Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.
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

1 Introduction .............................................. 1
   1.1 General background .................................. 1
   1.2 Objectives .......................................... 5
   1.3 Organisation of the thesis ............................ 5

2 Literature review ........................................... 7
   2.1 Campylobacter ......................................... 7
      2.1.1 General information .............................. 7
      2.1.2 Campylobacter epidemiology ..................... 8
      2.1.3 Molecular biology of Campylobacter ............. 12
      2.1.4 Flagella and the major outer membrane proteins . 14
   2.2 Typing methods ....................................... 15
   2.3 Multilocus sequence typing (MLST) .................. 17
      2.3.1 Selection of MLST .............................. 18
   2.4 Evolutionary methods and phylogenetics .............. 21
      2.4.1 Phylogenetic networks .......................... 21
      2.4.2 Assessing confidence in phylogenetic trees ....... 24
      2.4.3 Specific phylogenetic methods .................. 25
      2.4.4 Sequence based methods ........................ 26
      2.4.5 Bayesian methods ................................ 27
      2.4.6 Consensus trees and consensus split networks . 29
   2.5 Population genetics ................................... 30
      2.5.1 Coalescent theory ............................... 30
7 Conclusion and further directions

7.1 Conclusion

7.1.1 The analysis of SLVs

7.1.2 The role of geographical isolation in the evolution of Campylobacter

7.1.3 Analysis on targeted gene reference sets

7.2 Further directions

Bibliography
List of Figures

2.1 A portion of gene porA on Campylobacter jejuni strain NCTC11168 16
2.2 The positions of MLST loci on the strain NCTC 11168 20
2.3 Two trees of the same set of taxa, but with different tree shapes 23
2.4 Split network and reticulated network 24
3.1 An eBURST diagram 40
3.2 SLVs of PubMLST data for C. jejuni 48
3.3 SLVs of PubMLST data for C. coli 49
4.1 Number of nucleotide differences in SLVs 67
4.2 Number of nucleotide differences in SLVs for all tested bacteria 69
5.1 Maps of NZ and UK 85
5.2 Neighbor Net of 1-PSI matrix 90
5.3 Neighbor-Net plot of pairwise Fst values at different levels 91
5.4 Rarefaction plot for UK and NZ data on human host source 94
5.5 Rarefaction plot for UK and NZ data on poultry host source 95
5.6 Rarefaction plot for UK and NZ data on ruminant host source 96
5.7 Reconstruction of the phylogeny of some NZ specific strains 100
5.8 Structure analysis results 106
6.18 Mapping events on the phylogeny of ST-474 116
6.1 The clonal genealogy was generated by SimMLST 117
6.2 The UPGMA tree for the simulated dataset 118
6.3 The NJ tree for the simulated dataset 119
6.4 The strict consensus tree of MP for the simulated dataset 120
## List of Tables

<p>| 3.1 | Example one for an SLV | 38 |
| 3.2 | Example two for an SLV | 38 |
| 3.3 | Estimates for <em>C. jejuni</em> | 47 |
| 3.4 | Estimates for <em>C. coli</em> clade 1 | 50 |
| 3.5 | Comparison of different prior parameters <em>C. jejuni</em> | 53 |
| 3.6 | Comparison of different prior parameters <em>C. coli</em> | 54 |
| 4.1 | Number of STs, SLVs and ratio of SLVs to STs | 66 |
| 4.2 | Estimates of several bacteria for MLST | 70 |
| 4.3 | Estimates of several bacteria for MLST (Median) | 70 |
| 4.4 | Comparison between $\rho/\theta$ and $r/m$ | 71 |
| 4.5 | Estimates for <em>B. cereus</em> | 76 |
| 4.6 | Estimates for <em>E. faecium</em> | 76 |
| 4.7 | Estimates for <em>H. influenzae</em> | 77 |
| 4.8 | Estimates for <em>K. pneumoniae</em> | 77 |
| 4.9 | Estimates for <em>S. uberis</em> | 78 |
| 4.10 | Estimates for <em>S. zooepidemicus</em> | 78 |
| 4.11 | Estimates for <em>S. aureus</em> | 79 |
| 4.12 | Estimates for <em>N. lactamica</em> | 79 |
| 4.13 | Estimates for <em>N. gonorrhoeae</em> | 80 |
| 4.14 | Estimates for <em>N. meningitidis</em> | 80 |
| 5.1 | AMOVA with country defined as higher grouping | 93 |
| 5.2 | AMOVA with host defined as higher grouping | 93 |</p>
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.3</td>
<td>BEAST results of the mean of split time</td>
<td>98</td>
</tr>
<tr>
<td>6.1</td>
<td>Symmetric-difference matrix</td>
<td>113</td>
</tr>
<tr>
<td>6.2</td>
<td>Number of variants</td>
<td>134</td>
</tr>
</tbody>
</table>