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	Intr	oducti	ion	1
	1.1	Gener	al background	1
	1.2	Objec	tives	5
	1.3	Organ	isation of the thesis	5
2	Lite	erature	e review	7
	2.1	Camp	ylobacter	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 memberane proteins	14
	2.2	Typin	g methods	15
	2.3	Multil	ocus sequence typing (MLST)	17
		2.3.1	Selection of MLST	18
	2.4	Evolu	tionary 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	Popula	ation genetics	30
		2.5.1	Coalescent theory	30

		2.5.2	The comparison between phylogenetic model and coalescent methods	33	
	2.6	Softwa	are	35	
3	Est in t and	imatin he gen <i>Camp</i>	g the relative roles of recombination and point mutation neration of single locus variants in <i>Campylobacter jejuni</i> pylobacter coli	37	
	3.1	Backg	round	37	
4	The relative roles of recombination and point mutation to the gen- eration of single locus variants in a range of bacterial pathogens 60				
	4.1	Summ	ary	60	
	4.2	Introd	uction	61	
		4.2.1	A brief introduction into the selected bacteria	63	
	4.3	Mater	ial and Methods	65	
		4.3.1	Isolates	65	
		4.3.2	Modelling procedure	65	
	4.4	Result	js	67	
		4.4.1	The distribution of nucleotide differences within SLV for each bacterium	67	
		4.4.2	Estimates for several bacteria by loci	71	
	4.5	Discus	ssion	72	
	4.6	Supple	ementary material	75	
		4.6.1	Recombination and mutation models	75	
	4.7	Estim	ates for tested bacteria by loci (tables)	76	
5	Inve of C	estigat Campyl	ing the impact of geographical isolation on the evolution lobacter by comparing New Zealand and United Kingdom		
	data	asets		81	
	5.1	Summ	ary	81	
	5.2	Introd	uction	82	
	5.3	Mater	ial and Methods	84	
		5.3.1	Isolates	84	

		5.3.2	Analysis overview	3
		5.3.3	Population genetics tools and network methods 86	3
		5.3.4	Bayesian Phylogenetic analysis	3
	5.4	Result	s	3
		5.4.1	Fst and AMOVA at different levels)
		5.4.2	BEAST analysis	7
	5.5	Discus	ssion \ldots \ldots \ldots \ldots \ldots $$ 99	9
	5.6	Addit	100 on all structure analysis 100 100 100	4
		5.6.1	Bayesian cluster analysis $\ldots \ldots 10^4$	4
		5.6.2	Structure analysis results	õ
		5.6.3	Discussion about structure analysis	õ
6	\mathbf{Est}	imatin	g the clonal genealogy for ST-474, a commonly found	
	Nev	v Zeal	and <i>Campylobacter</i> sequence type 108	3
	6.1	Introd	luction $\ldots \ldots \ldots$	3
	6.2	Phylo	genetic analysis and methods)
		6.2.1	Data)
		6.2.2	Phylogenetic analysis and methods comparison	1
		6.2.3	Mapping events on the ST-474 branch	1
		6.2.4	Compatibility	1
	6.3	Result	ss	2
		6.3.1	The phylogeny of the simulated dataset	2
		6.3.2	Results for the targeted gene reference set	3
		6.3.3	Results for the MLST dataset	4
		6.3.4	Mapping events on ST-474 related phylogeny	3
		6.3.5	Compatibility	5
	6.4	Discus	ssion \ldots \ldots \ldots \ldots \ldots 139	9
	6.5	Ackno	wledgements $\ldots \ldots 145$	3
	6.6	Apper	ndix A: Variant loci on phylogeny of ST-474	5

7	Con	clusior	n and further directions	150
	7.1	Conclu	$sion \ldots \ldots$	150
		7.1.1	The analysis of SLVs	152
		7.1.2	The role of geographical isolation in the evolution of Campy- lobacter	152
		7.1.3	Analysis on targeted gene reference sets	153
	7.2	Furthe	r directions	154
Bi	bliog	raphy		156

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 \ldots	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 <i>C. jejuni</i>	48
3.3	SLVs of PubMLST data for <i>C. coli</i>	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 \ldots .	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

6.5	The ML tree for the simulated dataset \hdots
6.6	ClonalFrame result for the simulated dataset
6.7	UPGMA tree for TGRS data
6.8	Neighbor-Joining plot for TGRS data
6.9	Maximum parsimony plot for TGRS data
6.10	ML plot for TGRS data
6.11	Reconstruction of the phylogeny of TGRS by ClonalFrame (Result 1) 127
6.12	Reconstruction of the phylogeny of TGRS by ClonalFrame (Result 2) 128
6.13	UPGMA tree for 33 STs
6.14	Neighbor-Joining plot for 33 STs
6.15	Strict consensus tree of 96 Maximum parsimony trees for 33 STs $~$ 131 $$
6.16	ML plot for 33 STs
6.17	Phylogeny plot for 33 STs by ClonalFrame
6.20	Maximum parsimony tree for cluster one
6.21	Maximum parsimony tree for cluster two tree 1
6.19	Comparability plot
6.22	Maximum parsimony tree for cluster two tree 2
6.23	Maximum parsimony tree for cluster three
6.24	Maximum parsimony tree for cluster four

List of Tables

3.1	Example one for an SLV	38
3.2	Example two for an SLV	38
3.3	Estimates for <i>C. jejuni</i>	47
3.4	Estimates for <i>C. coli</i> clade 1	50
3.5	Comparison of different prior parameters C. jejuni	53
3.6	Comparison of different prior parameters <i>C.coli</i>	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 ρ/θ and r/m	71
4.5	Estimates for <i>B. cereus</i>	76
4.6	Estimates for <i>E. faecium</i>	76
4.7	Estimates for <i>H. influenzae</i>	77
4.8	Estimates for K. pneumoniae	77
4.9	Estimates for <i>S. uberis</i>	78
4.10	Estimates for S. zooepidemicus	78
4.11	Estimates for <i>S. aureus</i>	79
4.12	Estimates for <i>N. lactamica</i>	79
4.13	Estimates for N. gonorrhoeae	80
4.14	Estimates for <i>N. meningitidis</i>	80
5.1	AMOVA with country defined as higher grouping	93
5.2	AMOVA with host defined as higher grouping	93

5.3	BEAST results of the mean of split time	98
6.1	Symmetric-difference matrix	113
6.2	Number of variants	134