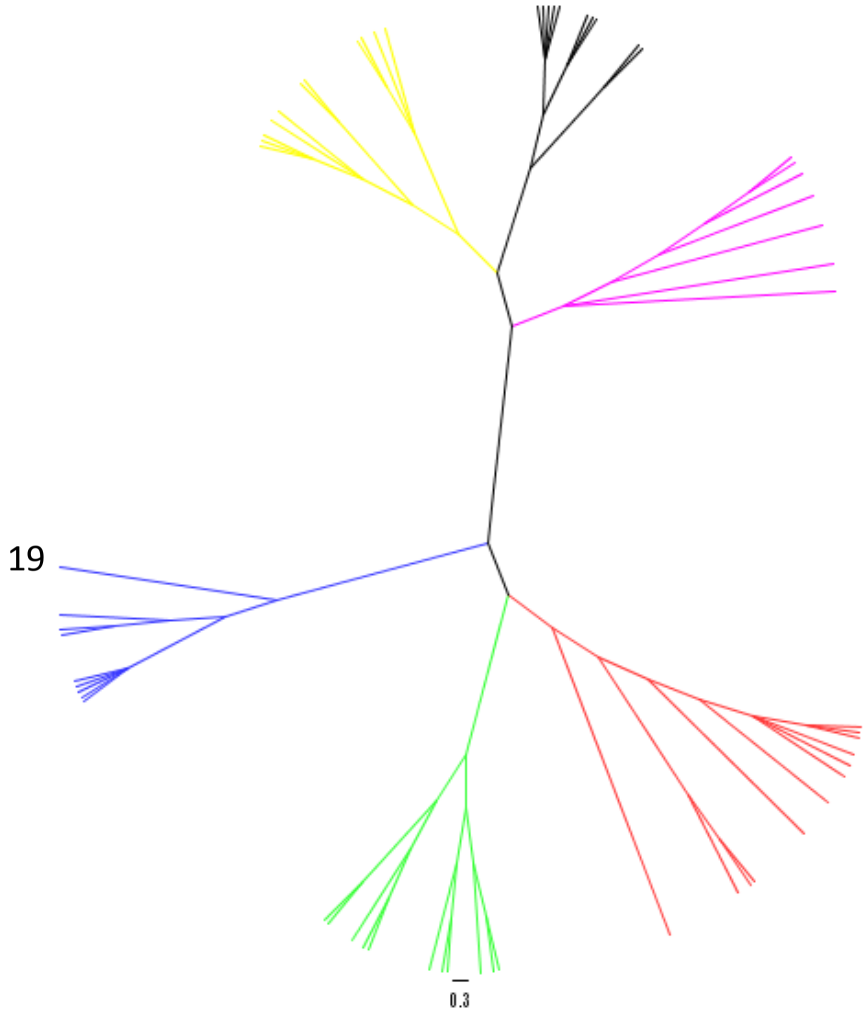


Figure 6.6: The clonal genealogy was estimated by ClonalFrame, version 1.1 [69]. The tree was calculated for the simulated dataset. Different colours represent different clonal genealogies/clusters. Six main clusters are consistent with the true genealogy, although the diversity for all clusters are less estimated compared to the true genealogy, and sequence type number 19 is misclassified into the blue cluster.

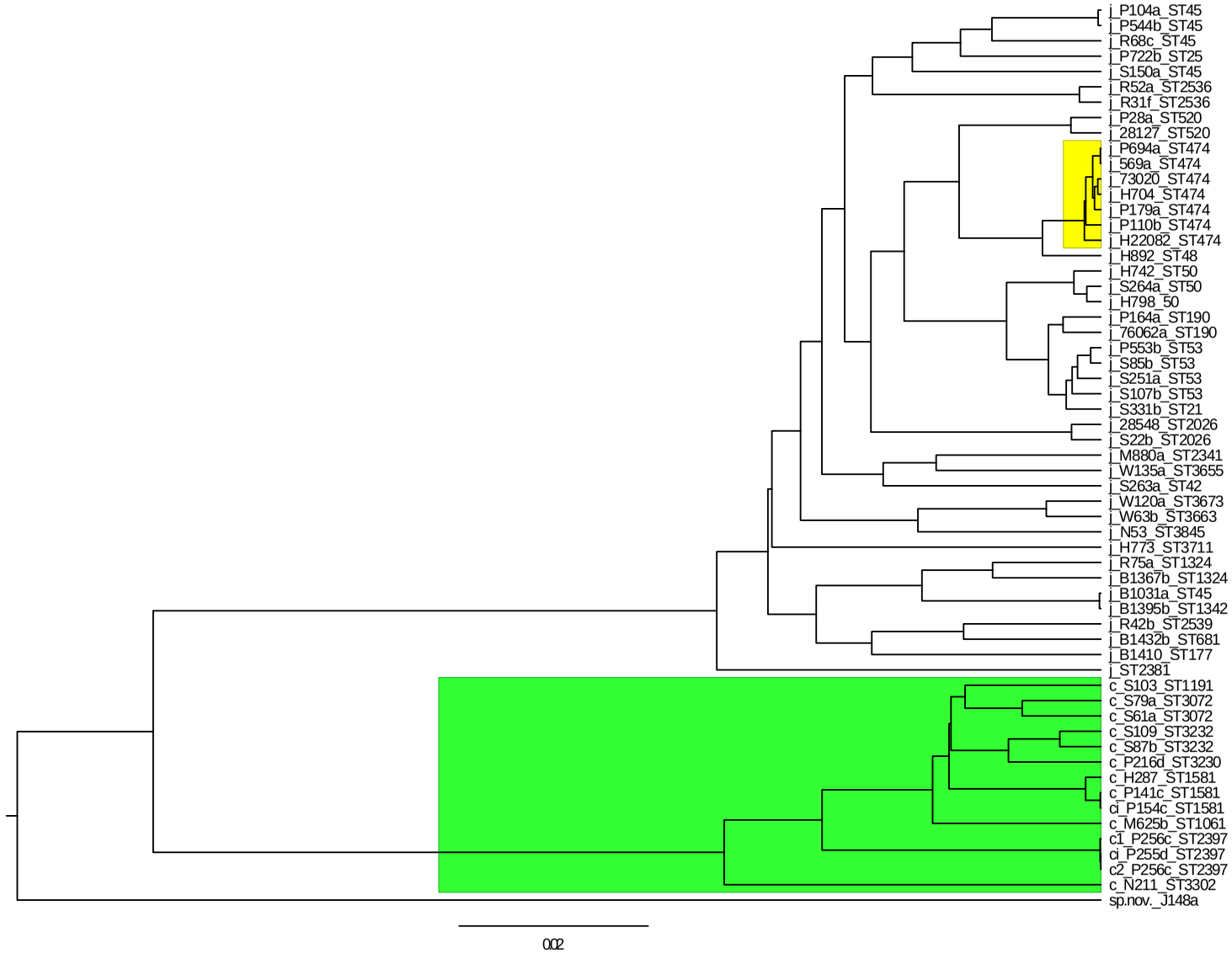


Figure 6.7: UPGMA tree for 59 *Campylobacter* isolates using concatenated TGRS data, implemented by Geneious Pro v5.6.5 [79].

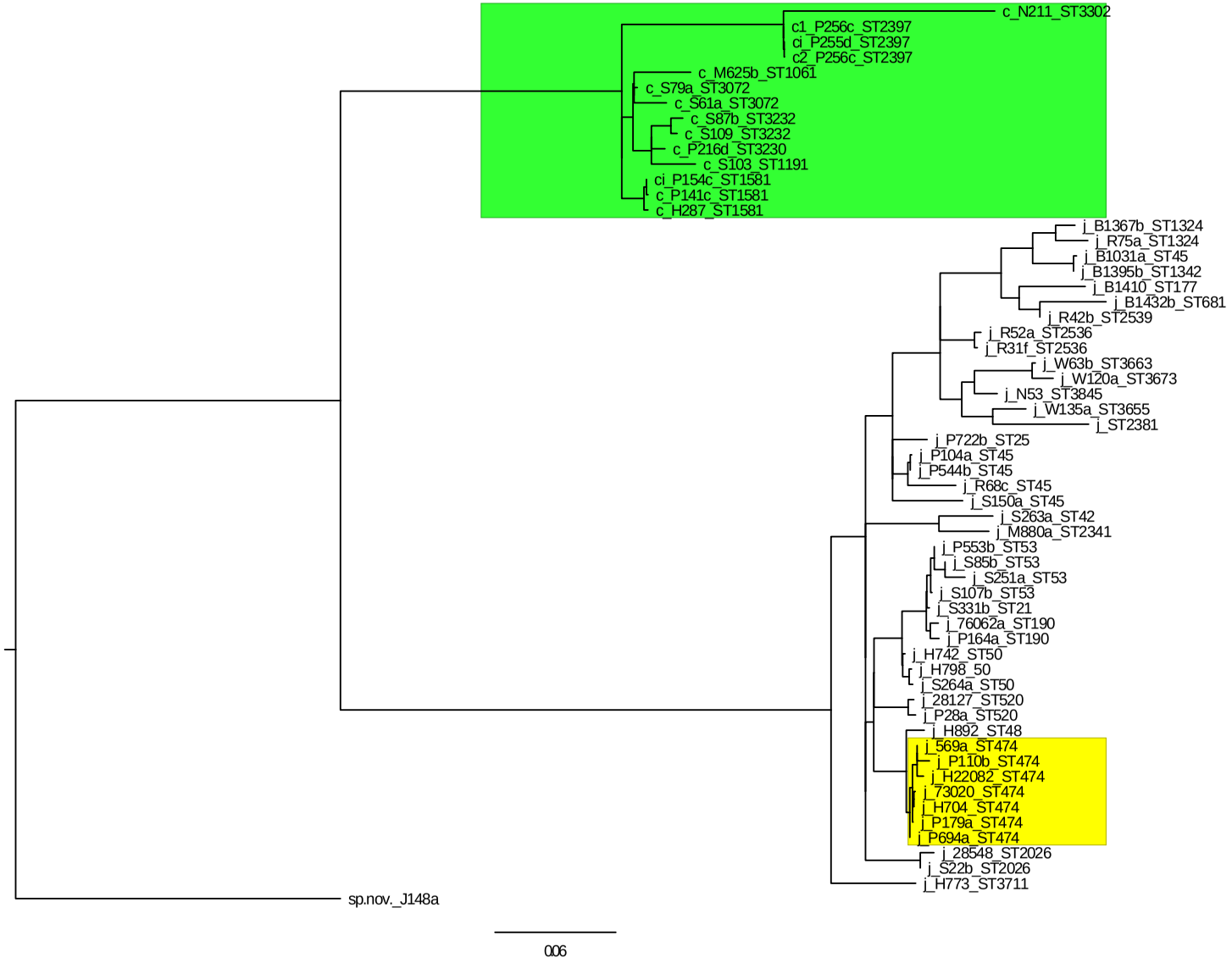


Figure 6.8: Neighbor-Joining plot for 59 *Campylobacter* isolates using concatenated TGRS data by PAUP*, version 4.0 [359].

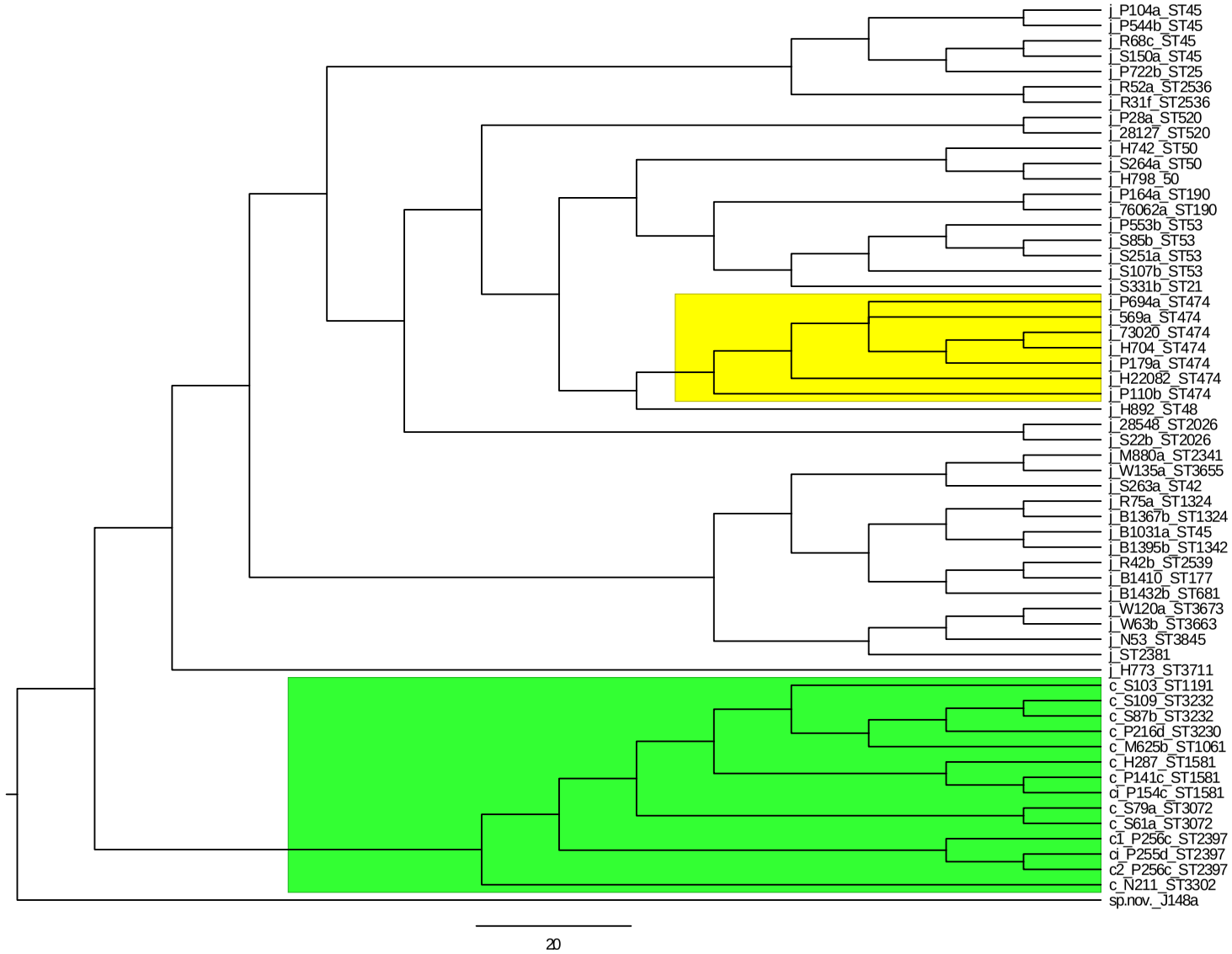


Figure 6.9: Maximum parsimony plot for 59 *Campylobacter* isolates using concatenated TGRS data by PAUP*, version 4.0 [359].

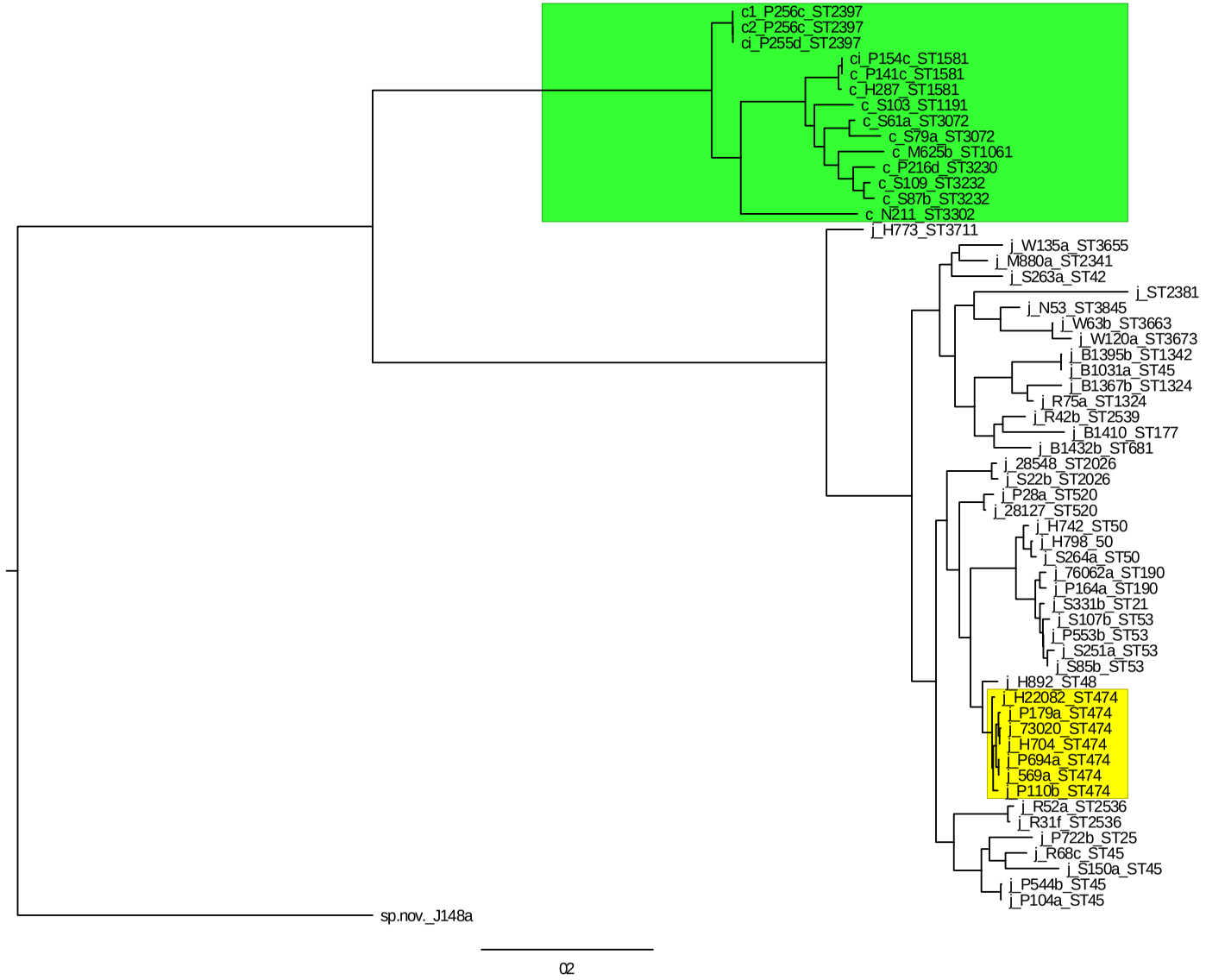


Figure 6.10: ML plot for 59 *Campylobacter* isolates using concatenated TGRS data implemented by RAxML web sever [346].

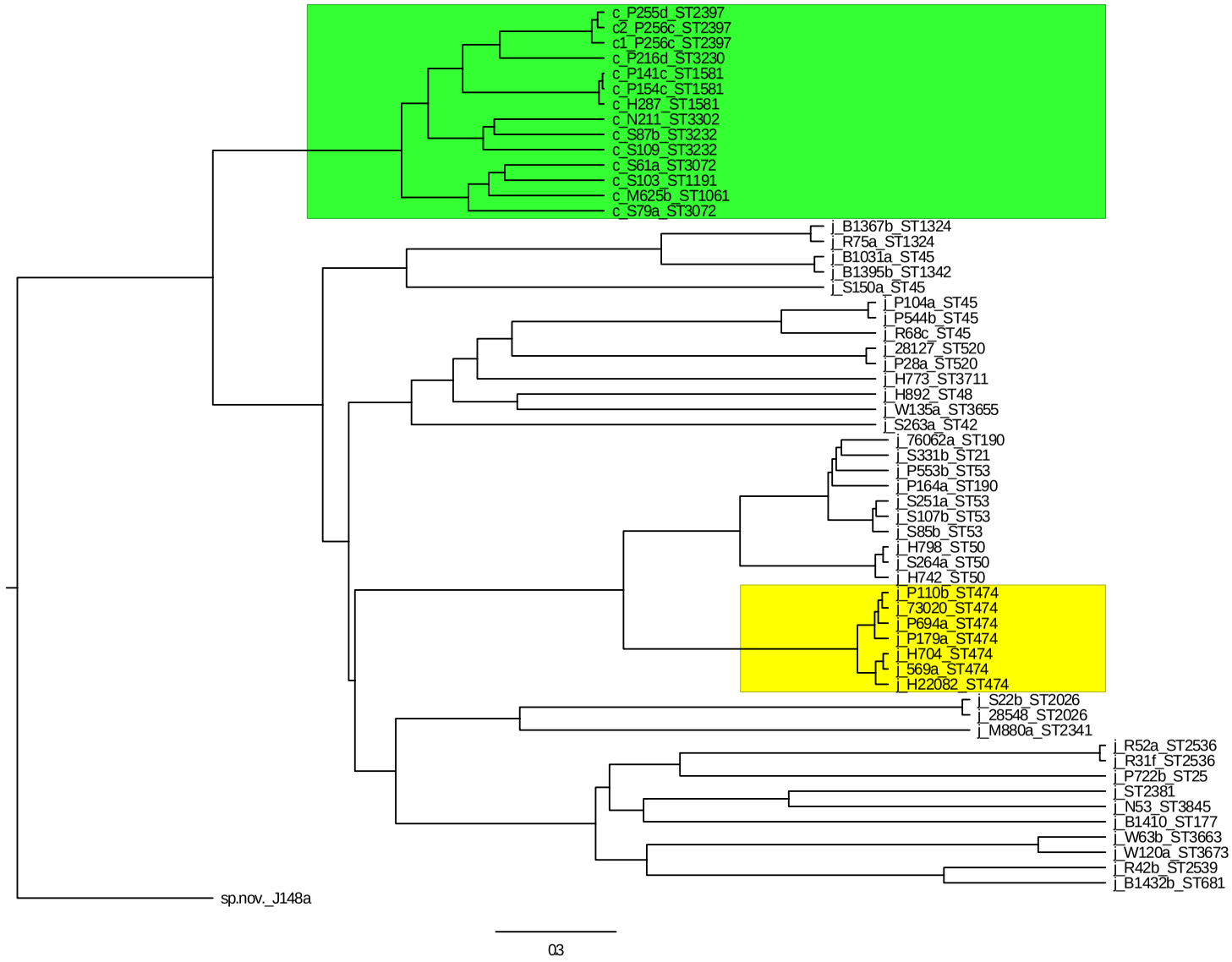


Figure 6.11: Reconstruction of the phylogeny of 59 *Campylobacter* isolates using concatenated TGRS data implemented by ClonalFrame [69] (result 1). This result is based on 10000 iterations after the burn-in, 10000 burn-in iterations, and 10 iterations performed between recording the parameter values in the posterior sample (also called the thinning interval).

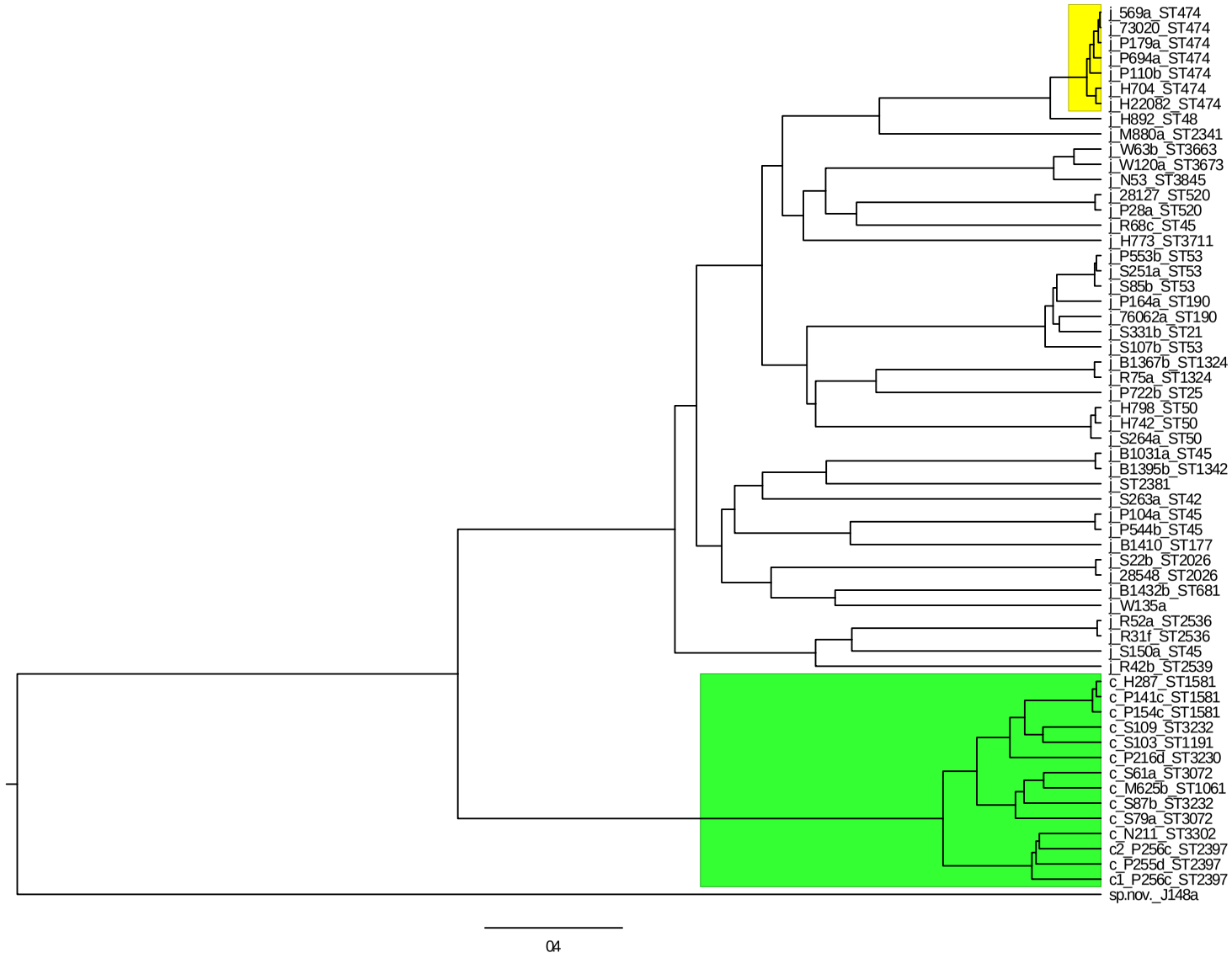


Figure 6.12: Reconstruction of the phylogeny of 59 *Campylobacter* isolates using concatenated TGRS data implemented by ClonalFrame [69] (result 2). This result is also based on 10000 iterations after the burn-in, 10000 burn-in iterations, and 10 iterations performed between recording the parameter values in the posterior sample (also called the thinning interval).

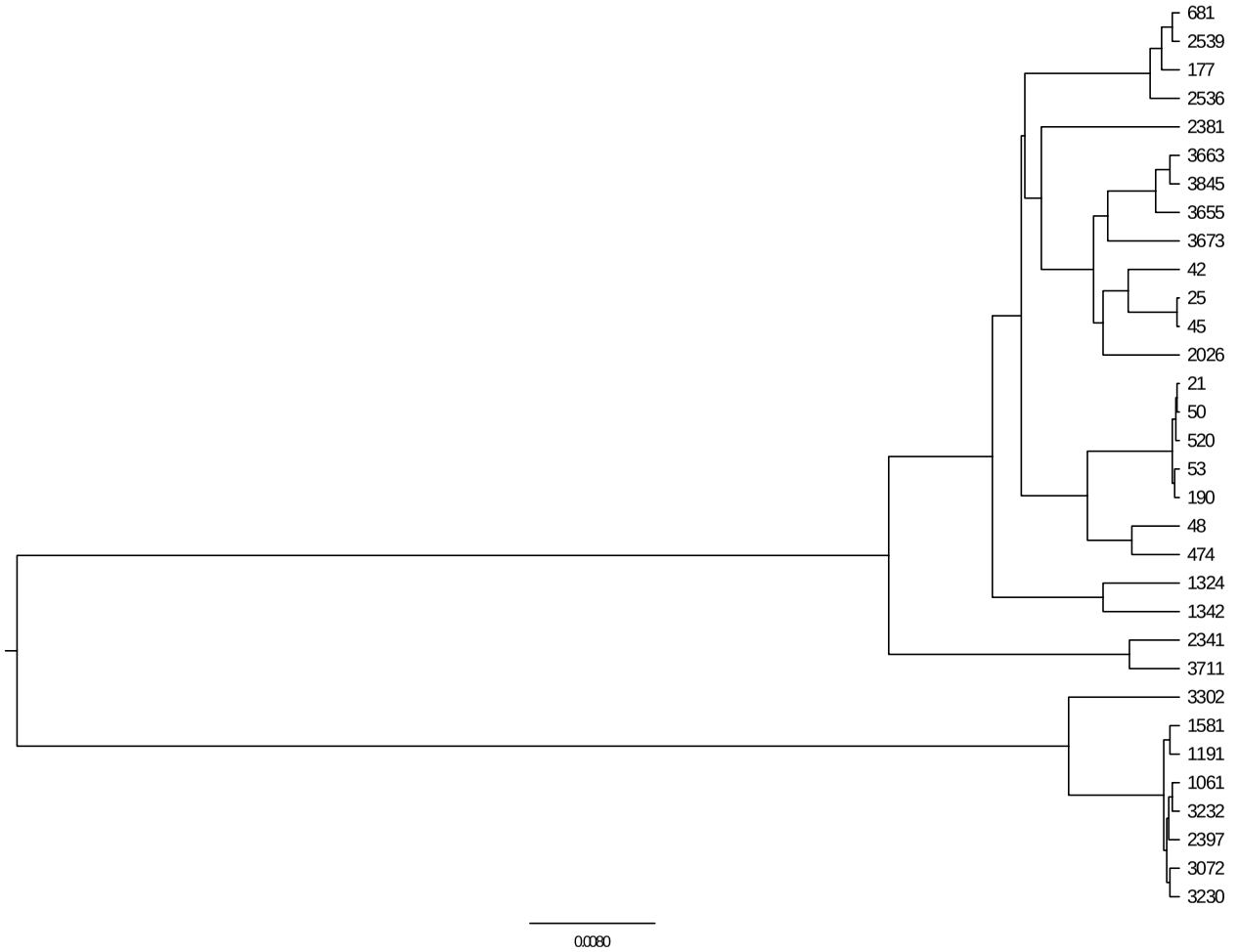


Figure 6.13: UPGMA tree for 33 STs of *Campylobacter* implemented by Geneious Pro v5.6.5 [79]. Numeric labels represent STs.

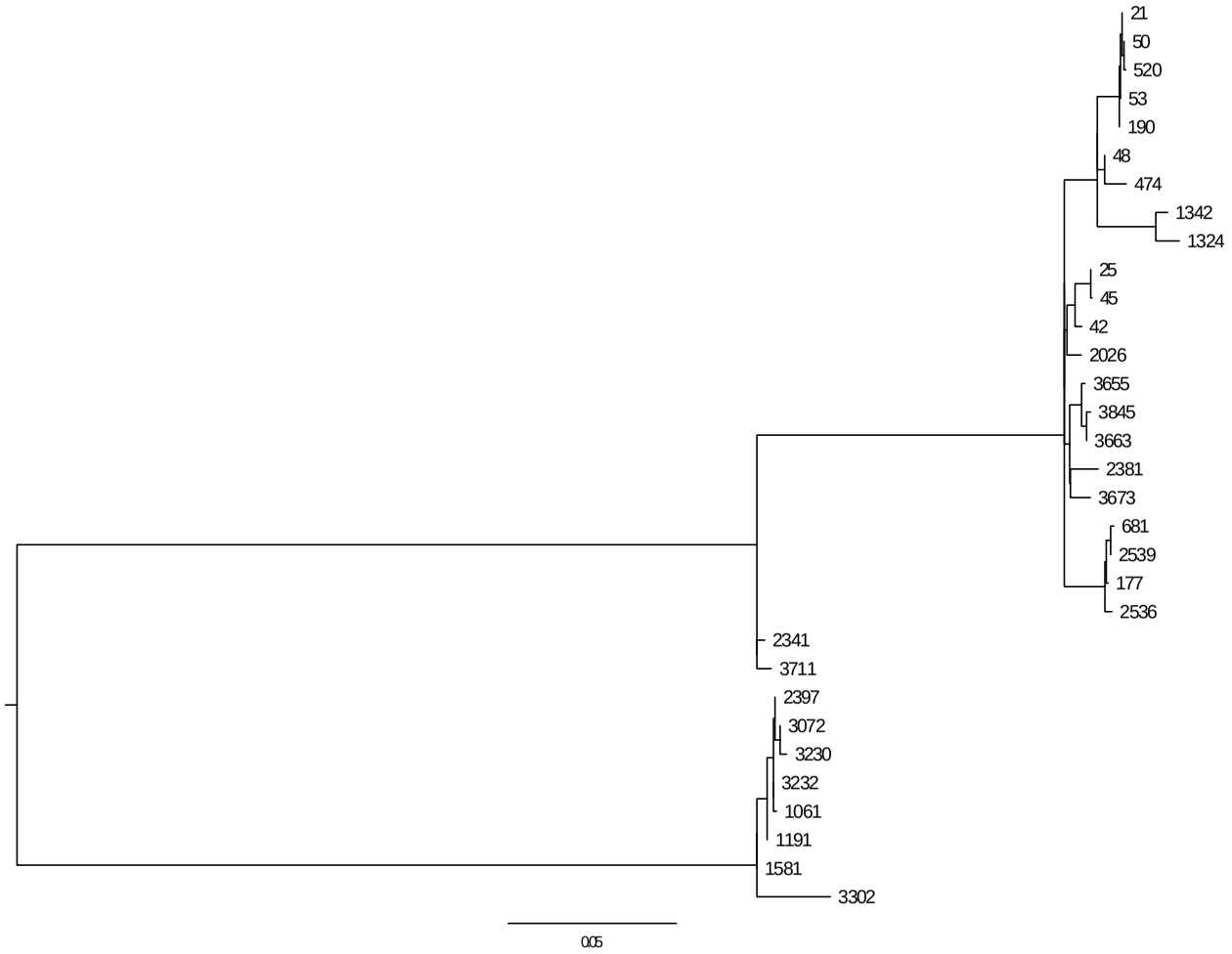


Figure 6.14: Neighbor-Joining plot for 33 STs of *Campylobacter* by PAUP*, version 4.0 [359]. Numeric labels represent STs.

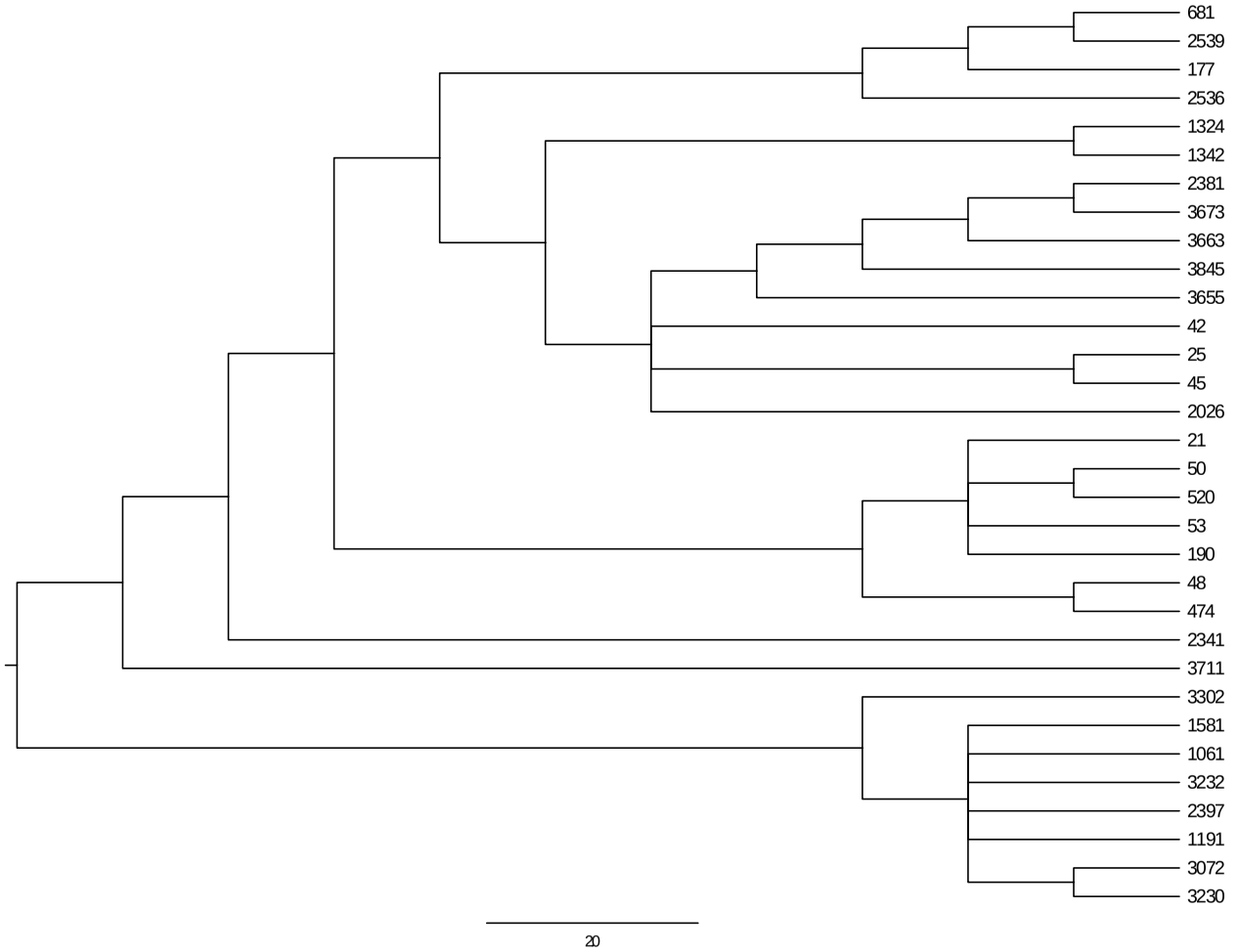


Figure 6.15: Strict consensus tree of 96 Maximum parsimony trees for 33 STs of *Campylobacter* by PAUP*, version 4.0 [359]. Numeric labels represent STs.

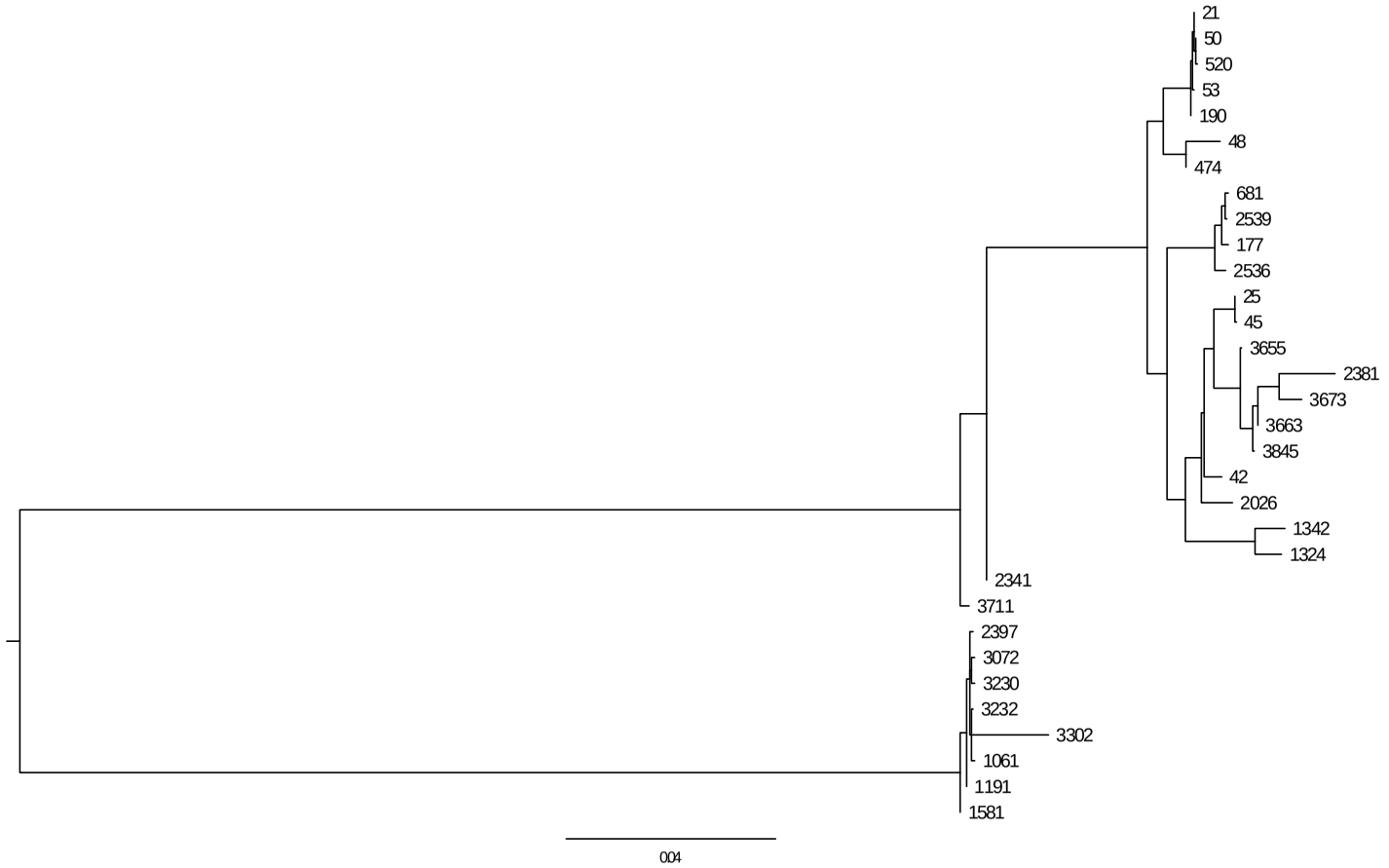


Figure 6.16: ML plot for 33 STs of *Campylobacter* implemented by RAxML web sever [346]. Numeric labels represent STs.

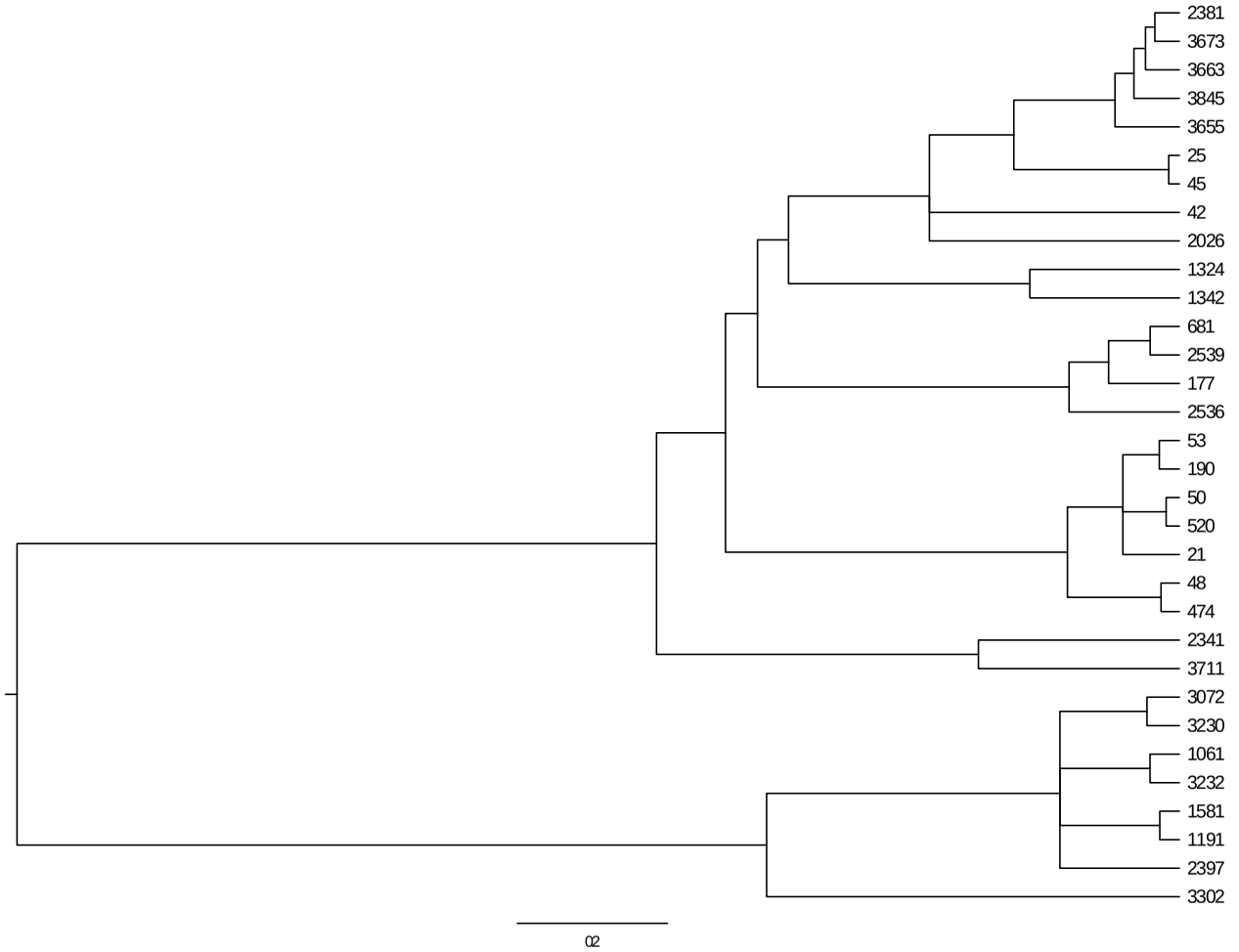


Figure 6.17: Phylogeny plot for 33 STs of *Campylobacter* implemented by ClonalFrame, version 1.1 [69]. Numeric labels represent STs.

Table 6.2: Number of variants for the given position

Position Name	Number of variants unique to that position
A position (P694a)	131
B position (H22082)	58
C position (P694a&H22082)	38
D position (P179a)	10
E position (H569a)	10
F position (P179a&H569a)	4
G position (P110b)	21
H position (H704)	68
I position (H73020)	47

In order to map possible events to a given phylogeny, all the loci which can cause split variants were listed (shown in Appendix 6.6). Split variants are defined as unique events that occurred on the branches, represented by letters from A to I. For the MP phylogeny (Figure 6.18), isolate H892 was plotted as the outgroup, because it is ST-48, and the other isolates belong to ST-474. Position A represents all the events that occurred along P694a branch because all the recorded loci (shown in Appendix 6.6) only appeared in isolate P694a. Similarly, letters B, D, E, G, H, and I represent the variant loci appear exclusively (appeared and only appeared) on those pendant edges. Letters C and F represent the variants for a marked non-pendant edges branch. For example, position F represents the allelic profile is the same for both P179a and H 569a, but differs from the other six isolates.

In Figure 6.18 and Table 6.2, Position A shows that there are 131 loci which differentiate P694a from the other six ST-474 isolates. Position B shows that there are 58 loci which differentiate H22082 from the other six ST-474 isolates, and position C shows that there are 38 loci which differentiate both H22082 and P694a from the other five ST-474 isolates. Position D shows that there are 10 loci which differentiate P179a from the other six ST-474 isolates, while position E shows that there are 10 loci which differentiate H569a from the other six ST-474 isolates. Position F shows that there are 4 loci which differentiate both H569a and P179 from the other five ST-474 isolates. Position G shows that there are 21 loci which differentiate P110b from the other six ST-474 isolates. Finally, position H shows that there are 68 loci which differentiate H704 from the other six ST-474 isolates, and position I shows that there are 47 loci which differentiate H73020 from the other six ST-474 isolates.

6.3.5 Compatibility

For seven ST-474 isolates, 78 loci of the 274 loci were informative. These 78 loci were the ones that are displayed in the compatibility plot (Figure 6.19). For each pair of loci, the number of extra changes is computed. These numbers are the number of events that would be required over and above the minimum for them to fit on the same tree (this minimum is just the number of unique alleles - 1 for each loci). If this value is 0 then the two loci are compatible. After reordering the compatibility plot, it can be found that blocks of mutually compatible characters (cliques) from the set of 78 parsimony informative loci. There are four cliques, which contain 37 loci (highlighted by green color in Figure 6.19), 21 loci (yellow), 10 loci (purple) and 5 loci (grey) separately. The maximum parsimony tree has been built from these four sets of characters separately (Figures 6.20 to 6.24). There are two tied trees which share the equal number of parsimony score for clique two with 21 loci (Figures 6.21 and 6.22).

These trees (Figures 6.20 to 6.24) have few splits in common. These completely different tree topologies are important because it shows that different blocks of loci have different evolutionary histories, and these evolutionary events are not compatible with the tree reconstruction. Furthermore, these genes are not contiguous with respect to the genome.

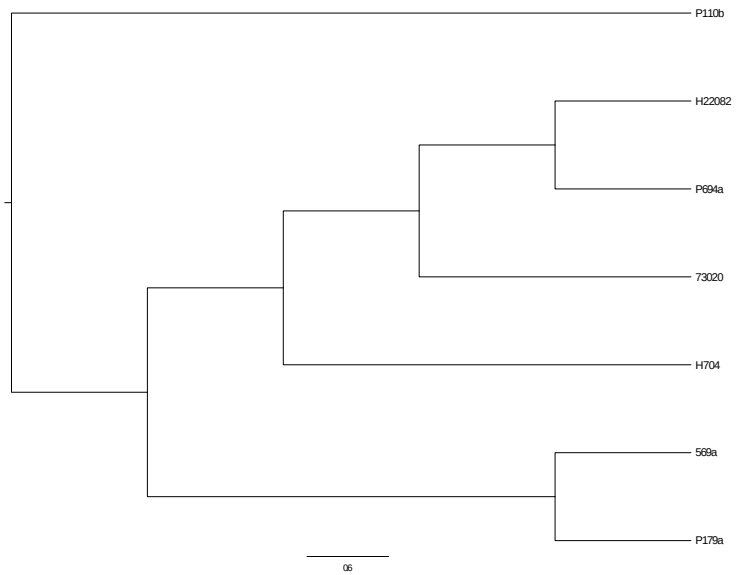


Figure 6.20: Maximum parsimony tree for for cluster one which contains 37 loci.

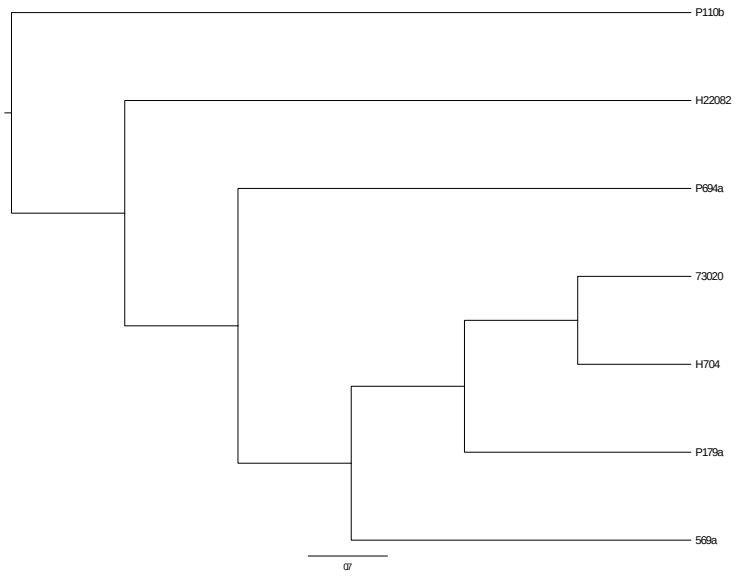


Figure 6.21: Maximum parsimony tree for cluster two which contains 27 loci (tree 1).

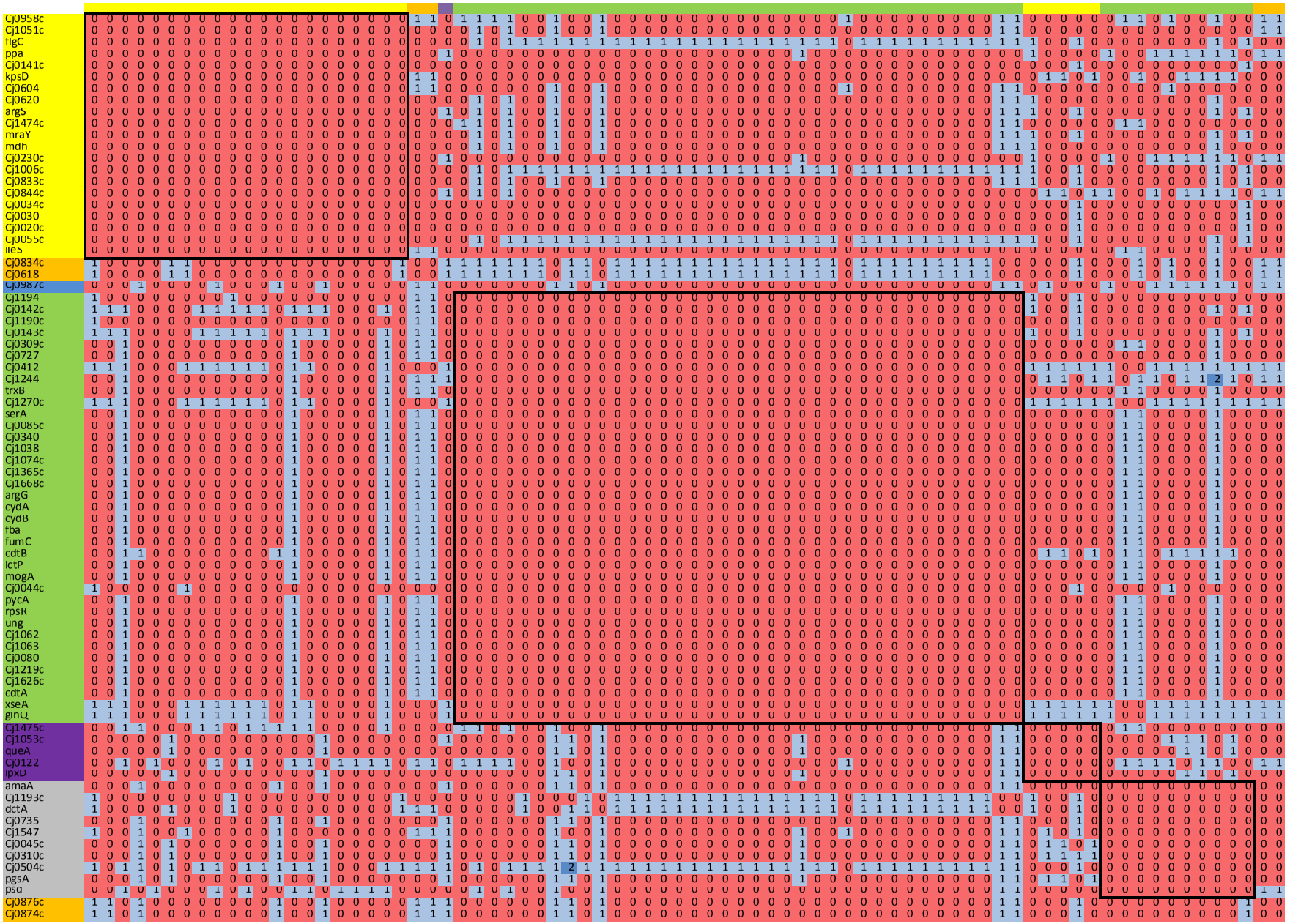


Figure 6.19: These numbers are the number of events that would be required over and above the minimum for them to fit on the same tree. If this value is 0 then the two loci are compatible, else this pair is incompatible.

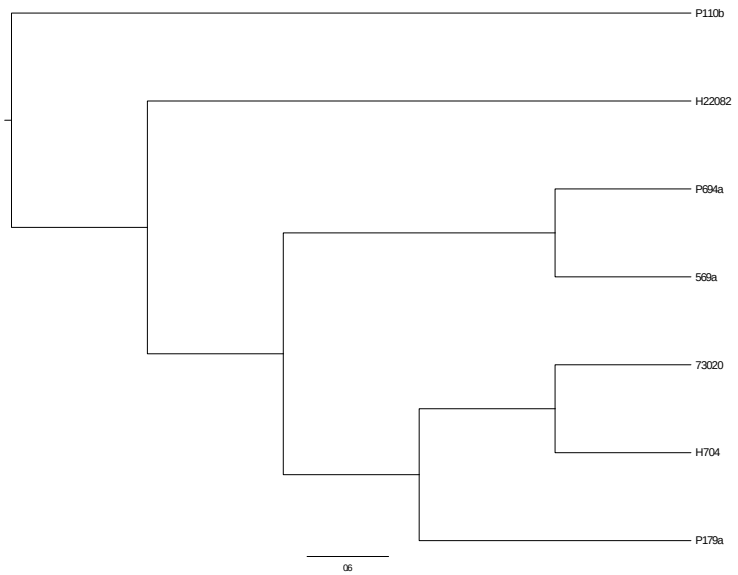


Figure 6.22: Maximum parsimony tree for cluster two which contains 27 loci (tree 2).

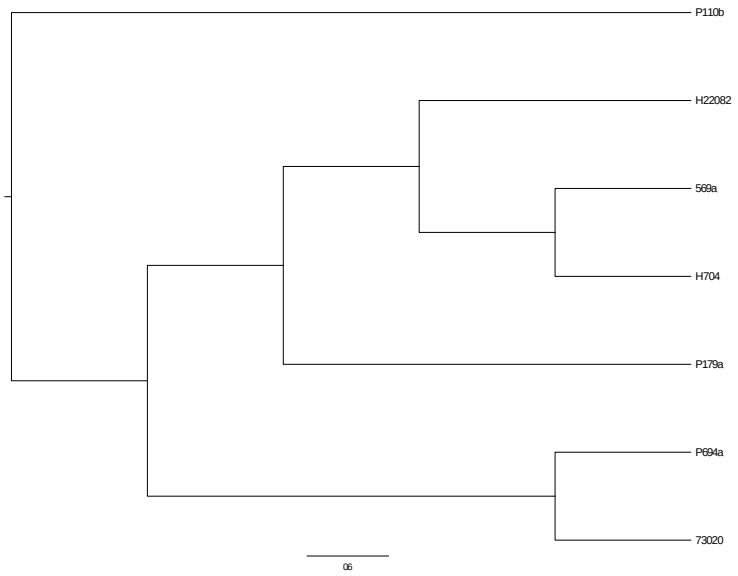


Figure 6.23: Maximum parsimony tree for cluster three which contains 10 loci.

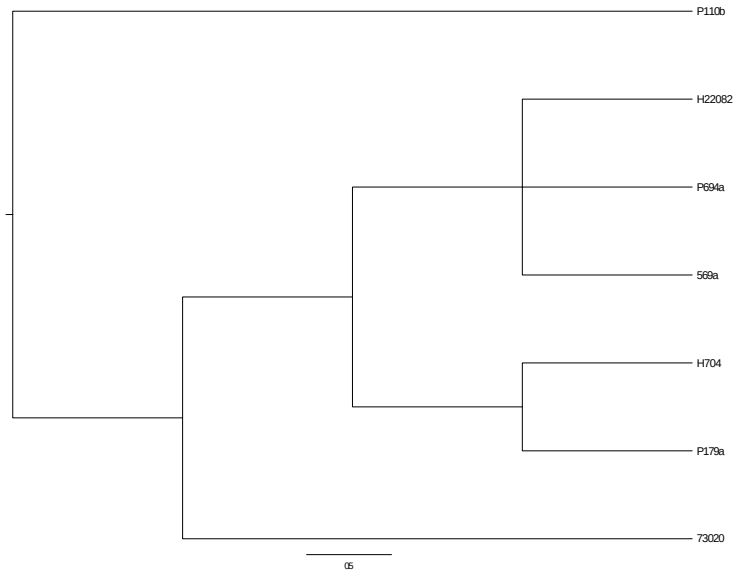


Figure 6.24: Maximum parsimony tree for cluster four which contains five loci.

6.4 Discussion

One goal of this study was to compare the application of a range of phylogenetic and population genetic tools, such as UPGMA, NJ, MP, ML, and Bayesian methods, while investigating the evolutionary genealogy of 59 *Campylobacter* whole genomes, with the focus on estimating the evolutionary clonal genealogy for the New Zealand-associated *Campylobacter* strain type 474 (ST-474). The divergence position for ST-474 is at the shallower part of all the reconstructed trees. This means that the divergence of ST-474 from ST-48 is a recent event, and this is consistent with the effect of geographical isolation discussed in Chapter 5.

The aim of mapping events onto branches of a tree topology was to identify genes which are compatible to a given tree, and further to identify the events that lead to the observed phylogeny of extant ST-474. Based on the results from the simulation study, it can be seen that all the applied phylogenetic methods have difficulty in accurately reconstructing the shallower part of the phylogeny. This analysis assumes that ST-474 share synteny with the reference genome (NCTC 11168). In addition, the phylogenies resulting from a range of phylogenetic methods on both the TGRS and MLST data sets differ, so it is very difficult to define one phylogeny/cladogram of ST-474. Therefore, due to its efficiency, the maximum parsimony tree was chosen to further investigate ST-474.

The applied analyses suggest that the phylogenetic analysis on concatenated sequences may not be the most appropriate way to reconstruct a phylogeny which reflects the evolution of *Campylobacter*. This is also supported by the largely different tree topologies inferred from four cliques for seven ST-474 isolates. Further

studies may focus on a subset of TGRS and their association with different selection pressures, as some quantitative analysis methods have now available [59, 273, 355]. For example, studies based on subsets of genes which are associated with common functions may provide an opportunity to build on the conclusions about the clonal relationships of ST-474 made from this research.

It can be seen that there are no apparent splits in common among the four cliques (Figures 6.20 to 6.24). This observation is consistent with the large amount of recombination that occurred within ST-474. The large occurrence of recombination is consistent with previous publications [25, 99, 335, 414]. These incompatible pairs are more likely to be introduced by recombination rather than mutation, because the analysis is carried out at the allelic profile level rather than nucleotide level. At the allelic profile level, the situation similar to parallel change or homoplasy, is unlikely to occur. This is because it would require exactly same mutation events to occur multiple times across several branch/isolates, and at the same time, for all the other nucleotide sites in this loci to stay unchanged.

In addition, the gene coordinates for each loci show that the loci in one clique are not close together on the reference genome. The large separation for the the gene coordinates in one clique is different from the observation in *Helicobacter pylori* [217]. For *Helicobacter pylori*, it has been found that the imported DNA of a donor can be interrupted by small fragment of sequences of the recipient [217]. One possible explanation of the separated gene coordinates within each cliques is that the DNA secondary structure could cause the similar evolutionary events occur among different parts of the genome [36, 162].

Whole genome analyses on fewer strain types has also been carried out [25, 77, 118, 131, 286]. Much research has also been done on the evolution of *Campylobacter* using different combinations of genes, in particular the MLST genes [237, 333, 414]. Compared to MLST datasets, the analyses with TGRS contain much longer sequences and fewer isolates. Consequently, the analysis of the TGRS provides more complete information on the evolution of *Campylobacter*.

The phylogeny results inferred through TGRS show that all the isolates containing the same ST are grouped together. This observation is consistent with the previous research which shows that seven MLST loci can be used to differentiate isolates and investigate evolutionary relationships among bacteria [116]. However, when comparing phylogenies inferred by TGRS and MLST, it can be seen that the topologies inferred by TGRS are quite different from those of MLST, although the same phylogenetic methods (UPGMA, NJ, MP, ML, or CF) were applied. This observation shows that different sets of genes have different evolutionary histories, and this is also supported by the different trees produced by the four cliques.

There are some common features inferred by different phylogenetic methods using the TGRS data. For example, *C. coli* and seven ST-474 isolates were always grouped together. All of the applied phylogenetic methods on TGRS data differentiate *C.* sp. nov. from *C. jejuni* and *C. coli*. For both MLST and TGRS, the different phylogenetic methods differentiated *C. jejuni* and *C. coli* and group ST-474 and ST48 together.

The different phylogenetic methods inferred different phylogenetic relationships for the 59 *Campylobacter* isolates on TGRS, and even with MLST data, different phylogenetic relationships were observed by different phylogenetic methods. The main reason for this is that different phylogenetic methods have different strengths and weaknesses.

Distance-based methods, such as UPGMA and NJ, rely on the distance matrix, which is calculated from pairwise nucleotide differences. However, the distance matrix ignores some information [291, 294], for example, different sets of sequences can have the same matrix [291]. The strengths of distance-based models are that they have lower computational requirements and can be applied to a wider range of types of data. However, missing data can affect the estimations of distance. In terms of the assumptions, UPGMA assumes a strict molecular clock in which all lineages should have a constant evolutionary rate, whereas NJ relaxes this assumption [47]. Therefore, compared to UPGMA, NJ is more robust.

Character-based methods, including MP, ML, and Bayesian methods, use calculations from the alignment of several sequences by comparing nucleotide sites in the alignment. The strengths of MP are its computational simplicity and the ease in interpreting the results. When investigating the evolutionary process, ML has some advantages over NJ and MP if the evolutionary model has been chosen correctly. ML produces the phylogeny that gives the largest probability of observed sequence data, based on the chosen substitution model and tree.

In terms of computational requirements, NJ and MP have a great advantage over ML and Bayesian methods, for which the simulated data requires 76 hours on a desktop computer for one run of 10,000 iterations (including 5000 burn-in iterations) in the ClonalFrame program. The TGRS requires at least 240 hours on a desktop computer for one run with the same setting in the ClonalFrame program. In contrast, NJ and MP require less than one hour on a desktop computer to be completed.

In general statistics, consistency and efficiency are the two main statistical properties to be considered when comparing different estimation methods to enable accurate inferences. Consistency in statistics [76, 380] means the estimator essentially converges to the underlying true value as the amount of data grows to infinity. In phylogeny, the consistency of a phylogeny reconstruction [108] means it converges

to the underlying true evolutionary phylogeny when the number of analysed nucleotide sites approaches infinity. NJ, ML, and Bayesian methods are all model-based methods. They are consistent when the chosen model is the correct underlying model [411], though the correct underlying model is difficult to guarantee.

Efficiency in statistics [76, 258] is a measurement of the optimal unbiased estimator. An efficient estimator means fewer samples are needed to achieve a given statistical power. In phylogeny, the efficiency measures the probability of reconstructing the correct phylogeny under a fixed number of nucleotide sites [411]. Previous research [112, 150] has demonstrated that ML has higher efficiency when compared to NJ and MP in terms of reconstructing the correct phylogeny.

As well as, accuracy is also important to statistical analysis. Accuracy can be measured by several methods, such as simulation [161, 174], comparing to known phylogeny and statistical tests [161]. The inference of phylogeny can be viewed as a combination of selected methods and sequences [364], so its the accuracy can be affected by selection of sequences and phylogenetic analysis. The selection of sequences/alignments are more important than the choice of phylogenetic analysis [218]. The analyses carried out in this research on TGRS and MLST across different methods also support this view, because there are more differences between results from different data than between results from different methods. Most current programs can only handle limited numbers of sequences, especially for ML and Bayesian analysis [364].

The underlying model assumptions made by CF are that recombination events introduce a constant rate of nucleotide changes to the contiguous regions of a sequence [69]. This method outperformed existing methods when applied to *Salmonella* and *Bacillus* MLST data [69], but for *Campylobacter*, this method faced difficulty as it did not converge and consequently it is hard to make inferences from this method's results. Moreover, even when convergence is attained for the simulated dataset, the largest number of incorrect splits was obtained for CF, as shown by symmetric-difference distances (Table 6.1). This is a very surprising result, given that CF is the only method with an appropriate model.

Inconsistent results are observed in the CF results (Figure 6.11 and Figure 6.12). Figure 6.11 shows that the main difference is the divergence time of *C.* sp. nov.. One explanation for this is that large recombination events on the *C.* sp. nov. genome result in different phylogenetic relationships from different genes.

The analysis in this chapter also reveals some contrasting branches of the evolution of *Campylobacter*. Three main phylogenetic groups were defined by UPGMA, NJ, MP, and ML for *C.* sp. nov., *C. jejuni* and *C. coli*. All the applied analyses show *C.* sp. nov. diverged earlier than *C. jejuni* and *C. coli*. Some results from CF

show the relationship of *C.* sp. nov. is closer to *C. coli*. This could be due to the selected regions of genomes or selected phylogenetic analyses. New approaches and selected alignment sets can benefit more detailed examination of the phylogeny of *Campylobacter*.

The aims of the analyses in Chapter 6 were: (1) comparing a range of phylogenetic tools, including UPGMA, NJ, MP, ML, and CF methods, to investigate evolutionary genealogy based on 59 *Campylobacter* whole genomes, with the focus on estimating the evolutionary clonal genealogy for New Zealand specific *Campylobacter* strain type 474 (ST-474); and (2) mapping events on branch and computing the compatibility plot is to locate the genes which contains information about the clonal phylogeny of ST-474.

The same phylogenetic methods on different datasets had different results. Furthermore, the different phylogenetics methods applied on for the same dataset, including a simulated data, MLST, and TGRS, have inferred the different tree topologies. Further development of methods that capture and model recombination, such as ClonalFrame is required.

The different phylogenies inferred by TGRS and MLST even using the same phylogenetic methods and at least four different phylogenies inferred from seven ST-474 isolates show different combinations of genes have different evolutionary histories, and concatenated sequence is not the most appropriate choice for phylogeny inferencing of *Campylobacter*, in particular TGRS data. Four cliques have been located for further investigation.

Phylogenetic studies on MLST and TGRS of 59 isolates have offered important insights into genome evolution of *Campylobacter* and shed new light on the origin of ST-474. The phylogenetic information of ST-474 is inferred based on the increasing abundance of molecular data. Despite different reconstructed phylogenies, evidence across different assumptions and methods suggests: (1) *C.* sp. nov. is a separate species of *Campylobacter*; and (2) Phylogeny analyses on concatenated sequences can reflect the major properties of clonal evolution of *Campylobacter*; and (3) ST-474 isolates are recent expansion strain types in the evolution of *Campylobacter*; and (4) Both MLST and TGRS scale data can distinguish *C. jejuni* and *C. coli* clearly.

6.5 Acknowledgements

I acknowledge Dr Barbara R. Holland for sharing her knowledge and expertise in phylogenetics and providing Python scripts to locate the incompatibility loci and produce compatibility plot, Prof Nigel French and Prof Paul Fearnhead for all the

guidance and suggestions, and Dr Patrick Biggs for the datasets. We acknowledge the Marsden Fund project 08-MAU-099 (Cows, starlings and *Campylobacter* in New Zealand: unifying phylogeny, genealogy and epidemiology to gain insight into pathogen evolution) for funding this project. BRH acknowledges the Australian Research Council (grant FT100100031).

6.6 Appendix A: Variant loci on phylogeny of ST-474

Position Name	Variants appeared on that position
A position (P694a)	*porA, Cj0339, Cj0497, Cj0737, Cj0741, Cj1257c, dnaJ, grpE, ppa, Cj0251c, Cj0454c, Cj0455c, Cj0607, Cj0610c, Cj0611c, Cj0619, Cj0967, Cj1442c, Cj1516, Cj1724c, Cj0685c, Cj0982c, dsbB, gatA, gidA, hypB, lpxD, priA, neuB3, pstA, pta, Cj0411, Cj1069, Cj1295, Cj1305c, Cj1377c, Cj1548c, accC, amaA, gpsA, waaE, Cj1337, pstB, pyrC2, queA, ruvB, Cj0230c, acnB, ispG, pabB, Cj1068, pgsA, rdxA, rpsF, ssb, gatB, sdaA, flaA, Cj0045c, Cj0621, Cj1298, Cj1506c, sucC, Cj0022c, Cj0038c, Cj0247c, cfrA, murC, Cj1170c, Cj1294, Cj1407c, Cj1051c, waaC, Cj0141c, Cj0265c, Cj0310c, Cj0456c, Cj0457c, Cj0563, Cj0620, Cj0728, Cj0735, Cj1053c, Cj1174, Cj1296, Cj1728c, argS, cft, Cj0368c, corA, flgH, glyS, livK, mraY, nrdF, pabA, wlaB, pnp, pstC, pstS, ruvC, Cj0184c, Cj1319, lysS, Cj0832c, Cj0833c, Cj0837c, Cj0843c, Cj0844c, Cj0846, Cj0848c, Cj0849c, Cj0850c, Cj0864, fliR, folD, hemL, murA, ogt, psd, xerD, Cj1618c, Cj1664, Cj1665,*

Position Name	Variants appeared on that position
A position (P694a) cont	*Cj1666c, cadF, chuA, lepP, Cj0458c, pbpB, serB, Cj0617, Cj0020c, Cj0030, Cj0034c, Cj0036, Cj0609c, Cj1005c, Cj1658, dnaX, mdh, Cj1431c,*
B position (H22082)	*porA, Cj0339, Cj0497, Cj0737, Cj0741, Cj1257c, dnaJ, grpE, ppa, Cj0251c, Cj0454c, Cj0455c, Cj0607, Cj0610c, Cj0611c, Cj0619, Cj0967, Cj1442c, Cj1516, Cj1724c, Cj0685c, Cj0982c, dsbB, gatA, gidA, hypB, lpxD, priA, neuB3, pstA, pta, Cj0411, Cj1069, Cj1295, Cj1305c, Cj1377c, Cj1548c, accC, amaA, gpsA, waaE, Cj1337, pstB, pyrC2, queA, ruvB, Cj0230c, acnB, ispG, pabB, Cj1068, pgsA, rdxA, rpsF, ssb, gatB, sdaA, flaA, Cj0045c, Cj0621, Cj1298, Cj1506c, sucC, Cj0022c, Cj0038c, Cj0247c, cfrA, murC, Cj1170c, Cj1294, Cj0495, Cj0496, Cj0930, Cj0958c, Cj1258, feoB, hypF, kdpB, napA, Cj0888c, Cj0889c, flaD, flhA, Cj0044c, Cj0604, Cj1056c, Cj1474c, Cj1194, Cj1256c, Cj1190c, Cj0170, fabD, murE, pfs, pheA, slyD, uvrA,*

Position Name	Variants appeared on that position
C position (P694a&H22082)	*Cj0727, Cj0724, Cj0019c, Cj0085c, Cj0340, Cj1038, Cj1074c, Cj1365c, Cj1668c, argG, cydA, cydB, fba, fumC, ileS, lctP, lon, mogA, pycA, rpsR, sdaC, ung, Cj1062, Cj1063, Cj0080, Cj1219c, Cj1626c, cdtA, cdtB, serA, trxB, Cj0309c, Cj1244, kpsD,*
D position (P179a)	*Cj0724, Cj0309c, Cj0339, Cj0737, Cj0741, Cj0619, Cj0967, Cj1442c, Cj1516, Cj1724c, Cj0685c, dsbB, gidA, neuB3, pstA, pta, Cj0411, Cj1295, Cj1305c, Cj1377c, Cj1548c, accC, amaA, waaE, Cj1337, pstB, acnB, ispG, flaA, Cj0621, Cj1298, Cj0022c, Cj0038c, cfrA, murC, Cj1170c, Cj1294, hypF, waaC, Cj0265c, Cj0735, Cj1296, mraY, Cj1319, lysS, Cj1028c, Cj0055c,*
E position (H569a)	*Cj0724, Cj0737, Cj0741, Cj0619, Cj1442c, Cj0685c, dsbB, gidA, neuB3, pstA, Cj0411, Cj1295, Cj1305c, Cj1377c, Cj1548c, accC, waaE, Cj1337, pstB, flaA, Cj0621, Cj0022c, Cj0038c, cfrA, murC, Cj1170c, Cj1294, Cj1319, ileS, Cj0454c, Cj0610c, Cj0247c, Cj1474c, Cj0620, Cj1728c, livK, Cj1666c, Cj0617, dnaX, Cj1007c, tig,*

Position Name	Variants appeared on that position
F position (P179a&H569a)	*Cj0724, Cj0737, Cj0741, Cj0619, Cj1442c, Cj0685c, dsbB, gidA, neuB3, pstA, Cj0411, Cj1295, Cj1305c, Cj1377c, Cj1548c, accC, waaE, Cj1337, pstB, flaA, Cj0621, Cj0022c, Cj0038c, cfrA, murC, Cj1170c, Cj1294, Cj1319, ileS, Cj0454c, Cj0610c, Cj0247c, Cj1474c, Cj0620, Cj1728c, livK, Cj1666c, Cj0617, dnaX, Cj1007c, tig, Cj0309c, Cj0339, Cj0967, Cj1516, Cj1724c, pta, amaA, acnB, ispG, Cj1298, hypF, waaC, Cj0265c, Cj0735, Cj1296, mraY, lysS, Cj1028c, Cj0055c, ppa, lpxD, queA, Cj0230c, pgsA, Cj0045c, Cj0958c, Cj0604, Cj1051c, Cj0310c, Cj1053c, psd, Cj0122, Cj1547, thyX,*
G position (P110b)	*Cj0724, Cj0737, Cj0741, Cj0619, Cj1442c, Cj0685c, dsbB, gidA, neuB3, pstA, Cj0411, Cj1295, Cj1305c, Cj1377c, Cj1548c, accC, waaE, Cj1337, pstB, flaA, Cj0621, Cj0022c, Cj0038c, cfrA, murC, Cj1170c, Cj1294, ileS, Cj0247c, Cj0617, Cj0339, Cj0967, Cj1724c, pta, acnB, hypF, Cj0265c, Cj1296, Cj1028c, lpxD, queA, Cj0604, Cj1051c, kpsD, porA, grpE, Cj0611c, Cj0982c, gatA, hypB, priA, rdxA, Cj1506c, sucC, Cj0044c, Cj1190c, pbpB, Cj1658, Cj1588c,*

Position Name	Variants appeared on that position
H position (H704)	*Cj0019c, Cj0724, Cj0737, Cj0741, Cj0619, Cj1442c, Cj0685c, gidA, neuB3, pstA, Cj0411, Cj1295, Cj1305c, Cj1548c, accC, waaE, Cj1337, pstB, flaA, Cj0621, Cj0038c, cfrA, murC, Cj1170c, Cj1294, Cj0247c, hypF, Cj0265c, Cj1028c, Cj1051c, porA, grpE, Cj0611c, hypB, priA, Cj1506c, sucC, pbpB, Cj1658, Cj1588c, Cj0454c, Cj1474c, Cj0620, dnaX, Cj1007c, tig, Cj1516, Cj1298, mraY, Cj0055c, psd, thyX, sdaC, Cj0607, sdaA, uvrA, cft, pstC, pstS, Cj0833c, Cj0020c, Cj0030, Cj0034c, Cj1005c, mdh, Cj1193c, Cj1475c, tonB2, Cj0025c, Cj0035c, Cj0198c, Cj0462, Cj0708, Cj0771c, Cj0800c, Cj1006c, Cj1587c, Cj1589, carB, fliE, icd, mreB, Cj0466, recJ, secA, sucD, Cj0229, Cj0373, Cj0463, Cj0530, Cj0465c, flgB, flgC, ilvC, metG, recG, Cj1453c,*
I position (H73020)	*Cj0019c, Cj0724, Cj0737, Cj0741, Cj0619, Cj1442c, Cj0685c, gidA, neuB3, pstA, Cj0411, Cj1295, Cj1305c, Cj1548c, accC, waaE, Cj1337, pstB, flaA, Cj0621, Cj0038c, cfrA, murC, Cj1170c, Cj1294, Cj1028c, porA, hypB, priA, Cj0454c, Cj1516, Cj1298, Cj0020c, Cj0030, Cj0034c, Cj0035c, Cj0967, Cj0604, Cj0044c, lysS, Cj1547, Cj0727, pyrC2, feoB, Cj0141c, Cj0849c, Cj0036, Cj0609c, Cj1099, Cj1214c, Cj1215, Cj1543, Cj1544c, Cj1545c, Cj1546, Cj1713, cheA, clpB, glcD, hydB, kdsB, pyrD, uvrB, rbn,*

Chapter 7

Conclusion and further directions

7.1 Conclusion

Improving our understanding of the evolution of *Campylobacter* has had important implications for a wide range of areas, including epidemiological investigations [335, 332, 334, 336], and policy development to minimise the impact of emerging pathogens [327, 328]. A better understanding of the evolutionary mechanisms of *Campylobacter* can also help gain insight into the emergence of new pathogenic strains.

Generally speaking, genetic exchange plays an important role in the evolution of *Campylobacter*, and can be affected by three factors:

1. The physical mechanism [83, 120] of genetic exchange, e.g. whether mutation or recombination is more prevalent.

2. Ecological separation of subpopulations in different hosts and different geographical locations.

3. Selection [419].

In this thesis, these factors were explored using three types of analyses: (1) an analysis which worked on single locus variant (SLV) datasets; and (2) analyses which investigated the physical barrier brought by geographical isolation; and (3) the application of a range of phylogenetic analyses to investigate the clonal genealogy of *Campylobacter* strains using a targeted gene reference set (TGRS) and full genome analysis. The first factor (physical mechanism) relates to the biological function of bacteria, the second factor (ecological separation) relates to the habitats in which the bacteria lives, and the last factor (selection pressure) relates to the interaction between the adaptive ability of bacteria and their environment. The analysis of

SLVs relates to the first and third factors, and the analysis of geographical isolation relates to the second and third factors, and the analysis on targeted gene reference sets relates to all three factors.

This research has made a contribution to our understanding of the evolution of *Campylobacter* species. It has put forward a new statistical approach to calculate the relative contribution of recombination and mutation to generating SLVs (Chapter 3), and this method has been successfully applied to a wider range of pathogenic bacteria (Chapter 4). Additionally the research has described the likely evolution of *C. jejuni* and *C. coli* in the unique environment of NZ and compared their evolution in NZ and the UK (Chapter 5). The detailed investigation of the phylogeny of 59 NZ isolates, based on both MLST and TGRS regions is described in Chapter 6.

Based on the above analyses, there are two main findings. The first is that both the SLV and geographical isolation analyses reveal patterns of genetic exchange that have changed over time. The results of the SLV analyses suggest that, in general, different bacteria, such as *C. jejuni* and *C. coli*, have different rates of genetic exchange, and even for the same species of bacteria, different clades have different rates of genetic exchange. The unique distribution of STs attributable to geographical isolation show that the distribution of genetic material also varied through time, and there is some evidence of clustering by geographical region at the ST and allelic profile levels, however at the nucleotide level there is relatively stronger clustering by host association than clustering by geographical region. The second finding shows that different genes have different evolutionary histories. This is supported by all of the three main analyses (the analysis of SLV, the analysis of geographical isolation, and the analysis of TGRS). The results of the SLV analyses show that different loci have different ratios of recombination to mutation when generating SLVs. In the analysis of geographical isolation, the different estimated time of divergence of several strains demonstrated different evolutionary patterns for different lineages. The different phylogenies inferred by five different methods on TGRS imply that there are the different evolutionary histories of different genes across the entire genome. The following sections will describe some specific conclusions from each of the three analyses.

7.1.1 The analysis of SLVs

Recombination and point mutation can both generate new alleles that lead to SLVs. Estimating the relative roles of recombination and mutation helps us understand how organisms evolve. A statistical method was proposed to estimate the relative rates of those two evolutionary processes and this was applied in a comparative way to *Campylobacter* and other bacteria.

There are three major findings from this study. The first is that recombination plays a more important role than mutation in generating SLVs for *Campylobacter jejuni*. Generally, the estimated ratios of recombination events to point mutations were larger for SLVs than those estimates for pairs of more distantly related strains for *C. jejuni*. For *C. jejuni*, the probability of recombination generating a new allele that leads to an SLV was estimated to be roughly seven times more than that of mutation, whereas for the most commonly isolated clade of *C. coli*, recombination and mutation were estimated to have a similar contribution to the generation of SLVs. Furthermore, the majority of nucleotide differences between strains that make up an SLV were attributable to recombination; for *C. jejuni*, 98% of nucleotide differences between SLVs were attributable to recombination, whereas for *C. coli*, 85% of nucleotide differences were attributable to recombination. The second finding was that purifying selection may act more stringently on recombination events than on point mutations. Purifying selection is a possible explanation for why research on more distantly related STs revealed relatively lower rates of recombination [388, 403, 414]. The third finding is that the more important role of recombination to mutation is not unique to *C. jejuni* and *C. coli*. For most of the test bacteria, recombination was estimated to have played a more important role than mutation in generating SLVs.

The proposed new statistical method for estimating the relative rates of recombination and mutation is a starting point for making inferences from genetic data, such as MLST databases by the use of mathematical models. The new method can be extended to larger scale analyses (more genes). The findings in this thesis provide better evidence for the importance of recombination and mutation in the evolution, and a better understanding of the evolutionary dynamics of *Campylobacter* to gain insight into the emergence of new, potentially more virulent, strains.

7.1.2 The role of geographical isolation in the evolution of *Campylobacter*

Host association, geographical location, and human agricultural activities have all played a role in the evolution of *Campylobacter*. Previous research [190, 332, 336]

has shown that *Campylobacter* is associated with many different hosts, and the divergence of *Campylobacter* species may have been caused by human agricultural activities. This thesis investigated the effect of geographical isolation on the evolution of *Campylobacter* by utilizing the unique geographical location of NZ and comparing datasets from NZ and the UK.

There are three main findings from this study. The first is that geographical isolation played a role in the evolution of *Campylobacter*. The second finding reveals that the distribution of STs have changed over time. The evidence showed that clustering by geographical region at the ST and allelic profile levels differed from that at the nucleotide level. The role of geographical isolation is much stronger at the sequence type level than at the nucleotide level. This is evidence that the effect of geographical isolation on the evolution of *Campylobacter* diversity decreases over time, i.e. variation among the analyses from the nucleotide level, to the allelic profile level, to the sequence type level implies that the effect of geographical isolation only plays a short-lived role in the evolution and diversity of *Campylobacter* genotypes. The third finding shows that despite the short-lived effect of geographical isolation, there are some NZ specific and associated lineages of *Campylobacter* strains (e.g ST-2381 and ST-474), and evidence of some unique lineages (e.g. *C.* sp. nov.) that existed in NZ before the arrival of Polynesian settlement and the introduction of livestock, whereas some evolved uniquely in NZ as recently as a few hundred years ago.

7.1.3 Analysis on targeted gene reference sets

ST-474 [25, 259] is widely distributed in different geographical locations in New Zealand [246] and caused more than a quarter notification of human campylobacteriosis cases in NZ between 2004- 2006.

In order to compare five different phylogenetic methods, such as UPGMA, NJ,MP, ML, and CF, a simulation study was carried out. I simulated the sequences and clonal genealogy of the sequences, and then applied five phylogenetic methods to infer the clonal genealogy and compare the inferred genealogy to the true underlying genealogy from which the sequences were generated in the simulation. Through the simulation study, it can be seen that the tips of clonal genealogy are basically consistent across a range of phylogenetic methods, such as UPGMA, NJ,MP, ML, and CF, although there are some minor conflicts at the deeper branches. By comparing the results of inferred clonal genealogy using the MLST region and the TGRS region, different methods all show that ST-474 diverged from a common ancestor with ST-48. The majority of the same STs are grouped together, although the evid-

ence from part of this study suggests that the clonal genealogy is inconclusive for different phylogenetic methods. The main reason could be that different genes have different evolutionary histories, and high-frequency recombination can make the inference more complex. Furthermore, the poor performance of CF on the TGRS was presumably due to the size of the dataset making it difficult for the MCMC to converge, despite as CF being the most appropriate underlying model of all those applied.

This research will hopefully serve as a basis for future studies estimating clonal genealogy using targeted gene reference sets. The study has gone some way towards enhancing our understanding of the complexity of the clonal genealogy of bacteria and the need for improving the accuracy and efficiency of current phylogenetic tools. More information on the selection criterion for some gene sets would help to establish a greater degree of accuracy.

Different tree topologies are inferred by different phylogenetic methods. The simulation study shows that all the applied phylogenetic methods can reconstruct the deeper branches accurately, but not the shallower part of the phylogeny. In addition, further development of methods that capture and model recombination, such as ClonalFrame is required.

Different tree topologies are inferred by different subset of genes. This conclusion is supported by the different results between TGRS and MLST inferred using the same methods. Furthermore, this conclusion is also supported by the completely different tree topologies inferred by different cliques among 78 informative loci within seven ST-474 isolates.

7.2 Further directions

It is recommended that further research be undertaken in the following areas: firstly, a future study extending the SLV work for analysis on targeted gene reference sets would be very interesting. The extension of SLV work can be done by adjusting the selected criterion for the investigated sequence types and altering the approaches for building mutation and recombination models. SLV analysis begins with choosing a subset of analysed sequences. This subset contains all pairs of sequences with SLVs, and then mutation and recombination models are applied on the subset of sequences. Increasing the number of loci will reduce the chance of finding SLVs, because SLVs are defined as a pair of sequences with one and only one locus variant. The script could be adjusted to find the pairs of sequences under a custom-defined criterion, such as the pairs of sequences with fewer than n loci variants, or the pairs

of sequences with fewer than m nucleotide sites variants. Even, the criterion for selecting subset of analysed sequences could be built on more dynamic measures (like sliding windows with a flexible width). In addition, mutation and recombination models would become more complex to allow more meaningful parameters to be added and estimated.

Secondly, TGRS preliminary reformatting can be extended to whole genome datasets. Then a wide range of existing methods from MLST analysis could be adjusted for whole genome analyses. The script produced in Chapter 3 can turn the long character sequence data into an allelic profile format. The allelic profile format contains fewer characters, but this format can extract and compress the useful information. This information represents the similarities and differences among a set of isolates, i.e. whether or not the isolates are identical. The compressed format ignores minor differences to focus on a more achievable level of analysis. In addition, ribosomal multilocus sequence typing (rMLST) [186] has been proposed recently to provide a universal approach to the classification of bacteria from domain to strain. This method integrates microbial genealogy and typing to efficiently identify gene variants [186]. A web-accessible database has been integrated in PubMLST. This method produces the similar format of data to MLST, although there are no STs assigned. SLV or a similar analysis can then be applied to the the rMLST database.

Thirdly, a further study could apply BEAST analysis to more selected genes across the whole genomes. The selected genes from a set of isolates can contain different evolutionary processes, but should not contain recombinant regions, because BEAST is not designed to cope with recombination. Compared to other Bayesian analytical software, BEAST can estimate the time of divergence of given strain types. START2 software can be applied to test whether the recombination exists. The existence of recombination and the average length of recombination are crucial questions for many biological analyses. Around two decades ago, the diversity of sequences was used to decide the existence of selection pressure [252, 47]. If the diversity of sequences is small, then selection pressure plays a role in the evolution of the analysed sequences. But later research [224] showed the presence or absence of recombination disrupted the simple relationship; only when recombination occurred frequently did the simple relationship hold. If recombination is infrequent, lower diversity could be a consequence of genetic hitch-hiking on a gene under positive selection.

Furthermore, more research is needed to better understand evolution using whole genome analyses. This analysis can help us understand how the bacterial pathogens evolve to infect human beings. *C.* sp. nov. is a very interesting species. So far, it has only been isolated from takahe and weka, two native bird species in NZ. Several variants of *C.* sp. nov. have been found at the date of writing. The genome

size for the proposed species is 2.1-2.2 million basepairs (2 Mbp), which is larger than the pathogenic species, *C. jejuni* and *C. coli* (1.6 Mbp). This is interesting because *C. jejuni* and *C. coli* are a major cause of acute gastroenteritis in humans, and together, *C. jejuni* and *C. coli* are responsible for approximately 95% of human infections caused by the *Campylobacter* species (around 32 species and 13 subspecies identified so far). Research in this thesis show the proposed species (*C.* sp. nov.) to be the closest relative species to *C. jejuni* and *C. coli*, when compared to other members of the *Campylobacteraceae* family.

Thus, one hypothesis for further study of the proposed species (*C.* sp. nov.) is that loss of genes (0.4 Mbp) is associated with pathogenicity in humans. This is counter to the notion that highly adapted species, such as multi-host pathogens, contain more genes than their ancestors [236, 304, 338], could be an area which deserves more attention.

A better understanding of pathogen phylogenomics will be helpful in designing disease control and intervention to reduce the impact caused by emerging pathogens. In this thesis, the findings have furthered our understanding of the role of recombination, mutation, selection, and geographical isolation in the evolution of *Campylobacter*. Based on the findings in this thesis, the directions listed in this chapter could lead to more insight into the evolution of this prominent human pathogen and potential emergence of new strains in *Campylobacter* and other emerging pathogens. These further analyses will be useful in understanding the evolution and emergence of other globally important human pathogens and emerging infectious diseases.

Bibliography

[1] Aarts, H., Lith, L., & Jacobs-Reitsma, W. (2008). Discrepancy between penner serotyping and polymerase chain reaction fingerprinting of *Campylobacter* isolated from poultry and other animal sources. *Letters in Applied Microbiology*, 20(6), 371–374.

[2] Achtman, M., Zurth, K., Morelli, G., Torrea, G., Guiyoule, A., & Carniel, E. (1999). Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis. *Proceedings of the National Academy of Sciences of the United States of America*, 96(24), 14043.

[3] Adak, G., Cowden, J., Nicholas, S., & Evans, H. (1995). The Public Health Laboratory Service national case-control study of primary indigenous sporadic cases of *Campylobacter* infection. *Epidemiology and Infection*, 115(1), 15.

[4] Adak, G., Meakins, S., Yip, H., Lopman, B., & O'Brien, S. (2005). Disease risks from foods, England and Wales, 1996–2000. *Emerging Infectious Diseases*, 11(3), 365–72.

[5] Adams, W., Deaver, K., Cochi, S., Plikaytis, B., Zell, E., Broome, C., Wenger, J., Stephens, D., Farley, M., Harvey, C., et al. (1993). Decline of childhood *Haemophilus influenzae* type b (hib) disease in the hib vaccine era. *JAMA: the Journal of the American Medical Association*, 269(2), 221–226.

[6] Allos, B. (1997). Association between *Campylobacter* infection and Guillain-Barré syndrome. *The Journal of Infectious Diseases*, 176(S2), 125–128.

[7] Allos, B. (2001). *Campylobacter jejuni* infections: update on emerging issues and trends. *Clinical Infectious Diseases*, 32, 1201–1206.

[8] Alm, R., Guerry, P., Power, M., Lior, H., & Trust, T. (1991). Analysis of the role of flagella in the heat-labile Lior serotyping scheme of thermophilic *Campylobacters* by mutant allele exchange. *Journal of Clinical Microbiology*, 29(11), 2438.

[9] Ashford, W., Adams, H., Johnson, S., Thornsberry, C., Potts, D., English, J., Biddle, J., & Jaffe, H. (1981). Spectinomycin-resistant penicillinase-producing *Neisseria gonorrhoeae*. *The Lancet*, 318(8254), 1035–1037.

[10] Baker, M., Ball, A., Devane, M., Garrett, N., Gilpin, B., Hudson, A., Klena, J., Nicol, C., Savill, M., Scholes, P., et al. (2002). Potential transmission routes of *Campylobacter* from environment to humans. *New Zealand Ministry of Health report*, FW0246.

[11] Baker, M., Sneyd, E., & Wilson, N. (2007a). Is the major increase in notified campylobacteriosis in New Zealand real? *Epidemiology and Infection*, 135(01), 163–170.

[12] Baker, M., Wilson, N., Edwards, R., & Lecturers, S. (2007b). *Campylobacter* infection and chicken: an update on New Zealand's largest 'common source outbreak'. *Journal of the New Zealand Medical Association*, 120, 1261.

[13] Baker, M., Wilson, N., Ikram, R., Chambers, S., Shoemack, P., & Cook, G. (2006). Regulation of chicken contamination urgently needed to control New Zealand's serious campylobacteriosis epidemic. *New Zealand Medical Journal*, 119.

[14] Bandelt, H. & Dress, A. (1992). A canonical decomposition theory for metrics on a finite set. *Advances in Mathematics*, 92(1), 47–105.

[15] Bandelt, H., Forster, P., Sykes, B., & Richards, M. (1995). Mitochondrial portraits of human populations using median networks. *Genetics*, 141(2), 743.

[16] Barr, C., Schulman, K., Iacuzio, D., & Bradley, J. (2007). Effect of oseltamivir on the risk of pneumonia and use of health care services in children with clinically diagnosed influenza. *Current Medical Research and Opinion*, 23(3), 523–531.

[17] Barrow, P., Huggins, M., & Lovell, M. (1994). Host specificity of *Salmonella* infection in chickens and mice is expressed in vivo primarily at the level of the reticuloendothelial system. *Infection and Immunity*, 62(10), 4602.

[18] Basic-Hammer, N., Vogel, V., Basset, P., & Blanc, D. (2010). Impact of recombination on genetic variability within *Staphylococcus aureus* clonal complexes. *Infection, Genetics and Evolution*, 10(7), 1117–1123.

[19] Bates, C., Hiett, K., & Stern, N. (2004). Relationship of *Campylobacter* isolated from poultry and from darkling beetles in New Zealand. *Journal Information*, 48(1).

[20] Beauchamp, A. & Worthy, T. (1988). Decline in distribution of the takahe *Porphyrio (= Notornis) mantelli*: a re-examination. *Journal of the Royal Society of New Zealand*, 18(1), 103–118.

[21] Beery, J., Hugdahl, M., & Doyle, M. (1988). Colonization of gastrointestinal tracts of chicks by *Campylobacter jejuni*. *Applied and Environmental Microbiology*, 54(10), 2365.

[22] Bennett, J., Griffiths, D., McCarthy, N., Sleeman, K., Jolley, K., Crook, D., & Maiden, M. (2005). Genetic diversity and carriage dynamics of *Neisseria lactamica*in infants. *Infection and Immunity*, 73(4), 2424–2432.

[23] Bennett, J., Jolley, K., Earle, S., Corton, C., Bentley, S., Parkhill, J., & Maiden, M. (2012). A genomic approach to bacterial taxonomy: an examination and proposed reclassification of species within the genus *Neisseria*. *Microbiology*, 158(Pt 6), 1570–1580.

[24] Berman, S. (1991). Epidemiology of acute respiratory infections in children of developing countries. *Review of Infectious Diseases*, 13(Supplement 6), S454–S462.

[25] Biggs, P., Fearnhead, P., Hotter, G., Mohan, V., Collins-Emerson, J., Kwan, E., Besser, T., Cookson, A., Carter, P., & French, N. (2011). Whole-genome comparison of two *Campylobacter jejuni* isolates of the same sequence type reveals multiple loci of different ancestral lineage. *PloS One*, 6(11), e27121.

[26] Birkenhead, D., Hawkey, P., Heritage, J., Gascoyne-Binzi, D., & Kite, P. (2008). PCR for the detection and typing of *Campylobacters*. *Letters in Applied Microbiology*, 17(5), 235–237.

[27] Bisgard, K., Kao, A., Leake, J., Strebel, P., Perkins, B., & Wharton, M. (1998). *Haemophilus influenzae* invasive disease in the United States, 1994-1995: near disappearance of a vaccine-preventable childhood disease. *Emerging Infectious Diseases*, 4(2), 229.

[28] Bisno, A. & Stevens, D. (1996). Streptococcal infections of skin and soft tissues. *New England Journal of Medicine*, 334(4), 240–246.

[29] Black, R., Levine, M., Clements, M., Hughes, T., & Blaser, M. (1988). Experimental *Campylobacter jejuni* infection in humans. *Journal of Infectious Diseases*, 157(3), 472–479.

[30] Blakebrough, I., Greenwood, B., Whittle, H., Bradley, A., & Gilles, H. (1982). The epidemiology of infections due to *Neisseria meningitidis* and *Neisseria*

lactamica in a northern Nigerian community. *Journal of Infectious Diseases*, 146(5), 626–637.

[31] Bottone, E. (2010). *Bacillus cereus*, a volatile human pathogen. *Clinical Microbiology Reviews*, 23(2), 382–398.

[32] Bozdogan, H. (1987). Model selection and Akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3), 345–370.

[33] Bruno, W., Socci, N., & Halpern, A. (2000). Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Molecular Biology and Evolution*, 17(1), 189.

[34] Bryant, D. & Moulton, V. (2004). Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, 21(2), 255.

[35] Buchanan, S. (1999). Beta-barrel proteins from bacterial outer membranes: structure, function and refolding. *Current Opinion in Structural Biology*, 9(4), 455–461.

[36] Buckler, E., Ippolito, A., & Holtsford, T. (1997). The evolution of ribosomal DNA divergent paralogues and phylogenetic implications. *Genetics*, 145(3), 821–832.

[37] Buzby, J., Allos, B., & Roberts, T. (1997). The economic burden of *Campylobacter*-associated Guillain-Barré syndrome. *Journal of Infectious Diseases*, 176(Supplement 2), S192.

[38] Carter, P., Collins-Emerson, J., Midwinter, A., Cookson, A., Biggs, P., & French, N. (2011). A new *Campylobacter* species identified in water and wild birds (rallidae) in New Zealand. *16th International Workshop on Campylobacter, Helicobacter and Related Organisms. Vancouver, Canada.*

[39] Carter, P., McTavish, S., Brooks, H., Campbell, D., Collins-Emerson, J., Midwinter, A., & French, N. (2009). Novel clonal complexes with an unknown animal reservoir dominate *Campylobacter jejuni* isolates from river water in New Zealand. *Applied and Environmental Microbiology*, 75(19), 6038–6046.

[40] Caughley, G. (1988). The colonisation of New Zealand by the Polynesians. *Journal of the Royal Society of New Zealand*, 18(3), 245–270.

[41] Cenci-Goga, B., Karama, M., Rossitto, P., Morgante, R., & Cullor, J. (2003). Enterotoxin production by *Staphylococcus aureus* isolated from mastitic cows. *Journal of Food Protection*, 66(9), 1693–1696.

[42] Chang, N. & Taylor, D. (1990). Use of pulsed-field agarose gel electrophoresis to size genomes of campylobacter species and to construct a sali map of campylobacter jejuni ua580. *Journal of Bacteriology*, 172(9), 5211–5217.

[43] Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, (pp. 265–270).

[44] Chao, A., Chazdon, R., Colwell, R., & Shen, T. (2004). A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecology Letters*, 8(2), 148–159.

[45] Chao, A. & Shen, T. (2010). Spade (species prediction and diversity estimation). *Program and User's Guide. Available at http://chao. stat. nthu. edu. tw. Accessed*, 2.

[46] Chao, A., Shen, T., & Hwang, W. (2006). Application of laplace's boundary-mode approximations to estimate species and shared species richness. *Australian & New Zealand Journal of Statistics*, 48(2), 117–128.

[47] Chaudhuri, R. & Henderson, I. (2012). The evolution of the *Escherichia coli* phylogeny. *Infection, Genetics and Evolution*.

[48] Clark, C., Beeston, A., Bryden, L., Wang, G., Barton, C., Cuff, W., Gilmour, M., & Ng, L. (2007). Phylogenetic relationships of *Campylobacter jejuni* based on pora sequences. *Canadian Journal of Microbiology*, 53(1), 27–38.

[49] Clark, C., Bryden, L., Cuff, W., Johnson, P., Jamieson, F., Ciebin, B., & Wang, G. (2005). Use of the Oxford multilocus sequence typing protocol and sequencing of the flagellin short variable region to characterize isolates from a large outbreak of waterborne *Campylobacter* sp. strains in Walkerton, Ontario, Canada. *Journal of Clinical Microbiology*, 43(5), 2080.

[50] Cochi, S., Fleming, D., Hightower, A., Limpakarnjanarat, K., Facklam, R., David Smith, J., Keith Sikes, R., & Broome, C. (1986). Primary invasive *Haemophilus influenzae* type b disease: A population-based assessment of risk factors. *The Journal of pediatrics*, 108(6), 887–896.

[51] Cockerham, C. (1969). Variance of gene frequencies. *Evolution*, (pp. 72–84).

[52] Coen, P., Cartwright, K., & Stuart, J. (2000). Mathematical modelling of infection and disease due to *Neisseria meningitidis* and *Neisseria lactamica*. *International Journal of Epidemiology*, 29(1), 180–188.

[53] Coffey, T., Pullinger, G., Urwin, R., Jolley, K., Wilson, S., Maiden, M., & Leigh, J. (2006). First insights into the evolution of *Streptococcus uberis*: a multilocus sequence typing scheme that enables investigation of its population biology. *Applied and Environmental Microbiology*, 72(2), 1420–1428.

[54] Coleman, J. (1988). Distribution, prevalence, and epidemiology of bovine tuberculosis in brushtail possums, *Trichosurus-Vulpecula*, in the Hohonu range, New Zealand. *Wildlife Research*, 15(6), 651–663.

[55] Colles, F., Jones, K., Harding, R., & Maiden, M. (2003). Genetic diversity of *Campylobacter jejuni* isolates from farm animals and the farm environment. *Applied and Environmental Microbiology*, 69(12), 7409–7413.

[56] Collins, M., Jones, D., Farrow, J., Kilpper-Balz, R., & Schleifer, K. (1984). *Enterococcus avium* nom. rev., comb. nov.; *E. casseliflavus* nom. rev., comb. nov.; *E. durans* nom. rev., comb. nov.; *E. gallinarum* comb. nov.; and *E. malodoratus* sp. nov. *International Journal of Systematic Bacteriology*, 34(2), 220–223.

[57] Cooper, R. & Millener, P. (1993). The New Zealand biota: historical background and new research. *Trends in Ecology & Evolution*, 8(12), 429–433.

[58] Corander, J. & Marttinen, P. (2006). Bayesian identification of admixture events using multilocus molecular markers. *Molecular Ecology*, 15(10), 2833–2843.

[59] Cornuet, J., Santos, F., Beaumont, M., Robert, C., Marin, J., Balding, D., Guillemaud, T., & Estoup, A. (2008). Inferring population history with *DIY ABC*: a user-friendly approach to approximate Bayesian computation. *Bioinformatics*, 24(23), 2713–2719.

[60] Cowan, S., Schirmer, T., Rummel, G., Steiert, M., Ghosh, R., Pauptit, R., Jansonius, J., & Rosenbusch, J. (1992). Crystal structures explain functional properties of two e. coli porins.

[61] Cowles, M. & Carlin, B. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, (pp. 883–904).

[62] Day, N., Moore, C., Enright, M., Berendt, A., Smith, J., Murphy, M., Peacock, S., Spratt, B., & Feil, E. (2001). A link between virulence and ecological abundance in natural populations of *Staphylococcus aureus*. *Science*, 292(5514), 114.

[63] Delcher, A., Harmon, D., Kasif, S., White, O., & Salzberg, S. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Research*, 27(23), 4636–4641.

[64] Delsuc, F., Brinkmann, H., & Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, 6(5), 361.

[65] Dempster, A., Laird, N., Rubin, D., et al. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1–38.

[66] Desai, M., Logan, J., Frost, J., & Stanley, J. (2001). Genome sequence-based fluorescent amplified fragment length polymorphism of *Campylobacter jejuni*, its relationship to serotyping, and its implications for epidemiological analysis. *Journal of Clinical Microbiology*, 39(11), 3823.

[67] Diamond, J. (1984). Distributions of New Zealand birds on real and virtual islands. *New Zealand Journal of Ecology*, 7, 37–55.

[68] Didelot, X., Barker, M., Falush, D., & Priest, F. (2009a). Evolution of pathogenicity in the *Bacillus cereus* group. *Systematic and Applied Microbiology*, 32(2), 81–90.

[69] Didelot, X. & Falush, D. (2007). Inference of bacterial microevolution using multilocus sequence data. *Genetics*, 175(3), 1251.

[70] Didelot, X., Lawson, D., & Falush, D. (2009b). SimMLST: simulation of multilocus sequence typing data under a neutral model. *Bioinformatics*, 25(11), 1442.

[71] Didelot, X. & Maiden, M. (2010). Impact of recombination on bacterial evolution. *Trends in Microbiology*, 18(7), 315–322.

[72] Dingle, K., Colles, F., Falush, D., & Maiden, M. (2005). Sequence typing and comparison of population biology of *Campylobacter coli* and *Campylobacter jejuni*. *Journal of Clinical Microbiology*, 43(1), 340.

[73] Dingle, K., McCarthy, N., Cody, A., Peto, T., & Maiden, M. (2008). Extended sequence typing of *Campylobacter spp.*, United Kingdom. *Emerging Infectious Diseases*, 14(10), 1620.

[74] Dingle, K. E., Colles, F. M., Wareing, D. R. A., Ure, R., Fox, A. J., Bolton, F. E., Bootsma, H. J., Willems, R. J. L., Urwin, R., & Maiden, M. C. J. (2001). Multilocus sequence typing system for *Campylobacter jejuni*. *Journal of Clinical Microbiology*, 39(1), 14.

[75] Djordjevic, S., Unicomb, L., Adamson, P., Mickan, L., Rios, R., et al. (2007). Clonal complexes of *Campylobacter jejuni* identified by multilocus sequence typing are reliably predicted by restriction fragment length polymorphism analyses of the *flaA* gene. *Journal of Clinical Microbiology*, 45(1), 102–108.

[76] Dodge, Y., Cox, D., & Commenges, D. (2006). *The Oxford dictionary of statistical terms*. Oxford University Press, USA.

[77] Dorrell, N., Mangan, J., Laing, K., Hinds, J., Linton, D., Al-Ghusein, H., Barrell, B., Parkhill, J., Stoker, N., Karlyshev, A., et al. (2001). Whole genome comparison of *Campylobacter jejuni* human isolates using a low-cost microarray reveals extensive genetic diversity. *Genome Research*, 11(10), 1706–1715.

[78] Drouin, G., Prat, F., Ell, M., & Clarke, G. (1999). Detecting and characterizing gene conversions between multigene family members. *Molecular Biology and Evolution*, 16(10), 1369–1390.

[79] Drummond, A., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Duran, C., Field, M., Heled, J., Kearse, M., Markowitz, S., et al. (2011). Geneious v5. 4. *Biomatters Ltd, Auckland, New Zealand*.

[80] Drummond, A., Nicholls, G., Rodrigo, A., & Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, 161(3), 1307.

[81] Eberhart-Phillips, J., Walker, N., Garrett, N., Bell, D., Sinclair, D., Rainger, W., & Bates, M. (1997). Campylobacteriosis in New Zealand: results of a case-control study. *Journal of Epidemiology and Community Health*, 51(6), 686.

[82] Edwards, A. & Sforza, C. (1963). The reconstruction of evolution. *Heredity*, 18.

[83] Eggleston, A. & West, S. (1997). Recombination initiation: Easy as a, b, c, d...χ? *Current Biology*, 7(12), R745–R749.

[84] Ehling-Schulz, M., Fricker, M., & Scherer, S. (2004). *Bacillus cereus*, the causative agent of an emetic type of food-borne illness. *Molecular Nutrition & Food Research*, 48(7), 479–487.

[85] Ekdahl, K., Normann, B., & Andersson, Y. (2005). Could flies explain the elusive epidemiology of campylobacteriosis? *BMC Infectious Diseases*, 5(1), 11.

[86] Ellis-Iversen, J., Ridley, A., Morris, V., Sowa, A., Harris, J., Atterbury, R., Sparks, N., & Allen, V. (2011). Persistent environmental reservoirs on farms as

risk factors for *Campylobacter* in commercial poultry. *Epidemiology and Infection*, 1(1), 1–9.

[87] Elvers, K., Morris, V., Newell, D., & Allen, V. (2011). Molecular tracking, through processing, of *Campylobacter* strains colonizing broiler flocks. *Applied and Environmental Microbiology*, 77(16), 5722–5729.

[88] Enright, M. & Spratt, B. (1998). A multilocus sequence typing scheme for streptococcus pneumoniae: identification of clones associated with serious invasive disease. *Microbiology*, 144(11), 3049.

[89] Enright, M. & Spratt, B. (1999). Multilocus sequence typing. *Trends in Microbiology*, 7(12), 482–487.

[90] Ericson, P., Christidis, L., Cooper, A., Irestedt, M., Jackson, J., Johansson, U., & Norman, J. (2002). A Gondwanan origin of passerine birds supported by DNA sequences of the endemic New Zealand wrens. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 269(1488), 235–241.

[91] Estabrook, G., Johnson, C., & McMorris, F. (1976). A mathematical foundation for the analysis of cladistic character compatibility. *Mathematical Biosciences*, 29(1), 181–187.

[92] Evans, S. & Sayers, A. (2000). A longitudinal study of *Campylobacter* infection of broiler flocks in Great Britain. *Preventive Veterinary Medicine*, 46(3), 209–223.

[93] Excoffier, L. & Lischer, H. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, 10(3), 564–567.

[94] Excoffier, L., Smouse, P., & Quattro, J. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, 131(2), 479.

[95] Falush, D., Kraft, C., Taylor, N., Correa, P., Fox, J., Achtman, M., & Suerbaum, S. (2001). Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: estimates of clock rates, recombination size, and minimal age. *Proceedings of the National Academy of Sciences*, 98(26), 15056.

[96] Falush, D., Stephens, M., & Pritchard, J. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4), 1567.

[97] Falush, D., Stephens, M., & Pritchard, J. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes*, 7(4), 574–578.

[98] Fearnhead, P. (2007). On the choice of genetic distance in spatial-genetic studies. *Genetics*, 177(1), 427–434.

[99] Fearnhead, P., Smith, N., Barrigas, M., Fox, A., & French, N. (2005). Analysis of recombination in *Campylobacter jejuni* from MLST population data. *Journal of Molecular Evolution*, 61(3), 333–340.

[100] Feil, E., Cooper, J., Grundmann, H., Robinson, D., Enright, M., Berendt, T., Peacock, S., Smith, J., Murphy, M., & Spratt, B. (2003). How clonal is *Staphylococcus aureus*? *Journal of Bacteriology*, 185(11), 3307.

[101] Feil, E., Enright, M., & Spratt, B. (2000a). Estimating the relative contributions of mutation and recombination to clonal diversification: a comparison between *Neisseria meningitidis* and *Streptococcus pneumoniae*. *Research in Microbiology*, 151(6), 465–469.

[102] Feil, E., Holmes, E., Bessen, D., Chan, M., Day, N., Enright, M., Goldstein, R., Hood, D., Kalia, A., Moore, C., et al. (2001). Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proceedings of the National Academy of Sciences*, 98(1), 182.

[103] Feil, E., Li, B., Aanensen, D., Hanage, W., & Spratt, B. (2004). eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *Journal of Bacteriology*, 186(5), 1518.

[104] Feil, E., Maiden, M., Achtman, M., & Spratt, B. (1999). The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. *Molecular Biology and Evolution*, 16(11), 1496.

[105] Feil, E., Smith, J., Enright, M., & Spratt, B. (2000b). Estimating recombinational parameters in *Streptococcus pneumoniae* from multilocus sequence typing data. *Genetics*, 154(4), 1439.

[106] Feil, E. & Spratt, B. (2001). Recombination and the population structures of bacterial pathogens. *Annual Review of Microbiology*, 55, 561–590.

[107] Feinsinger, P., Spears, E., & Poole, R. (1981). A simple measure of niche breadth. *Ecology*, 62(1), 27–32.

[108] Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Biology*, 27(4), 401–410.

[109] Felsenstein, J. (1981). Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6), 368–376.

[110] Felsenstein, J. (1983). Statistical inference of phylogenies. *Journal of the Royal Statistical Society. Series A (General)*, (pp. 246–272).

[111] Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39(4), 783–791.

[112] Felsenstein, J. (2004). Inferring phytogenies. *Sunderland, Massachusetts: Sinauer Associates*.

[113] Fitch, W. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Biology*, 20(4), 406–416.

[114] Fitzgerald, C., Helsel, L., Nicholson, M., Olsen, S., Swerdlow, D., Flahart, R., Sexton, J., & Fields, P. (2001). Evaluation of methods for subtyping *Campylobacter jejuni* during an outbreak involving a food handler. *Journal of Clinical Microbiology*, 39(7), 2386.

[115] Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., Bult, C., Tomb, J., Dougherty, B., Merrick, J., et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269(5223), 496–512.

[116] Foley, S., White, D., McDermott, P., Walker, R., Rhodes, B., Fedorka-Cray, P., Simjee, S., & Zhao, S. (2006). Comparison of subtyping methods for differentiating *Salmonella enterica* serovar Typhimurium isolates obtained from food animal sources. *Journal of Clinical Microbiology*, 44(10), 3569–3577.

[117] Foster, G., Holmes, B., Steigerwalt, A. G., Lawson, P. A., Thorne, P., Byrer, D. E., Ross, H. M., Xerry, J., Thompson, P. M., & Collins, M. D. (2004). *Campylobacter insulaenigrae* sp. nov., isolated from marine mammals. *International Journal of Systematic and Evolutionary Microbiology*, 54(6), 2369–2373.

[118] Fouts, D., Mongodin, E., Mandrell, R., Miller, W., Rasko, D., Ravel, J., Brinkac, L., DeBoy, R., Parker, C., Daugherty, S., et al. (2005). Major structural differences and novel potential virulence mechanisms from the genomes of multiple *Campylobacter* species. *PLoS Biol*, 3(1), e15.

[119] Foxwell, A., Kyd, J., & Cripps, A. (1998). Nontypeable *Haemophilus influenzae*: pathogenesis and prevention. *Microbiology and Molecular Biology Reviews*, 62(2), 294–308.

[120] Fraser, C., Hanage, W., & Spratt, B. (2007). Recombination and the nature of bacterial speciation. *Science*, 315(5811), 476–480.

[121] French, N., M. J. (2010). Source attribution July 2009 to June 2010 of human *Campylobacter jejuni* cases from the Manawatu. *NZFSA Agreement 11777, Schedule 1A.*

[122] French, N., Barrigas, M., Brown, P., Ribiero, P., Williams, N., Leatherbarrow, H., Birtles, R., Bolton, E., Fearnhead, P., & Fox, A. (2005). Spatial epidemiology and natural population structure of *Campylobacter jejuni* colonizing a farmland ecosystem. *Environmental Microbiology*, 7(8), 1116–1126.

[123] French, N., Midwinter, A., Holland, B., Collins-Emerson, J., Pattison, R., Colles, F., & Carter, P. (2009). Molecular epidemiology of *Campylobacter jejuni* isolates from wild-bird fecal material in children's playgrounds. *Applied and Environmental Microbiology*, 75(3), 779.

[124] Friedman, C., Neimann, J., Wegener, H., & Tauxe, R. (2000). Epidemiology of *Campylobacter jejuni* infections in the United States and other industrialized nations. *Campylobacter*, 2, 121–138.

[125] Frost, J., Oza, A., Thwaites, R., & Rowe, B. (1998). Serotyping scheme for *Campylobacter jejuni* and *Campylobacter coli* based on direct agglutination of heat-stable antigens. *Journal of Clinical Microbiology*, 36(2), 335.

[126] Garrett, N., Devane, M., Hudson, J., Nicol, C., Ball, A., Klena, J., Scholes, P., Baker, M., Gilpin, B., & Savill, M. (2007). Statistical comparison of *Campylobacter jejuni* subtypes from human cases and environmental sources. *Journal of Applied Microbiology*, 103(6), 2113–2121.

[127] Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14(7), 685.

[128] Gascuel, O. & Steel, M. (2006). Neighbor-joining revealed. *Molecular Biology and Evolution*, 23(11), 1997.

[129] Gaviria, J. & Bisno, A. (2000). Group C and G *streptococci*. *Streptococcal Infections: Clinical Aspects, Microbiology and Molecular Pathogenesis. Oxford University Press, New York, NY*, (pp. 238–254).

[130] Gaykema, R., Goehler, L., & Lyte, M. (2004). Brain response to cecal infection with *Campylobacter jejuni*: analysis with fos immunohistochemistry. *Brain, Behavior, and Immunity*, 18(3), 238–245.

[131] Gaynor, E., Cawthraw, S., Manning, G., MacKichan, J., Falkow, S., & Newell, D. (2004). The genome-sequenced variant of *Campylobacter jejuni* NCTC 11168 and the original clonal clinical isolate differ markedly in colonization, gene expression, and virulence-associated phenotypes. *Journal of Bacteriology*, 186(2), 503–517.

[132] Gelman, A. & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4), 457–472.

[133] Genigeorgis, C., Hassuneh, M., & Collins, P. (1986). *Campylobacter jejuni* infection on poultry farms and its effect on poultry meat contamination during slaughtering. *Journal of Food Protection*, 49(11), 895–903.

[134] Georgsson, F., Şorkelsson, Á., Geirsdóttir, M., Reiersen, J., & Stern, N. (2006). The influence of freezing and duration of storage on *Campylobacter* and indicator bacteria in broiler carcasses. *Food Microbiology*, 23(7), 677–683.

[135] Gibson, J., Fitzgerald, C., & Owen, R. (1995). Comparison of PFGE, ribotyping and phage-typing in the epidemiological analysis of *Campylobacter jejuni* serotype HS2 infections. *Epidemiology and Infection*, 115(2), 215.

[136] Gilks, W. (1996). Markov Chain Monte Carlo. *Encyclopedia of Biostatistics*.

[137] Gillespie, I., O'Brien, S., Frost, J., Adak, G., Horby, P., Swan, A., Painter, M., Neal, K., et al. (2002). A case-case comparison of *Campylobacter coli* and *Campylobacter jejuni* infection: a tool for generating hypotheses. *Emerging Infectious Diseases*, 8(9), 937.

[138] Gilpin, B., Robson, B., Lin, S., Scholes, P., & On, S. (2012). Pulsed-field gel electrophoresis analysis of more than one clinical isolate of *Campylobacter* spp. from each of 49 patients in New Zealand. *Journal of Clinical Microbiology*, 50(2), 457–459.

[139] Glynn, P. & Iglehart, D. (1989). Importance sampling for stochastic simulations. *Management Science*, (pp. 1367–1392).

[140] Gold, R., Goldschneider, I., Lepow, M., Draper, T., & Randolph, M. (1978). Carriage of *Neisseria meningitidis* and *Neisseria lactamica* in infants and children. *Journal of Infectious Diseases*, 137(2), 112–121.

[141] Golden, M., Whittington, W., Handsfield, H., Hughes, J., Stamm, W., Hogben, M., Clark, A., Malinski, C., Helmers, J., Thomas, K., et al. (2005). Effect of expedited treatment of sex partners on recurrent or persistent gonorrhea or chlamydial infection. *New England Journal of Medicine*, 352(7), 676–685.

[142] Goossens, H., Giesendorf, B., Vandamme, P., Vlaes, L., Van den Borre, C., Koeken, A., Quint, W., Blomme, W., Hanicq, P., Koster, D., et al. (1995). Investigation of an outbreak of *Campylobacter* upsaliensis in day care centers in brussels: analysis of relationships among isolates by phenotypic and genotypic typing methods. *The Journal of Infectious Diseases*, 172(5), 1298–1305.

[143] Gotelli, N. & Colwell, R. (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, 4(4), 379–391.

[144] Graur, D. & Li, W. (2000). *Fundamentals of molecular evolution*, volume 2. Sinauer Associates Sunderland, MA.

[145] Grenfell, B., Pybus, O., Gog, J., Wood, J., Daly, J., Mumford, J., & Holmes, E. (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, 303(5656), 327.

[146] Grove-White, D., Leatherbarrow, A., Cripps, P., Diggle, P., & French, N. (2010). Temporal and farm-management-associated variation in the faecal-pat prevalence of *Campylobacter jejuni* in ruminants. *Epidemiology and Infection*, 138(04), 549–558.

[147] Guerry, P. (2007). *Campylobacter* flagella: not just for motility. *Trends in Microbiology*, 15(10), 456–461.

[148] Guerry, P., Alm, R., Power, M., Logan, S., & Trust, T. (1991). Role of two flagellin genes in *Campylobacter* motility. *Journal of Bacteriology*, 173(15), 4757.

[149] Guttman, D. & Dykhuizen, D. (1994). Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science*, 266(5189), 1380.

[150] Hall, B. (2005). Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. *Molecular Biology and Evolution*, 22(3), 792.

[151] Hall, G., Kirk, M., Becker, N., Gregory, J., Unicomb, L., Millard, G., Stafford, R., Lalor, K., et al. (2005). Estimating foodborne gastroenteritis, Australia. *Emerging Infectious Diseases*, 11(8), 1257–64.

[152] Hanninen, M., Hakkinen, M., & Rautelin, H. (1999). Stability of related human and chicken *Campylobacter jejuni* genotypes after passage through chick intestine studied by pulsed-field gel electrophoresis. *Applied and Environmental Microbiology*, 65(5), 2272.

[153] Harrington, C., Moran, L., Ridley, A., Newell, D., & Madden, R. (2003). Inter-laboratory evaluation of three flagellin pcr/rflp methods for typing *Campylobacter jejuni* and *C. coli*: the campynet experience. *Journal of Applied Microbiology*, 95(6), 1321–1333.

[154] Harrington, C., Thomson-Carter, F., & Carter, P. (1997). Evidence for recombination in the flagellin locus of *Campylobacter jejuni*: implications for the flagellin gene typing scheme. *Journal of Clinical Microbiology*, 35(9), 2386.

[155] Hartl, D. & Clark, A. (1997). *Principles of population genetics*, volume 116. Sinauer associates Sunderland, Massachusetts.

[156] Hearnden, M., Skelly, C., Eyles, R., & Weinstein, P. (2003). The regionality of campylobacteriosis seasonality in New Zealand. *International Journal of Environmental Health Research*, 13(4), 337–348.

[157] Heck Jr, K., van Belle, G., & Simberloff, D. (1975). Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. *Ecology*, (pp. 1459–1461).

[158] Hein, J., Schierup, M., & Wiuf, C. (2005). *Gene genealogies, variation and evolution: a primer in coalescent theory*. Oxford University Press, USA.

[159] Helms, M., Vastrup, P., Gerner-Smidt, P., Molbak, K., & Evans, S. (2003). Short and long term mortality associated with foodborne bacterial gastrointestinal infections: registry based study* Commentary: matched cohorts can be useful. *British Medical Journal*, 326(7385), 357.

[160] Hendrixson, D. & DiRita, V. (2003). Transcription of σ54-dependent but not σ28-dependent flagellar genes in *Campylobacter jejuni* is associated with formation of the flagellar secretory apparatus. *Molecular microbiology*, 50(2), 687–702.

[161] Hillis, D. (1995). Approaches for assessing phylogenetic accuracy. *Systematic Biology*, 44(1), 3–16.

[162] Hillis, D. & Dixon, M. (1991). Ribosomal DNA: molecular evolution and phylogenetic inference. *Quarterly Review of Biology*, (pp. 411–453).

[163] Hofshagen, M. & Kruse, H. (2005). Reduction in flock prevalence of *Campylobacter spp.* in broilers in Norway after implementation of an action plan. *Journal of Food Protection 174;*, 68(10), 2220–2223.

[164] Hogg, A., Higham, T., Lowe, D., Palmer, J., Reimer, P., & Newnham, R. (2002). A wiggle-match date for Polynesian settlement of New Zealand. *Antiquity*, 77 (295), 116–125.

[165] Holland, B., Delsuc, F., & Moulton, V. (2005). Visualizing conflicting evolutionary hypotheses in large collections of trees: using consensus networks to study the origins of placentals and hexapods. *Systematic Biology*, 54(1), 66–76.

[166] Holland, B., Huber, K., Moulton, V., & Lockhart, P. (2004). Using consensus networks to visualize contradictory evidence for species phylogeny. *Molecular Biology and Evolution*, 21(7), 1459.

[167] Holmes, K., Johnson, D., & Trostle, H. (1970). An estimate of the risk of men acquiring gonorrhea by sexual contact with infected females. *American Journal of Epidemiology*, 91(2), 170–174.

[168] Holsinger, K. & Weir, B. (2009). Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature Reviews Genetics*, 10(9), 639–650.

[169] Hubisz, M., Falush, D., Stephens, M., & Pritchard, J. (2009). Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, 9(5), 1322–1332.

[170] Hudson, R. (1983a). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23(2), 183–201.

[171] Hudson, R. (1983b). Testing the constant-rate neutral allele model with protein sequence data. *Evolution*, 37(1), 203–217.

[172] Hudson, R. (1990). Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, 7(1), 44.

[173] Hudson, R. & Kaplan, N. (1988). The coalescent process in models with selection and recombination. *Genetics*, 120(3), 831.

[174] Huelsenbeck, J. (1995). Performance of phylogenetic methods in simulation. *Systematic Biology*, 44(1), 17–48.

[175] Huelsenbeck, J., Ronquist, F., Nielsen, R., & Bollback, J. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294(5550), 2310.

[176] Humphrey, T., O'Brien, S., & Madsen, M. (2007). *Campylobacters* as zoonotic pathogens: A food production perspective. *International Journal of Food Microbiology*, 117(3), 237–257.

[177] Huson, D. (1998). SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*, 14(1), 68.

[178] Huson, D. & Bryant, D. (2004). Estimating phylogenetic trees and networks using SplitsTree 4. *manuscript in preparation, software available from www-ab. informatik. uni-tuebingen. de/software*.

[179] Huson, D. & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2), 254.

[180] Huson, D., Rupp, R., & Scornavacca, C. (2011). *Phylogenetic networks: concepts, algorithms and applications*. Cambridge University Press.

[181] Huyer, M., Parr Jr, T., Hancock, R., & Page, W. (1986). Outer membrane porin protein of *Campylobacter jejuni*. *FEMS Microbiology Letters*, 37(3), 247–250.

[182] Istre, G., Conner, J., Broome, C., Hightower, A., & Hopkins, R. (1985). Risk factors for primary invasive *Haemophilus influenzae* disease: increased risk from day care attendance and school-aged household members. *The Journal of pediatrics*, 106(2), 190–195.

[183] Iwase, T., Uehara, Y., Shinji, H., Tajima, A., Seo, H., Takada, K., Agata, T., & Mizunoe, Y. (2010). *Staphylococcus epidermidis* Esp inhibits *Staphylococcus aureus* biofilm formation and nasal colonization. *Nature*, 465(7296), 346–349.

[184] Janda, W., Gaydos, C., Murray, P., Baron, E., Jorgensen, J., Landry, M., Pfaller, M., et al. (2006). *Neisseria*. *Manual of Clinical Microbiology*, 1(Ed. 9), 601–620.

[185] Jensen, E. & Miller, C. (2001). *Staphylococcus* infections in broiler breeders. *AviaTech*, 1, 1–6.

[186] Jolley, K., Bliss, C., Bennett, J., Bratcher, H., Brehony, C., Colles, F., Wimalarathna, H., Harrison, O., Sheppard, S., Cody, A., et al. (2012). Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology*, 158(Pt 4), 1005–1015.

[187] Jolley, K., Feil, E., Chan, M., & Maiden, M. (2001). Sequence type analysis and recombinational tests (START). *Bioinformatics*, 17(12), 1230–1231.

[188] Jolley, K. & Maiden, M. (2010). BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*, 11(1), 595.

[189] Jolley, K., Wilson, D., Kriz, P., McVean, G., & Maiden, M. (2005). The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in *Neisseria meningitidis*. *Molecular Biology and Evolution*, 22(3), 562.

[190] Jorgensen, F., Ellis-Iversen, J., Rushton, S., Bull, S., Harris, S., Bryan, S., Gonzalez, A., & Humphrey, T. (2011). Influence of season and geography on *Campylobacter jejuni* and *C. coli* subtypes in housed broiler flocks reared in Great Britain. *Applied and Environmental Microbiology*, 77(11), 3741.

[191] Josefsen, M., Carroll, C., Rudi, K., Engvall, E., Hoorfar, J., et al. (2011). *Campylobacter* in poultry, pork, and beef. *Rapid Detection, characterization, and Enumeration of Foodborne Pathogens*, (pp. 209–227).

[192] Kaplan, N., Darden, T., & Hudson, R. (1988). The coalescent process in models with selection. *Genetics*, 120(3), 819.

[193] Karenlampi, R., Rautelin, H., Schonberg-Norio, D., Paulin, L., & Hanninen, M. (2006). Longitudinal study of finnish human *Campylobacter jejuni* and *C. coli* isolates using multilocus sequence typing, including comparison with epidemiological data, and poultry and cattle isolates. *Applied and Environmental Microbiology*, (pp. AEM–01488).

[194] Karlyshev, A., Champion, O., Churcher, C., Brisson, J., Jarrell, H., Gilbert, M., Brochu, D., St Michael, F., Li, J., Wakarchuk, W., et al. (2005). Analysis of *Campylobacter jejuni* capsular loci reveals multiple mechanisms for the generation of structural diversity and the ability to form complex heptoses. *Molecular Microbiology*, 55(1), 90–103.

[195] Kass, R. & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, (pp. 1343–1370).

[196] Kelly, D. (2001). The physiology and metabolism of *Campylobacter jejuni* and *Helicobacter pylori*. *Journal of Applied Microbiology*, 90(S6), 16S–24S.

[197] Ketley, J. & Konkel, M. (2005). *Campylobacter: molecular and cellular biology*. Taylor & Francis.

[198] Kimura, M. (1985). *The neutral theory of molecular evolution.* Cambridge University Press.

[199] Kinana, A., Cardinale, E., Bahsoun, I., Tall, F., Sire, J., Breurec, S., Garin, B., Saad-Bouh Boye, C., & Perrier-Gros-Claude, J. (2007). *Campylobacter coli* isolates derived from chickens in Senegal: diversity, genetic exchange with *Campylobacter jejuni* and quinolone resistance. *Research in Microbiology*, 158(2), 138–142.

[200] Kinana, A., Cardinale, E., Tall, F., Bahsoun, I., Sire, J., Garin, B., Breurec, S., Boye, C., & Perrier-Gros-Claude, J. (2006). Genetic diversity and quinolone resistance in *Campylobacter jejuni* isolates from poultry in Senegal. *Applied and Environmental Microbiology*, 72(5), 3309.

[201] Kingman, J. (1982a). The coalescent. *Stochastic Processes and Their Applications*, 13(3), 235–248.

[202] Kingman, J. (1982b). On the genealogy of large populations. *Journal of Applied Probability*, 19, 27–43.

[203] Kingman, J. (2000). Origins of the coalescent. 1974-1982. *Genetics*, 156(4), 1461.

[204] Kinsella, N., Guerry, P., Cooney, J., & Trust, T. (1997). The *flgE* gene of *Campylobacter coli* is under the control of the alternative sigma factor σ^{54}. *Journal of Bacteriology*, 179(15), 4647.

[205] Kist, M. (1985). The historical background of *Campylobacter* infection: new aspects. In *Proceedings of the 3rd International Workshop on Campylobacter infection: 1985; Ottawa* (pp. 23–27).

[206] Kluytmans, J., Van Belkum, A., & Verbrugh, H. (1997). Nasal carriage of *Staphylococcus aureus*: epidemiology, underlying mechanisms, and associated risks. *Clinical Microbiology Reviews*, 10(3), 505–520.

[207] Kohn, A. & Riggs, A. (1982). Sample size dependence in measures of proportional similarity. *Marine Ecology Progress Series. Oldendorf*, 9(2), 147–151.

[208] Konkel, M., Gray, S., Kim, B., Garvis, S., & Yoon, J. (1999). Identification of the enteropathogens *Campylobacter jejuni* and *Campylobacter coli* based on the *cadF* virulence gene and its product. *Journal of Clinical Microbiology*, 37(3), 510.

[209] Korczak, B., Zurfluh, M., Emler, S., Kuhn-Oertli, J., & Kuhnert, P. (2009). Multiplex strategy for MLST, *fla*-typing and genetic determination of antimicrobial resistance of Swiss *Campylobacter jejuni* and *Campylobacter coli* isolates. *Journal of Clinical Microbiology*, (pp. JCM–00237).

[210] Kossaibati, M. & Esslemont, R. (1997). The costs of production diseases in dairy herds in England. *The Veterinary Journal*, 154(1), 41–51.

[211] Kotiranta, A., Lounatmaa, K., & Haapasalo, M. (2000). Epidemiology and pathogenesis of *Bacillus cereus* infections. *Microbes and Infection*, 2(2), 189–198.

[212] Kreitman, M. (2000). Methods to detect selection in populations with applications to the human. *Annual Review of Genomics and Human Genetics*, 1(1), 539–559.

[213] Kuchenmüller, T., Hird, S., Stein, C., Kramarz, P., Nanda, A., Havelaar, A., et al. (2009). Estimating the global burden of foodborne diseases-a collaborative effort. *Eurosurveillance*, 14(18).

[214] Kuhner, M., Yamato, J., & Felsenstein, J. (1998). Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, 149(1), 429.

[215] Kühnert, D., Wu, C., & Drummond, A. (2011). Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. *Infection, Genetics and Evolution*.

[216] Kuhnert, P. & Christensen, H. (2008). *Pasteurellaceae: biology, genomics and molecular aspects*. ister Academic Press.

[217] Kulick, S., Moccia, C., Didelot, X., Falush, D., Kraft, C., & Suerbaum, S. (2008). Mosaic dna imports with interspersions of recipient sequence after natural transformation of *Helicobacter pylori*. *PLoS One*, 3(11), e3797.

[218] Kumar, S. & Filipski, A. (2007). Multiple sequence alignment: in pursuit of homologous DNA positions. *Genome Research*, 17(2), 127–135.

[219] Kuusi, M., Lahti, E., Virolainen, A., Hatakka, M., Vuento, R., Rantala, L., Vuopio-Varkila, J., Seuna, E., Karppelin, M., Hakkinen, M., et al. (2006). An outbreak of *Streptococcus equi* subspecies *zooepidemicus* associated with consumption of fresh goat cheese. *BMC Infectious Diseases*, 6(1), 36.

[220] Kwan, P., Barrigas, M., Bolton, F., French, N., Gowland, P., Kemp, R., Leatherbarrow, H., Upton, M., & Fox, A. (2008a). Molecular epidemiology of

Campylobacter jejuni populations in dairy cattle, wildlife, and the environment in a farmland area. *Applied and Environmental Microbiology*, 74(16), 5130–5138.

[221] Kwan, P., Birtles, A., Bolton, F., French, N., Robinson, S., Newbold, L., Upton, M., & Fox, A. (2008b). Longitudinal study of the molecular epidemiology of *Campylobacter jejuni* in cattle on dairy farms. *Applied and Environmental Microbiology*, 74(12), 3626–3633.

[222] Lane, L. & Baker, M. (1993). Are we experiencing an epidemic of campylobacter infection. *Comm Dis NZ*, 93, 7–63.

[223] Leatherbarrow, A., Hart, C., Kemp, R., Williams, N., Ridley, A., Sharma, M., Diggle, P., Wright, E., Sutherst, J., & French, N. (2004). Genotypic and antibiotic susceptibility characteristics of a *Campylobacter coli* population isolated from dairy farmland in the United Kingdom. *Applied and Environmental Microbiology*, 70(2), 822.

[224] Levin, B. (1981). Periodic selection, infectious gene exchange and the genetic structure of *E. coli* populations. *Genetics*, 99(1), 1–23.

[225] Levin, B. & Bergstrom, C. (2000). Bacteria are different: observations, interpretations, speculations, and opinions about the mechanisms of adaptive evolution in prokaryotes. *Proceedings of the National Academy of Sciences*, 97(13), 6981–6985.

[226] Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14), 1754–1760.

[227] Li, W. & Olmstead, R. (1997). *Molecular evolution*. Sinauer Associates Sunderland, MA.

[228] Librado, P. & Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, 25(11), 1451.

[229] Lindmark, H., Nilsson, M., & Guss, B. (2001). Comparison of the fibronectin-binding protein FNE from *Streptococcus equi* subspecies *equi* with FNZ from *S. equi* subspecies *zooepidemicus* reveals a major and conserved difference. *Infection and Immunity*, 69(5), 3159–3163.

[230] Lindstedt, B., Heir, E., Vardund, T., & Kapperud, G. (2000). Fluorescent amplified-fragment length polymorphism genotyping of *Salmonella enterica* subsp. *enterica serovars* and comparison with pulsed-field gel electrophoresis typing. *Journal of Clinical Microbiology*, 38(4), 1623.

[231] Lior, H., Woodward, D., Edgar, J., Laroche, L., & Gill, P. (1982). Serotyping of *Campylobacter jejuni* by slide agglutination based on heat-labile antigenic factors. *Journal of Clinical Microbiology*, 15(5), 761.

[232] Litrup, E., Torpdahl, M., & Nielsen, E. (2007). Multilocus sequence typing performed on *Campylobacter coli* isolates from humans, broilers, pigs and cattle originating in Denmark. *Journal of Applied Microbiology*, 103(1), 210–218.

[233] Long, J. (1986). The allelic correlation structure of Gainj-and Kalam-speaking people. i. the estimation and interpretation of Wright's F-statistics. *Genetics*, 112(3), 629–647.

[234] Lorenz, M. & Wackernagel, W. (1994). Bacterial gene transfer by natural genetic transformation in the environment. *Microbiological Reviews*, 58(3), 563.

[235] Lowy, F. (1998). *Staphylococcus aureus* infections. *New England Journal of Medicine*, 339(8), 520–532.

[236] Mahan, M., Slauch, J., Mekalanos, J., et al. (1993). Selection of bacterial virulence genes that are specifically induced in host tissues. *Science*, 259, 686–686.

[237] Maiden, M., Bygraves, J., Feil, E., Morelli, G., Russell, J., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., & Caugant, D. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America*, 95(6), 3140.

[238] Manly, B. (1985). *The statistics of natural selection on animal populations*. Chapman and Hall.

[239] Manning, G., Dowson, C., Bagnall, M., Ahmed, I., West, M., & Newell, D. (2003). Multilocus sequence typing for comparison of veterinary and human isolates of *Campylobacter jejuni*. *Applied and Environmental Microbiology*, 69(11), 6370.

[240] Martin, D., Lemey, P., Lott, M., Moulton, V., Posada, D., & Lefeuvre, P. (2010). RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics*, 26(19), 2462.

[241] Mascini, E., Troelstra, A., Beitsma, M., Blok, H., Jalink, K., Hopmans, T., Fluit, A., Hene, R., Willems, R., Verhoef, J., et al. (2006). Genotyping and preemptive isolation to control an outbreak of vancomycin-resistant *Enterococcus faecium*. *Clinical Infectious Diseases*, 42(6), 739.

[242] Matthews, K., Almeida, R., & Oliver, S. (1994). Bovine mammary epithelial cell invasion by *Streptococcus uberis*. *Infection and Immunity*, 62(12), 5641–5646.

[243] McCarthy, N., Colles, F., Dingle, K., Bagnall, M., Manning, G., Maiden, M., & Falush, D. (2007). Host-associated genetic import in *Campylobacter jejuni*. *Emerging infectious diseases*, 13(2), 267.

[244] McGlone, M. (1983). Polynesian deforestation of New Zealand: a preliminary synthesis. *Archaeology in Oceania*, 18(1), 11–25.

[245] McGlone, M. (1989). The Polynesian settlement of New Zealand in relation to environmental and biotic changes. *New Zealand Journal of Ecology*, 12((Supplement)), 115–129.

[246] McTavish, S., Pope, C., Nicol, C., Sexton, K., French, N., & Carter, P. (2008). Wide geographical distribution of internationally rare *Campylobacter* clones within New Zealand. *Epidemiology and Infection*, 136(09), 1244–1252.

[247] McWethy, D., Whitlock, C., Wilmshurst, J., McGlone, M., Fromont, M., Li, X., Dieffenbacher-Krall, A., Hobbs, W., Fritz, S., & Cook, E. (2010). Rapid landscape transformation in South Island, New Zealand, following initial Polynesian settlement. *Proceedings of the National Academy of Sciences*, 107(50), 21343–21348.

[248] Meinersmann, R., Helsel, L., Fields, P., & Hiett, K. (1997). Discrimination of *Campylobacter jejuni* isolates by *fla* gene sequencing. *Journal of Clinical Microbiology*, 35(11), 2810.

[249] Meinersmann, R. & Hiett, K. (2000). Concerted evolution of duplicate *fla* genes in *Campylobacter*. *Microbiology*, 146(9), 2283.

[250] Meinersmann, R., Patton, C., Evins, G., Wachsmuth, I., & Fields, P. (2002). Genetic diversity and relationships of *Campylobacter* species and subspecies. *International Journal of Systematic and Evolutionary Microbiology*, 52(5), 1789.

[251] Meinersmann, R., Phillips, R., Hiett, K., & Fedorka-Cray, P. (2005). Differentiation of *Campylobacter* populations as demonstrated by flagellin short variable region sequences. *Applied and Environmental Microbiology*, 71(10), 6368.

[252] Milkman, R. (1973). Electrophoretic variation in *Escherichia coli* from natural sources. *Science (New York, NY)*, 182(116), 1024.

[253] Milkman, R. & Bridges, M. (1990). Molecular evolution of the *Escherichia coli* chromosome. III. Clonal Frames. *Genetics*, 126(3), 505.

[254] Millener, P. R. (1981). *The Quaternary Avifauna of New Zealand*. PhD thesis, Geology Department, University of Auckland, New Zealand.

[255] Miller, W., Englen, M., Kathariou, S., Wesley, I., Wang, G., Pittenger-Alley, L., Siletz, R., Muraoka, W., Fedorka-Cray, P., & Mandrell, R. (2006). Identification of host-associated alleles by multilocus sequence typing of *Campylobacter coli* strains from food animals. *Microbiology*, 152(1), 245.

[256] Miller, W., On, S., Wang, G., Fontanoz, S., Lastovica, A., & Mandrell, R. (2005). Extended multilocus sequence typing system for *Campylobacter coli*, *C. lari*, *C. upsaliensis*, and *C. helveticus*. *Journal of Clinical Microbiology*, 43(5), 2315.

[257] Moriarty, E., Mackenzie, M., Karki, N., & Sinton, L. (2011). Survival of *Escherichia coli*, enterococci, and *Campylobacter* spp. in sheep feces on pastures. *Applied and Environmental Microbiology*, 77(5), 1797–1803.

[258] Mosteller, F. (2006). On some useful "inefficient" statistics. *Selected Papers of Frederick Mosteller*, (pp. 69–100).

[259] Müllner, P., Collins-Emerson, J., Midwinter, A., Carter, P., Spencer, S., van der Logt, P., Hathaway, S., & French, N. (2010). Molecular epidemiology of *Campylobacter jejuni* in a geographically isolated country with a uniquely structured poultry industry. *Applied and Environmental Microbiology*, 76(7), 2145–2154.

[260] Mullner, P., Jones, G., Noble, A., Spencer, S., Hathaway, S., & French, N. (2009a). Source attribution of food-borne zoonoses in New Zealand: a modified hald model. *Risk Analysis*, 29(7), 970–984.

[261] Mullner, P., Shadbolt, T., Collins-Emerson, J., Midwinter, A., Spencer, S., Marshall, J., Carter, P., Campbell, D., Wilson, D., Hathaway, S., et al. (2010). Molecular and spatial epidemiology of human campylobacteriosis: source association and genotype-related risk factors. *Epidemiology and Infection*, 138(10), 1372–1383.

[262] Mullner, P., Spencer, S., Wilson, D., Jones, G., Noble, A., Midwinter, A., Collins-Emerson, J., Carter, P., Hathaway, S., & French, N. (2009b). Assigning the source of human campylobacteriosis in New Zealand: a comparative genetic and epidemiological approach. *Infection, Genetics and Evolution*, 9(6), 1311–1319.

[263] Munson Jr, R., Kabeer, M., Lenoir, A., & Granoff, D. (1989). Epidemiology and prospects for prevention of disease due to *Haemophilus influenzae* in developing countries. *Review of Infectious Diseases*, 11(Supplement 3), S588–S597.

[264] Nachamkin, I. (2002). Chronic effects of *Campylobacter* infection. *Microbes and Infection*, 4(4), 399–403.

[265] Nachamkin, I., Blaser, M., & Tompkins, L. (1992). *Campylobacter jejuni: current status and future trends*. American Society for Microbiology.

[266] Nachamkin, I., Bohachick, K., & Patton, C. (1993). Flagellin gene typing of *Campylobacter jejuni* by restriction fragment length polymorphism analysis. *Journal of Clinical Microbiology*, 31(6), 1531.

[267] Nachamkin, I., Szymanski, C. M., & Blaser, M. J., Eds. (2008). *Campylobacter*. ASM Press; 3rd edition.

[268] Navarro, F., Llovet, T., Echeita, M., Coll, P., Aladuena, A., Usera, M., & Prats, G. (1996). Molecular typing of *Salmonella enterica* serovar typhi. *Journal of Clinical Microbiology*, 34(11), 2831–2834.

[269] Nelson, W. & Harris, B. (2006). Flies, fingers, fomites, and food. campylobacteriosis in New Zealand food-associated rather than food-borne. *Journal of the New Zealand Medical Association*, 119, 1240.

[270] Neuhauser, C. & Krone, S. (1997). The genealogy of samples in models with selection. *Genetics*, 145(2), 519.

[271] Newell, D. & Fearnley, C. (2003). Sources of *Campylobacter* colonization in broiler chickens. *Applied and Environmental Microbiology*, 69(8), 4343.

[272] Nichols, G. et al. (2005). Fly transmission of *Campylobacter*. *Emerging Infectious Diseases*, 11(3), 361–4.

[273] Nielsen, R. & Beaumont, M. (2009). Statistical inferences in phylogeography. *Molecular Ecology*, 18(6), 1034–1047.

[274] Nielsen, R. & Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148(3), 929–936.

[275] Nordborg, M. & Tavaré, S. (2002). Linkage disequilibrium: what history has to tell us. *Trends in Genetics*, 18(2), 83–90.

[276] Nordmann, P., Cuzon, G., & Naas, T. (2009). The real threat of *Klebsiella pneumoniae* carbapenemase-producing bacteria. *The Lancet Infectious Diseases*, 9(4), 228–236.

[277] Nuijten, P., Bartels, C., Bleumink-Pluym, N., Gaastra, W., & van der Zeijst, B. (1990a). Size and physical map of the *Campylobacter jejuni* chromosome. *Nucleic Acids Research*, 18(21), 6211.

[278] Nuijten, P., Van Asten, F., Gaastra, W., & Van der Zeijst, B. (1990b). Structural and functional analysis of two *Campylobacter jejuni* flagellin genes. *Journal of Biological Chemistry*, 265(29), 17798.

[279] Nylen, G., Dunstan, F., Palmer, S., Andersson, Y., Bager, F., Cowden, J., Feierl, G., Galloway, Y., Kapperud, G., Megraud, F., et al. (2002). The seasonal distribution of *Campylobacter* infection in nine European countries and New Zealand. *Epidemiology and Infection*, 128(03), 383–390.

[280] Oberhelman, R. & Taylor, D. (2000). *Campylobacter* infections in developing countries. *Campylobacter, 2nd ed. ASM Press, Washington, DC*, (pp. 139–153).

[281] Ogden, I., Dallas, J., MacRae, M., Rotariu, O., Reay, K., Leitch, M., Thomson, A., Sheppard, S., Maiden, M., Forbes, K., et al. (2009). *Campylobacter* excreted into the environment by animal sources: prevalence, concentration shed, and host association. *Foodborne Pathogens and Disease*, 6(10), 1161–1170.

[282] Olivier, M. (2003). A haplotype map of the human genome. *Physiological Genomics*, 13(1), 3.

[283] On, S., Nielsen, E., Engberg, J., & Madsen, M. (1998). Validity of *Sma*I-defined genotypes of *Campylobacter jejuni* examined by *Sal*I, *Kpn*I, and *Bam*HI polymorphisms: evidence of identical clones infecting humans, poultry, and cattle. *Epidemiology and Infection*, 120(03), 231–237.

[284] Owen, R., Hernandez, J., & Bolton, F. (1990). DNA restriction digest and ribosomal RNA gene patterns of *Campylobacter jejuni*: a comparison with bio-, sero-, and bacteriophage-types of United Kingdom outbreak strains. *Epidemiology and Infection*, 105(2), 265.

[285] Owen, R. & Leaper, S. (1981). Base composition, size and nucleotide sequence similarities of genome deoxyribonucleic acids from speies of the genus *Campylobacter*. *FEMS Microbiology Letters*, 12(4), 395–400.

[286] Parkhill, J., Wren, B., Mungall, K., Ketley, J., Churcher, C., Basham, D., Chillingworth, T., Davies, R., Feltwell, T., Holroyd, S., et al. (2000). The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature*, 403(6770), 665–668.

[287] Parsons, B., Cody, A., Porter, C., Stavisky, J., Smith, J., Williams, N., Leatherbarrow, A., Hart, C., Gaskell, R., Dingle, K., et al. (2009). Typing of *Campylobacter jejuni* isolates from dogs by use of multilocus sequence typing and pulsed-field gel electrophoresis. *Journal of Clinical Microbiology*, 47(11), 3466–3471.

[288] Patton, C., Wachsmuth, I., Evins, G., Kiehlbauch, J., Plikaytis, B., Troup, N., Tompkins, L., & Lior, H. (1991). Evaluation of 10 methods to distinguish epidemic-associated *Campylobacter* strains. *Journal of Clinical Microbiology*, 29(4), 680.

[289] Peltola, H. (2000). Worldwide *Haemophilus influenzae* type b disease at the beginning of the 21st century: global analysis of the disease burden 25 years after the use of the polysaccharide vaccine and a decade after the advent of conjugates. *Clinical Microbiology Reviews*, 13(2), 302–317.

[290] Penner, J. & Hennessy, J. (1980). Passive hemagglutination technique for serotyping *Campylobacter fetus* subsp. *jejuni* on the basis of soluble heat-stable antigens. *Journal of Clinical Microbiology*, 12(6), 732.

[291] Penny, D. (1982). Towards a basis for classification: the incompleteness of distance measures, incompatibility analysis and phenetic classification. *Journal of Theoretical Biology*, 96(2), 129–142.

[292] Penny, D. & Hendy, M. (1985a). Testing methods of evolutionary tree construction. *Cladistics*, 1(3), 266–278.

[293] Penny, D. & Hendy, M. (1985b). The use of tree comparison metrics. *Systematic Zoology*, (pp. 75–82).

[294] Penny, D. & Hendy, M. (1986). Estimating the reliability of evolutionary trees. *Molecular Biology and Evolution*, 3(5), 403–417.

[295] Petersen, L. & Newell, D. (2001). The ability of *Fla*-typing schemes to discriminate between strains of *Campylobacter jejuni*. *Journal of applied microbiology*, 91(2), 217–224.

[296] Pochon, X., Montoya-Burgos, J., Stadelmann, B., & Pawlowski, J. (2006). Molecular phylogeny, evolutionary rates, and divergence timing of the symbiotic dinoflagellate genus *Symbiodinium*. *Molecular Phylogenetics and Evolution*, 38(1), 20–30.

[297] Pohlner, J., Halter, R., Beyreuther, K., & Meyer, T. (1987). Gene structure and extracellular secretion of *Neisseria gonorrhoeae* IgA protease.

[298] Pole, M. (1994). The New Zealand flora-entirely long-distance dispersal? *Journal of Biogeography*, (pp. 625–635).

[299] Posada, D. (2006). Modeltest server: a web-based tool for the statistical selection of models of nucleotide substitution online. *Nucleic Acids Research*, 34(suppl 2), W700–W703.

[300] Posada, D. & Crandall, K. (1998). Modeltest: testing the model of DNA substitution. *Bioinformatics*, 14(9), 817.

[301] Posada, D. & Crandall, K. (2002). The effect of recombination on the accuracy of phylogeny estimation. *Journal of Molecular Evolution*, 54(3), 396–402.

[302] Potter, R., Kaneene, J., & Hall, W. (2003). Risk factors for sporadic *Campylobacter jejuni* infections in rural Michigan: a prospective case-control study. *American Journal of Public Health*, 93(12), 2118.

[303] Pritchard, J., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2), 945.

[304] Reid, A., Taubenberger, J., & Fanning, T. (2001). The 1918 spanish influenza: integrating history and biology. *Microbes and Infection*, 3(1), 81–87.

[305] Reynolds, J., Weir, B., & Cockerham, C. (1983). Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics*, 105(3), 767.

[306] Richman, A., Herrera, L., Nash, D., & Schierup, M. (2003). Relative roles of mutation and recombination in generating allelic polymorphism at an MHC class II locus in *Peromyscus maniculatus*. *Genetics Research*, 82(02), 89–99.

[307] Robinson, D. (1981). Infective dose of *Campylobacter jejuni* in milk. *British Medical Journal (Clinical research ed.)*, 282(6276), 1584–1584.

[308] Robinson, D. & Foulds, L. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1), 131–147.

[309] Rosef, O., Kapperud, G., Lauwers, S., & Gondrosen, B. (1985). Serotyping of *Campylobacter jejuni*, *Campylobacter coli*, and *Campylobacter laridis* from domestic and wild animals. *Applied and Environmental Microbiology*, 49(6), 1507.

[310] Rosenberg, N. & Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, 3(5), 380–390.

[311] Rosenquist, H., Sommer, H., Nielsen, N., & Christensen, B. (2006). The effect of slaughter operations on the contamination of chicken carcasses with thermotolerant *Campylobacter*. *International Journal of Food Microbiology*, 108(2), 226–232.

[312] Rosenstein, N., Perkins, B., Stephens, D., Popovic, T., & Hughes, J. (2001). Meningococcal disease. *New England Journal of Medicine*, 344(18), 1378–1388.

[313] Ryan, P. (2010). Sherris medical microbiology. *Recherche*, 67, 02.

[314] Saigh, J., Sanders, C., & Sanders Jr, W. (1978). Inhibition of *Neisseria gonorrhoeae* by aerobic and facultatively anaerobic components of the endocervical flora: evidence for a protective effect against infection. *Infection and Immunity*, 19(2), 704–710.

[315] Sails, A., Swaminathan, B., & Fields, P. (2003). Clonal complexes of *Campylobacter jejuni* identified by multilocus sequence typing correlate with strain associations identified by multilocus enzyme electrophoresis. *Journal of Clinical Microbiology*, 41(9), 4058.

[316] Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4), 406.

[317] Salzberg, S., Delcher, A., Kasif, S., & White, O. (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Research*, 26(2), 544–548.

[318] Sandberg, M., Bergsjo, B., Hofshagen, M., Skjerve, E., & Kruse, H. (2002). Risk factors for *Campylobacter* infection in Norwegian cats and dogs. *Preventive Veterinary Medicine*, 55(4), 241–253.

[319] Sari Kovats, R., Edwards, S., Charron, D., Cowden, J., D'Souza, R., Ebi, K., Gauci, C., Gerner-Smidt, P., Hajat, S., Hales, S., et al. (2005). Climate variability and *Campylobacter* infection: an international study. *International Journal of Biometeorology*, 49(4), 207–214.

[320] Sarkar, S. & Guttman, D. (2004). Evolution of the core genome of *Pseudomonas syringae*, a highly clonal, endemic plant pathogen. *Applied and Environmental Microbiology*, 70(4), 1999.

[321] Savill, M., Hudson, A., Devane, M., Garrett, N., Gilpin, B., & Ball, A. (2003). Elucidation of potential transmission routes of *Campylobacter* in New Zealand. *Water Science and Technology: a Journal of the International Association on Water Pollution Research*, 47(3), 33.

[322] Savill, M., Hudson, J., Ball, A., Klena, J., Scholes, P., Whyte, R., McCormick, R., & Jankovic, D. (2001). Enumeration of *Campylobacter* in New Zealand recreational and drinking waters. *Journal of Applied Microbiology*, 91(1), 38–46.

[323] Sawyer, S. (1989). Statistical tests for detecting gene conversion. *Molecular Biology and Evolution*, 6(5), 526–538.

[324] Schifman, R. & Ryan, K. (1983). *Neisseria lactamica* septicemia in an immunocompromised patient. *Journal of Clinical Microbiology*, 17(5), 934–935.

[325] Schouls, L., Reulen, S., Duim, B., Wagenaar, J., Willems, R., Dingle, K., Colles, F., & Van Embden, J. (2003). Comparative genotyping of *Campylobacter jejuni* by amplified fragment length polymorphism, multilocus sequence typing, and short repeat sequencing: strain diversity, host range, and recombination. *Journal of Clinical Microbiology*, 41(1), 15.

[326] Scott, W., Scott, H., Lake, R., & Baker, M. (2000). Economic cost to New Zealand of foodborne infectious disease. *The New Zealand Medical Journal*, 113(1113), 281.

[327] Sears, A. (2009). Campylobacteriosis in New Zealand 1997-2008 describing the recent decline in notifications. In *NZFSA Attribution Workshop, 28th October 2009; Wellington.*

[328] Sears, A., Baker, M., Wilson, N., Marshall, J., Muellner, P., Campbell, D., Lake, R., & French, N. (2011). Marked campylobacteriosis decline after interventions aimed at poultry, New Zealand. *Emerging Infectious Diseases*, 17(6), 1007.

[329] Selander, R., Caugant, D., Ochman, H., Musser, J., Gilmour, M., & Whittam, T. (1986). Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. *Applied and Environmental Microbiology*, 51(5), 873.

[330] Senior, K. (2009). Estimating the global burden of foodborne disease. *The Lancet Infectious Diseases*, 9(2), 80–81.

[331] Shen, T., Chao, A., & Lin, C. (2003). Predicting the number of new species in further taxonomic sampling. *Ecology*, 84(3), 798–804.

[332] Sheppard, S., Colles, F., Richardson, J., Cody, A., Elson, R., Lawson, A., Brick, G., Meldrum, R., Little, C., Owen, R., et al. (2010a). Host association of *Campylobacter* genotypes transcends geographic variation. *Applied and Environmental Microbiology*, 76(15), 5269.

[333] Sheppard, S., Dallas, J., MacRae, M., McCarthy, N., Sproston, E., Gormley, F., Strachan, N., Ogden, I., Maiden, M., & Forbes, K. (2009). *Campylobacter* genotypes from food animals, environmental sources and clinical disease in Scotland 2005/6. *International Journal of Food Microbiology*, 134(1-2), 96–103.

[334] Sheppard, S., Dallas, J., Wilson, D., Strachan, N., McCarthy, N., Jolley, K., Colles, F., Rotariu, O., Ogden, I., Forbes, K., et al. (2010b). Evolution of an agriculture-associated disease causing *Campylobacter coli* clade: evidence from national surveillance data in Scotland. *PloS One*, 5(12), e15708.

[335] Sheppard, S., McCarthy, N., Falush, D., & Maiden, M. (2008). Convergence of *Campylobacter* species: implications for bacterial evolution. *Science*, 320(5873), 237.

[336] Sheppard, S., McCarthy, N., Jolley, K., & Maiden, M. (2011). Introgression in the genus *Campylobacter*: generation and spread of mosaic alleles. *Microbiology*, 157(4), 1066.

[337] Simmons, D. (1969). Economic change in New Zealand prehistory. *The Journal of the Polynesian Society*, (pp. 3–34).

[338] Simon, R., Priefer, U., & Pühler, A. (1983). A broad host range mobilization system for in vivo genetic engineering: transposon mutagenesis in gram negative bacteria. *Nature Biotechnology*, 1(9), 784–791.

[339] Skirrow, M. & Blaser, M. (2000). Clinical aspects of *Campylobacter* infection. *Campylobacter*, 2, 69–88.

[340] Skirrow, M., Jones, D., Sutcliffe, E., & Benjamin, J. (1993). *Campylobacter* bacteraemia in England and Wales, 1981-91. *Epidemiology and infection*, 110(03), 567–573.

[341] Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, 139(1), 457.

[342] Smith, D., Jamieson, I., & Peach, R. (2005). Importance of ground weta (hemiandrus spp.) in stoat (mustela erminea) diet in small montane valleys and alpine grasslands. *New Zealand Journal of Ecology*, 29(2), 207–214.

[343] Smith, J., Smith, N., O'rourke, M., & Spratt, B. (1993). How clonal are bacteria? *Proceedings of the National Academy of Sciences of the United States of America*, 90(10), 4384.

[344] Sokal, R. & Michener, C. (1958). A statistical method for evaluating systematic relationships. *Kansas University science bulletin*, 38, 1409–1438.

[345] Sopwith, W., Birtles, A., Matthews, M., Fox, A., Gee, S., Painter, M., Regan, M., Syed, Q., & Bolton, E. (2008). Identification of potential environmentally adapted *Campylobacter jejuni* strain, United Kingdom. *Emerging Infectious Diseases*, 14(11), 1769.

[346] Stamatakis, A., Hoover, P., & Rougemont, J. (2008). A rapid bootstrap algorithm for the RAxML web servers. *Systematic Biology*, 57(5), 758–771.

[347] Stanley, K. & Jones, K. (2003). Cattle and sheep farms as reservoirs of *Campylobacter*. *Journal of Applied microbiology*, 94, 104–113.

[348] Steel, M., Hendy, M., & Penny, D. (1993). Parsimony can be consistent! *Systematic biology*, (pp. 581–587).

[349] Steele, M., McNab, B., Fruhner, L., DeGrandis, S., Woodward, D., & Odumeru, J. (1998). Epidemiological typing of *Campylobacter* isolates from meat processing plants by pulsed-field gel electrophoresis, fatty acid profile typing, serotyping, and biotyping. *Applied and Environmental Microbiology*, 64(7), 2346.

[350] Stephens, D., Hoffman, L., & McGee, Z. (1983). Interaction of *Neisseria meningitidis* with human nasopharyngeal mucosa: attachment and entry into columnar epithelial cells. *Journal of Infectious Diseases*, 148(3), 369–376.

[351] Stephens, M. (2001). Inference under the coalescent. *Handbook of Statistical Genetics*.

[352] Stephens, M. & Donnelly, P. (2000). Inference in molecular population genetics. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, (pp. 605–655).

[353] Stern, N., Hernandez, M., Blankenship, L., Deibel, K., Doores, S., Doyle, M., Ng, H., Pierson, M., Sofos, J., Sveum, W., et al. (1985). Prevalence and distribution of *Campylobacter jejuni* and *Campylobacter coli* in retail meats. *Journal of Food Protection (USA)*.

[354] Stevens, D. & Kaplan, E. (2000). *Streptococcal infections: clinical aspects, microbiology, and molecular pathogenesis*. Oxford University Press, USA.

[355] Stinchcombe, J. & Hoekstra, H. (2007). Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity*, 100(2), 158–170.

[356] Struelens, M. (1998). Molecular epidemiologic typing systems of bacterial pathogens: current issues and perpectives. *Memórias do Instituto Oswaldo Cruz*, 93, 581–586.

[357] Suerbaum, S., Lohrengel, M., Sonnevend, A., Ruberg, F., & Kist, M. (2001). Allelic diversity and recombination in *Campylobacter jejuni*. *Journal of Bacteriology*, 183(8), 2553.

[358] Suerbaum, S., Smith, J., Bapumia, K., Morelli, G., Smith, N., Kunstmann, E., Dyrek, I., & Achtman, M. (1998). Free recombination within helicobacter pylori. *Proceedings of the National Academy of Sciences*, 95(21), 12619.

[359] Swofford, D. (2003). PAUP*: phylogenetic analysis using parsimony, version 4.0 b10.

[360] Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105(2), 437.

[361] Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3), 585.

[362] Tamura, K., Nei, M., & Kumar, S. (2004). Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proceedings of the National Academy of Sciences of the United States of America*, 101(30), 11030.

[363] Tauxe, R., Nachamkin, I., Blaser, M., & Tompkins, L. (1992). Epidemiology of *Campylobacter jejuni* infections in the United States and other industrialized nations. *American Society for Microbiology, Washington, DC(USA).*

[364] Te Velthuis, A. & Bagowski, C. (2008). Linking fold, function and phylogeny: A comparative genomics view on protein (domain) evolution. *Current Genomics*, 9(2), 88.

[365] Teunis, P. & Havelaar, A. (2000). The beta poisson dose-response model is not a single-hit model. *Risk Analysis*, 20(4), 513–520.

[366] Teunis, P., Van den Brandhof, W., Nauta, M., Wagenaar, J., Van den Kerkhof, H., Van Pelt, W., et al. (2005). A reconsideration of the *Campylobacter* dose-response relation. *Epidemiology and Infection*, 133(4), 583–592.

[367] Thakur, S. & Gebreyes, W. (2005). *Campylobacter coli* in swine production: antimicrobial resistance mechanisms and molecular epidemiology. *Journal of Clinical Microbiology*, 43(11), 5705.

[368] Thomas, C. & Nielsen, K. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature Reviews Microbiology*, 3(9), 711–721.

[369] Tibayrenc, M. (1998). Beyond strain typing and molecular epidemiology: integrated genetic epidemiology of infectious diseases. *Parasitology Today*, 14(8), 323–329.

[370] Timoney, J. (2004). The pathogenic equine streptococci. *Veterinary Research*, 35(4), 397–409.

[371] Tomita, T., Meehan, B., Wongkattiya, N., Malmo, J., Pullinger, G., Leigh, J., & Deighton, M. (2008). Identification of *Streptococcus uberis* multilocus sequence types highly associated with mastitis. *Applied and Environmental Microbiology*, 74(1), 114.

[372] Towle, A. (1999). *Modern biology*. Holt Rinehart & Winston.

[373] Trewick, S. (1997). Flightlessness and phylogeny amongst endemic rails (aves: Rallidae) of the New Zealand region. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 352(1352), 429.

[374] Trewick, S. & Gibb, G. (2010a). Vicars, tramps and assembly of the New Zealand avifauna: a review of molecular phylogenetic evidence. *Ibis*, 152, 226.

[375] Trewick, S. & Gibb, G. (2010b). Vicars, tramps and assembly of the New Zealand avifauna: a review of molecular phylogenetic evidence. *Ibis*, 152, 226.

[376] Trewick, S., Worthy, T., William, G., & Jamieson, I. (2000). Origins and prehistoric ecology of takahe: flightless porphyrio (aves: Rallidae). *The takahe*, 50, 31–48.

[377] Turk, D. (1984). The pathogenicity of *Haemophilus influenzae*. *Journal of Medical Microbiology*, 18(1), 1–16.

[378] Turnbull, P. (1996). *Bacillus*: Barron's medical microbiology. *University of Texas Medical Branch*.

[379] Unicomb, L., Dalton, C., Gilbert, G., Becker, N., & Patel, M. (2008). Age-specific risk factors for sporadic *Campylobacter* infection in regional australia. *Foodborne Pathogens and Disease*, 5(1), 79–85.

[380] Upton, G. & Cook, I. (2002). Oxford dictionary of statistics.

[381] Van Bergen, M., Dingle, K., Maiden, M., Newell, D., van der Graaf-Van Bloois, L., van Putten, J., & Wagenaar, J. (2005). Clonal nature of *Campylobacter fetus* as defined by multilocus sequence typing. *Journal of Clinical Microbiology*, 43(12), 5888.

[382] van Lieshout, M., Blok, D., Wieland, C., de Vos, A., van't Veer, C., & van der Poll, T. (2012). Differential roles of MyD88 and TRIF in hematopoietic and resident cells during murine gram-negative pneumonia. *Journal of Infectious Diseases*.

[383] Vandamme, P. (2000). Taxonomy of the family campylobacteraceae. *Campylobacter*, 2, 3–26.

[384] Vandamme, P. & De Ley, J. (1991). Proposal for a new family, campylobacteraceae. *International Journal of Systematic Bacteriology*, 41(3), 451–455.

[385] Vandamme, P., Debruyne, L., De Brandt, E., & Falsen, E. (2010). Reclassification of bacteroides ureolyticus as *Campylobacter ureolyticus* comb. nov., and emended description of the genus *Campylobacter*. *International Journal of Systematic and Evolutionary Microbiology*, 60(9), 2016–2022.

[386] Vandamme, P., Falsen, E., Rossau, R., Hoste, B., Segers, P., Tytgat, R., & De Ley, J. (1991). Revision of *Campylobacter*, *Helicobacter*, and *Wolinella* taxonomy: emendation of generic descriptions and proposal of *Arcobacter* gen. nov. *International Journal of Systematic Bacteriology*, 41(1), 88–103.

[387] Veron, M. & Chatelain, R. (1973). Taxonomic study of the genus *Campylobacter* sebald and véron and designation of the neotype strain for the type species, *Campylobacter fetus* (Smith and Taylor) Sebald and Véron. *International Journal of Systematic and Evolutionary Microbiology*, 23(2), 122.

[388] Vos, M. & Didelot, X. (2008). A comparison of homologous recombination rates in bacteria and archaea. *The ISME Journal*, 3(2), 199–208.

[389] Wallis, G. & Trewick, S. (2009). New Zealand phylogeography: evolution on a small continent. *Molecular Ecology*, 18(17), 3548–3580.

[390] Wang, Y. & Taylor, D. (1990). Natural transformation in *Campylobacter* species. *Journal of Bacteriology*, 172(2), 949.

[391] Ward, M. & Watt, P. (1972). Adherence of *Neisseria gonorrhoeae* to urethral mucosal cells: an electron-microscopic study of human gonorrhea. *Journal of Infectious Diseases*, 126(6), 601–605.

[392] Wassenaar, T. & Blaser, M. (1999). Pathophysiology of *Campylobacter jejuni* infections of humans. *Microbes and Infection*, 1(12), 1023–1033.

[393] Wassenaar, T. & Newell, D. (2000). Genotyping of *Campylobacter* spp. *Applied and Environmental Microbiology*, 66(1), 1.

[394] Weinstock, H., Berman, S., & Cates Jr, W. (2004). Sexually transmitted diseases among American youth: incidence and prevalence estimates, 2000. *Perspectives on Sexual and Reproductive Health*, 36(1), 6–10.

[395] Weir, B. & Cockerham, C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, (pp. 1358–1370).

[396] Wells, S., Ott, S., & Hillberg Seitzinger, A. (1998). Key health issues for dairy cattle – new and old. *Journal of Dairy Science*, 81(11), 3029–3035.

[397] Wempe, J., Genigeorgis, C., Farver, T., & Yusufu, H. (1983). Prevalence of *Campylobacter jejuni* in two California chicken processing plants. *Applied and Environmental Microbiology*, 45(2), 355.

[398] Whittaker, R. (1952). A study of summer foliage insect communities in the great smoky mountains. *Ecological Monographs*, 22(1), 2–44.

[399] WHO (2001). *The Increasing Incidence of Human Campylobacteriosis: Report and Proceedings of a WHO Consultation of Experts, Copenhagen, Denmark, 21-25 November 2000.* World Health Organization. Dept. of Communicable Disease Surveillance and Response.

[400] Wilkinson, M. (1994a). Common cladistic information and its consensus representation: reduced adams and reduced cladistic consensus trees and profiles. *Systematic Biology*, 43(3), 343.

[401] Wilkinson, M. (1994b). The permutation method and character compatibility. *Systematic Biology*, (pp. 274–277).

[402] Wilkinson, M. & Thorley, J. (2001). Efficiency of strict consensus trees. *Systematic Biology*, 50(4), 610–613.

[403] Wilson, D., Gabriel, E., Leatherbarrow, A., Cheesbrough, J., Gee, S., Bolton, E., Fox, A., Hart, C., Diggle, P., & Fearnhead, P. (2009). Rapid evolution and the importance of recombination to the gastroenteric pathogen *Campylobacter jejuni*. *Molecular Biology and Evolution*, 26(2), 385.

[404] Wilson, D. & McVean, G. (2006). Estimating diversifying selection and functional constraint in the presence of recombination. *Genetics*, 172(3), 1411–1425.

[405] Wilson, N., Baker, M., Simmons, G., & Shoemack, P. (2006). New Zealand should control *Campylobacter* in fresh poultry before worrying about flies. *Journal of the New Zealand Medical Association*, 119(1242).

[406] Wim Ang, C., Jacobs, B., & Laman, J. (2004). The Guillain-Barré syndrome: a true case of molecular mimicry. *Trends in Immunology*, 25(2), 61–66.

[407] Withington, S. & Chambers, S. (1997). The cost of campylobacteriosis in New Zealand in 1995. *The New Zealand Medical Journal*, 110(1046), 222.

[408] Woodhead, M., Macfarlane, J., McCracken, J., Rose, D., & Finch, R. (1987). Prospective study of the aetiology and outcome of pneumonia in the community. *The Lancet*, 329(8534), 671–674.

[409] Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution*, 11(9), 367–372.

[410] Yang, Z. (1998). On the best evolutionary rate for phylogenetic analysis. *Systematic Biology*, 47(1), 125.

[411] Yang, Z. & Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nature Reviews Genetics*, 13(5), 303–314.

[412] Yorke, J., Hethcote, H., & Nold, A. (1978). Dynamics and control of the transmission of gonorrhea. *Sexually Transmitted Diseases*, 5(2), 51.

[413] Young, H., Harris, A., & Tapsall, J. (1984). Differentiation of gonococcal and non-gonococcal neisseriae by the superoxol test. *The British journal of venereal diseases*, 60(2), 87–89.

[414] Yu, S., Fearnhead, P., Holland, B., Biggs, P., Maiden, M., & French, N. (2012). Estimating the relative roles of recombination and point mutation in the generation of single locus variants in *Campylobacter jejuni* and *Campylobacter coli*. *Journal of Molecular Evolution*, (pp. 1–8).

[415] Yule, G. (1925). A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 213, 21–87.

[416] Zerbino, D. & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5), 821–829.

[417] Zhang, Q., Meitzler, J., Huang, S., & Morishita, T. (2000). Sequence polymorphism, predicted secondary structures, and surface-exposed conformational epitopes of *Campylobacter* major outer membrane protein. *Infection and Immunity*, 68(10), 5679.

[418] Zhong, B., Deusch, O., Goremykin, V., Penny, D., Biggs, P., Atherton, R., Nikiforova, S., & Lockhart, P. (2011). Systematic error in seed plant phylogenomics. *Genome Biology and Evolution*, 3, 1340.

[419] Zhu, P., Van Der Ende, A., Falush, D., Brieske, N., Morelli, G., Linz, B., Popovic, T., Schuurman, I., Adegbola, R., Zurth, K., et al. (2001). Fit genotypes and escape variants of subgroup III *Neisseria meningitidis* during three pandemics of epidemic meningitis. *Proceedings of the National Academy of Sciences*, 98(9), 5234.

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Shoukai Yu

Name/Title of Principal Supervisor: Prof Nigel French

Name of Published Research Output and full reference:

Title:Estimating the Relative Roles of Recombination and Point Mutation in the Generation of Single Locus Variants in Campylobacter jejuni and Campylobacter coli

author: Yu, S. and Fearnhead, P. and Holland, B.R. and Biggs, P. and Maiden, M. and French, N.
journal: Journal of molecular evolution
pages: 1--8
year: 2012
publisher: Springer

In which Chapter is the Published Work: Chapter 3

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate:

  and / or

- Describe the contribution that the candidate has made to the Published Work:

  The candidate conducted all the analysis under the suggestions from all the supervisors

Shoukai Yu
Digitally signed by Shoukai Yu
DN: cn=Shoukai Yu, o=Massey University, ou=IVABS, email=s.yu1@massey.ac.nz, c=NZ
Date: 2012.11.27 16:31:28 +13'00'

Candidate's Signature

2012/11/27
Date

Nigel French
Digitally signed by Nigel French
DN: cn=Nigel French, o=Massey University, ou=IDReC, email=n.p.french@massey.ac.nz, c=NZ
Date: 2012.12.02 14:12:08 +13'00'

Principal Supervisor's signature

2/12/12
Date