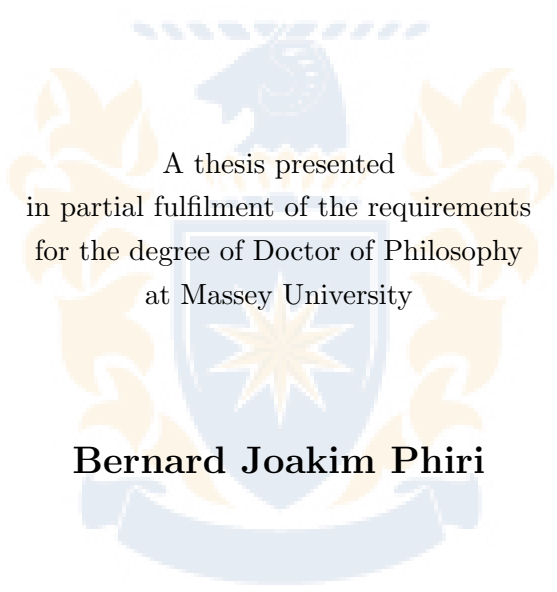


Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

# Estimating the public health risk associated with drinking water in New Zealand

The crest of Massey University is centered on the page. It features a blue shield with a white star, flanked by golden laurel branches. Above the shield is a blue horse head, and below it is a blue banner.

A thesis presented  
in partial fulfilment of the requirements  
for the degree of Doctor of Philosophy  
at Massey University

**Bernard Joakim Phiri**

Institute of Veterinary, Animal and Biomedical Sciences  
Massey University  
Palmerston North, New Zealand

2015  
(submitted 20 January 2015)



# Preamble

It always seems impossible until its done.

— *Nelson Rolihlahla Mandela (1918–2013)*



# Abstract

This thesis is concerned with the application of both epidemiological and molecular tools to assess the drinking water safety in New Zealand. Compromised drinking water safety is commonly manifested as gastrointestinal illness. The studies in this thesis were motivated by the desire to find ways of reducing the burden of such illness in the human population. Although the studies were conducted in the New Zealand setting the methodologies can be readily applied elsewhere.

The first study investigated the factors associated with the presence of microbes in raw water intended for public consumption. Random forest, an established non-parametric statistical method, was used to model data with possible complex interactions and identified variables that were predictive of the presence of microbes in raw drinking water. *E. coli*, which is widely used as a microbial contamination indicator in the water industry, was found to be a better predictor of the presence/absence of *Campylobacter* (bacteria) than protozoan microbes (*Cryptosporidium* and *Giardia*). This suggests that alternative methods of determining the presence/absence of pathogens in water should be developed. In the second study, the relationship between river flow and reports of cases of gastrointestinal illness was described using the distributed lag modelling approach. This revealed a positive relationship that peaked around 10 days after high flow. Further, the river flow-gastrointestinal illness relationship was stronger in small drinking distribution networks than in large ones. The small drinking water distribution networks could be targeted for facility upgrade in order to enhance their ability to deliver microbiologically safer drinking water.

The third study utilised culture-dependent methods to assess the public health risk associated with drinking water supplied at outdoor recreation facilities — campgrounds. Water treatment using methods such as ultra violet and chemical treatment were found to be highly beneficial for the campgrounds to deliver drinking water that was microbiologically safe and compliant with water safety regulations. The profiles and functional factors of drinking water microbial communities are described in the fourth study. Techniques from the fast-growing field of metagenomics were employed for this purpose. The capability of metagenomic techniques to detect multiple pathogens in a single assay was demonstrated. This has the potential to greatly enhance the specificity and sensitivity of microbial water quality testing.



# Acknowledgements

I could not have accomplished the research work presented in this thesis without the excellent help and guidance that I received from my PhD supervisors. Thank you Nigel French for inviting me to carry out this research work and for teaching me a great deal of things about science. You always had a suggestion on how to move forward when faced with challenges. To Patrick Biggs, thank you for your kind and enthusiastic support even at short notice. You have been inspirational in my approach to bioinformatics and the presentation of genomic information. Thank you Mark Stevenson for your attention to detail and keeping me reminded of the need to apply the epidemiological principles appropriately in my work. To Deb Prattley, thank you for your unfailing support and kind guidance in presenting my research as a coherent story. To Paul Rainey, you provided that critical suggestion that got things moving again when the metagenomic DNA extraction was stalling, thank you.

Thank you to all my fellow postgraduate students at the Epicentre and Hopkirk Research Institute for being part of my journey and sharing your experiences with me along the way. To Christine Cunningham, Wendy Maharey, Jacque Mackenzie and Simon Verschaffelt, thank you for your administrative and computational support. Thank you to the mEpiLab team that provided me with the much needed laboratory support, in particular Angie Reynolds, Anthony Pita, Niluka Velathanthiri, Julie Collins-Emerson, Ann Midwinter, Neville Haack, Errol Kwan, Rhuksana Akhter, Lynn Rogers and Sarah Moore. Special thanks to the New Zealand Genomic Limited team, Lorraine Berry, Richard Fong and Trish McLenachan, for going beyond the call of duty to help me resolve the metagenomic sequencing issues.

I sincerely thank the Allan Wilson Centre and Massey University for funding this project. Thanks to the Department of Conservation for allowing me to conduct my research on their campgrounds. In particular, thanks go to the various campground managers that provided me with campground information and kindly showed me the routes to the water abstraction sites. Thank you to the National Institute of Water and Atmospheric Research as well as the sixteen regional councils that emailed me the river flow data. Thank you to the Institute of Environmental Science and Research Limited for providing me with disease case and drinking water supply data.

Most importantly, many thanks go to my family for being understanding and patient with me as I carried out this work. To my partner Eve, heartfelt thanks for making our home a warm and loving place to live. Completion of this work would have been extremely difficult



## Acknowledgements

---

without your loving support. To Joe and Sam, thank you for letting your dad complete his PhD research trouble free. It has been a pleasure watching you blossom into fine young men during the four years of my PhD work, I love you very much.

# Acronyms

- <sup>m</sup>EpiLab** molecular epidemiology and public health laboratory. 47, 48, 50, 110, 112, 114, 116, 138, 165
- aspA*** aspartase. 116
- glnA*** glutamine synthetase. 116
- gltA*** citrate synthase. 116
- glyA*** serine hydroxy methyl transferase. 116
- pgm*** phospho glucomutase. 116
- tkt*** transketolase. 116
- uncA*** adenosine triphosphate synthase alpha subunit. 116
- BLAST** basic local alignment search tool. 143, 161
- FLASH** fast length adjustment of short reads. 141, 143, 161
- MEGAN** metagenome analyzer. 143, 161
- PAUDA** protein alignment using a DNA aligner. 143, 161
- QIIME** quantitative insights into microbial ecology. 141, 161, 166
- AIC** Akaike information criterion. 86
- ATP** adenosine triphosphate. 8
- BIOM** biological observation matrix. 142
- BLUE** best linear unbiased estimator. 83
- CART** classification and regression trees. 53, 54
- CCA** canonical correspondence analysis. 79, 142, 145, 155, 156, 158
- DAF** dissolved air floatation. 20
- DAPI** 4',6-diamidino-2-phenylindole. 51, 52
- dATP** deoxyadenosine triphosphate. 5, 25
- dCTP** deoxycytidine triphosphate. 5, 25
- ddATP** dideoxyadenosine triphosphate. 5, 6
- ddCTP** dideoxycytidine triphosphate. 5, 6
- ddGTP** dideoxyguanosine triphosphate. 6
- ddNTP** dideoxyribonucleotide triphosphate. 5, 6
- ddTTP** dideoxythymidine triphosphate. 6
- dGTP** deoxyguanosine triphosphate. 5, 25
- DLM** distributed lag model. 84–86
- DLNM** distributed lag non-linear model. 85–87, 94, 95, 100–104, 174–180
- DNA** deoxyribonucleic acid. 5–10, 25, 26, 35, 40, 112, 114, 115, 136–140, 149, 157, 158, 160, 164, 165, 194–197
- dNTP** deoxyribonucleotide triphosphate. 5, 7, 8, 25
- DOC** Department of Conservation. 105–107, 121, 127–130, 137, 138, 155, 158, 159
- dsDNA** double stranded DNA. 6
- dTTP** deoxythymidine triphosphate. 5, 25
- DWSNZ** drinking water standards for New Zealand. 17, 28, 33, 41, 115, 120, 122, 128, 129, 160
- ELISA** enzyme-linked immunosorbent assay. 24–26, 42
- emPCR** emulsion polymerase chain reaction. 8
- ESR** Institute of Environmental Science and Research Limited. 17, 78
- ESRI** Environmental Systems Research Institute. 53, 108
- FC** faecal coliform. 34
- FISH** fluorescence *in situ* hybridisation. 25, 42
- GDH** glutamate dehydrogenase. 114
- GDP** gross domestic product. 105
- GLM** generalised linear model. 119, 120, 125, 127
- GLMM** generalised linear mixed model. 57, 58, 67, 68, 71, 73, 119, 120, 127, 128
- gp60** 60-kDa glycoprotein. 26
- GPS** global positioning system. 84, 107, 108, 110, 162, 163, 166
- HACCP** hazard analysis critical control point. 4
- HAdV** human adenovirus. 22
- HPyV** human polyomavirus. 22
- HSP** heat-shock protein. 35
- IMS** immunomagnetic separation. 51
- ISFET** ion-sensitive field-effect transistors. 8
- LAMP** loop-mediated isothermal amplification. 26, 42, 43
- LDA** linear discriminant analysis. 79
- LSU** large subunit. 34

## Acronyms

---

- MANOVA** multivariate analysis of variance. 79
- MAV** maximum acceptable value. 115, 120, 122
- mCCDA** modified charcoal cefoperazone deoxy-cholate agar. 50, 51, 112, 164
- MfE** Ministry for the Environment. 1
- MFT** membrane filter technique. 23, 42
- MLST** multilocus sequence typing. 110, 115, 124, 128, 160, 194, 201
- MoH** Ministry of Health. 16, 17, 46, 78
- MPN** most probable number. 23, 33, 53, 60, 106, 116, 122
- MST** microbial source tracking. 29, 30, 42
- MTF** multiple-tube fermentation. 23, 42
- NCBI** National Center for Biotechnology Information. 143, 149, 156
- NGS** next-generation sequencing. 5, 11, 36, 140
- NIWA** National Institute of Water and Atmospheric Research. 78, 165
- NZGL** New Zealand Genomics Limited. 138, 165
- OD** optical density. 138
- OOB** out-of-bag. 55
- OTU** operational taxonomic unit. 141, 142
- PBS** phosphate buffered saline. 51, 114, 198, 199
- PCA** principal component analysis. 46, 79, 80, 90, 94, 95
- PCR** polymerase chain reaction. 5, 9, 25, 26, 29, 35, 42, 51, 110, 112, 114, 115, 120, 122–124, 128, 139, 140, 156, 165, 167, 194
- PSI** proportional similarity index. 142, 143, 149, 155
- QMRA** quantitative microbiological risk assessment. 22
- qPCR** quantitative real-time polymerase chain reaction. 22, 26, 43
- RAM** random-access memory. 166
- rDNA** ribosomal deoxyribonucleic acid. 34
- REC** river environment classification. 53
- RF** random forest. 46, 53–57, 65, 70–74
- RNA** ribonucleic acid. 25, 40, 138
- rRNA** ribosomal ribonucleic acid. 25, 26, 30, 34–37, 112, 136–140, 145, 155, 165
- SMRT** single molecule real-time. 9
- SNP** single nucleotide polymorphism. 136
- SPInDel** species identification by insertions/deletions. 35
- ssDNA** single stranded deoxyribonucleic acid. 5, 6, 9
- SSU** small subunit. 34
- ST** sequence type. 110, 112, 117, 124, 125
- STEC** shiga toxin-producing *E. coli*. 16, 72
- TC** total coliform. 34
- UK** United Kingdom. 37
- USA** United States of America. 16, 37, 45, 46, 75, 76
- USEPA** United States Environmental Protection Agency. 51, 114
- UV** ultra violet. 17, 21, 97, 120, 121, 127, 129, 130, 159
- VTEC** verocytotoxin-producing *E. coli*. 16, 72
- WGS** whole genome shotgun. 34, 136, 137, 139, 140, 143, 145, 149, 155, 157, 165, 166
- WHO** World Health Organization. 13, 15, 31, 41, 142, 149, 157
- YLL** years of life lost. 75
- ZMW** zero-mode waveguide. 9

# Contents

	i
<b>Preamble</b>	<b>iii</b>
<b>Abstract</b>	<b>v</b>
<b>Acknowledgements</b>	<b>vii</b>
<b>Acronyms</b>	<b>ix</b>
<b>General introduction</b>	
1.1 Background . . . . .	1
1.2 Water quality . . . . .	1
1.2.1 The chemical aspect of water quality . . . . .	2
1.2.2 The physical aspect of water quality . . . . .	2
1.2.3 The biological aspect of water quality . . . . .	3
1.2.4 Genomic sequencing . . . . .	5
1.3 The structure of this thesis . . . . .	10
<b>Literature review</b>	
2.1 Background . . . . .	13
2.2 Drinking water sources and supply in New Zealand . . . . .	16
2.2.1 Drinking water sources . . . . .	16
2.2.2 Drinking Water supply system . . . . .	16
2.3 Drinking water treatment processes . . . . .	20
2.4 Common methods for detecting indicator organisms in drinking water . . . . .	21
2.4.1 Organism isolation-based methods . . . . .	23
2.4.2 Immunological methods . . . . .	24
2.4.3 Gene sequence-based methods . . . . .	25
2.4.4 Microbial compliance criteria for New Zealand . . . . .	28
2.5 Microbial source tracking . . . . .	29
2.6 Indicator organism detection in recreational water . . . . .	30
2.7 Pathogens in drinking water — New Zealand . . . . .	31
2.8 Metagenomics . . . . .	34
2.8.1 Metagenomics in drinking water . . . . .	36
2.8.2 Metagenomic research trends . . . . .	36
2.8.3 Microbial community profiles . . . . .	39
2.8.4 Microbial community functional genes . . . . .	39
2.9 Summary . . . . .	40
<b>Factors associated with the presence of pathogens in drinking water sources of New Zealand</b>	
3.1 Background . . . . .	45
3.2 Materials and methods . . . . .	47
3.2.1 Study sites . . . . .	47
3.2.2 Sample collection . . . . .	47

## CONTENTS

---

3.2.3	Laboratory procedures . . . . .	49
3.2.4	Data . . . . .	52
3.2.5	Statistical techniques . . . . .	52
3.2.6	Data analysis . . . . .	55
3.3	Results . . . . .	58
3.3.1	Descriptive statistics . . . . .	58
3.3.2	Random forest analysis . . . . .	63
3.3.3	Regression analysis . . . . .	66
3.4	Discussion . . . . .	69
<b>The relationship between river flow and notified cases of gastroenteritis in New Zealand</b>		
4.1	Background . . . . .	75
4.2	Materials and methods . . . . .	77
4.2.1	Study units . . . . .	77
4.2.2	Data . . . . .	78
4.2.3	Multivariate data analysis . . . . .	79
4.2.4	Geostatistical exploration . . . . .	80
4.2.5	Statistical modelling . . . . .	84
4.3	Results . . . . .	87
4.3.1	Descriptive statistics . . . . .	87
4.3.2	Multivariate analysis . . . . .	90
4.3.3	Geostatistical analysis . . . . .	91
4.3.4	Distributed lag analysis . . . . .	93
4.4	Discussion . . . . .	94
<b>The culture-based microbiology of drinking water on campgrounds in New Zealand</b>		
5.1	Background . . . . .	105
5.2	Materials and methods . . . . .	107
5.2.1	Study campground selection . . . . .	107
5.2.2	Campground water catchment geospatial characteristics . . . . .	108
5.2.3	Sample collection . . . . .	109
5.2.4	Laboratory techniques . . . . .	110
5.2.5	Laboratory processing: Faecal samples . . . . .	112
5.2.6	Laboratory processing: Water samples . . . . .	114
5.2.7	<i>Campylobacter</i> MLST . . . . .	115
5.2.8	Public health risk assessment . . . . .	115
5.3	Data analysis . . . . .	116
5.3.1	Regression analysis . . . . .	117
5.4	Results . . . . .	121
5.4.1	Campground descriptive statistics . . . . .	121
5.4.2	Geospatial descriptives . . . . .	121
5.4.3	Water samples . . . . .	122
5.4.4	Faecal samples . . . . .	123
5.4.5	Multilocus sequence typing analysis . . . . .	124
5.4.6	Regression analysis . . . . .	125
5.5	Discussion . . . . .	127
<b>The metagenome of drinking water on campgrounds in New Zealand</b>		
6.1	Background . . . . .	135
6.2	Materials and methods . . . . .	138
6.2.1	Study sites and sample collection . . . . .	138
6.2.2	Laboratory processing . . . . .	138

---

6.2.3	Metagenomic DNA sequencing . . . . .	139
6.2.4	Sequence Data . . . . .	140
6.2.5	Data analysis . . . . .	141
6.3	Results . . . . .	145
6.3.1	Descriptive statistics . . . . .	145
6.3.2	Public health hazard assessment . . . . .	145
6.4	Discussion . . . . .	155
<b>General discussion</b>		
7.1	Background . . . . .	159
7.1.1	Types of data . . . . .	160
7.2	Challenges and pitfalls . . . . .	162
7.2.1	Sample collection . . . . .	162
7.2.2	Sample processing . . . . .	164
7.2.3	Data management and analysis . . . . .	165
7.2.4	Future research work . . . . .	166
<b>Appendix</b>		
A.1	Literature review . . . . .	169
A.2	River flow study . . . . .	172
A.3	Catchment study . . . . .	181
A.4	Campground study . . . . .	192
References		

## List of Figures

2.1	The Waitakere and Waikato public drinking water catchments . . . . .	18
2.2	Schematic representation of the Wellington area drinking water distribution network . . .	19
2.3	Schematic diagrams showing the three major parts of a nucleotide . . . . .	27
2.4	Schematic representation of the polymerase chain reaction process. . . . .	28
2.5	Number of 16S and metagenomic publications per calendar year . . . . .	37
2.6	Number of 16S and metagenomic publications in the top fifteen countries . . . . .	38
2.7	Top twenty peer-reviewed journals publishing articles on 16S and metagenomics articles .	38
3.1	Location of the twenty study drinking water sources . . . . .	49
3.2	Schematic representation of a basic decision tree . . . . .	54
3.3	Drinking water catchments with high <i>E. coli</i> concentrations . . . . .	61
3.4	Concentrations of <i>Cryptosporidium</i> and <i>Giardia</i> in study catchment samples . . . . .	62
3.5	Percentage of positive samples for the four study microbes for each season . . . . .	63
3.6	Variable importance scores for drinking water catchment geospatial attributes . . . . .	65
3.7	Random effects for the generalised linear mixed models . . . . .	68
4.1	Water distribution zones and abstraction points . . . . .	78
4.2	Number of gastrointestinal cases 1997–2006, New Zealand . . . . .	88
4.3	Twenty zones with the highest incidence rates during the study period . . . . .	90
4.4	Location of water distribution zones with the highest gastroenteritis incidence rates . . .	91
4.5	PCA biplot of gastrointestinal illness annual incidence rates . . . . .	92
4.6	Median annual gastrointestinal illness case incidence rates, 1997–2006, New Zealand . . .	98
4.7	Kriged median annual gastrointestinal illness case incidence rates . . . . .	99
4.8	Relationship between distributed lag river flow and gastrointestinal illness, New Zealand .	100
4.9	Relationship between distributed lag river flow and gastrointestinal illness, S00079 . . . .	101
4.10	Relationship between distributed lag river flow and gastrointestinal illness, S00118 . . . .	102
4.11	Relationship between distributed lag river flow and gastrointestinal illness, S00217 . . . .	103
4.12	Relationship between distributed lag river flow and gastrointestinal illness, S00735 . . . .	104
5.1	Map showing the location of study campgrounds in New Zealand . . . . .	109
5.2	Types of samples collected from the study campgrounds . . . . .	111
5.3	Flow diagram showing the <i>Campylobacter</i> taxonomic designation process . . . . .	113
5.4	Median most probable number of <i>E. coli</i> in campground water samples . . . . .	124
5.5	Minimum spanning tree of campground <i>Campylobacter jejuni</i> and <i>C. coli</i> . . . . .	126
6.1	Flow diagram showing how 16S rRNA gene metagenomes were analysed. . . . .	144
6.2	Flow diagram showing how whole genome shotgun metagenomes were analysed. . . . .	144
6.3	Taxa richness indices for 16S metagenomes . . . . .	146
6.4	Canonical correspondence plot of 16S metagenomes . . . . .	148
6.5	<i>Campylobacteraceae</i> phylogenetic tree constructed using 16S metagenomes . . . . .	150
6.6	NeighborNet trees illustrating divergence of metagenome sources . . . . .	152
6.7	Bubble plot showing the abundance of virulence factors found in WGS metagenomes . . .	153
6.8	Bubble plot showing the abundance of resistance factors found in WGS metagenomes . . .	153
6.9	NeighborNet tree illustrating divergence of virulence factors found in WGS metagenomes .	154

---

6.10	NeighborNet tree illustrating divergence of resistance factors found in WGS metagenomes	154
A.1	Schematic representation of table connections in a MySQL relational database . . . . .	172
A.2	Bubble plots of gastrointestinal illness cases, 1997–2006, New Zealand . . . . .	173
A.3	Relationship between distributed lag river flow and gastrointestinal illness, S00041 . . . . .	174
A.4	Relationship between distributed lag river flow and gastrointestinal illness, S00082 . . . . .	175
A.5	Relationship between distributed lag river flow and gastrointestinal illness, S00106 . . . . .	176
A.6	Relationship between distributed lag river flow and gastrointestinal illness, S00123 . . . . .	177
A.7	Relationship between distributed lag river flow and gastrointestinal illness, S00200 . . . . .	178
A.8	Relationship between distributed lag river flow and gastrointestinal illness, S00233 . . . . .	179
A.9	Relationship between distributed lag river flow and gastrointestinal illness, S00268 . . . . .	180
A.10	Location of drinking water abstraction sites used in the distributed lag analysis . . . . .	181
A.11	Percentage of positive samples for the four study pathogens for each calendar month . . . . .	185
A.12	Land cover for the first six study catchments supplying surface raw water . . . . .	186
A.13	Land cover for the second six study catchments supplying surface raw water . . . . .	187
A.14	Land cover for the last four study catchments supplying surface raw water . . . . .	188
A.15	Lithology for the first six study catchments supplying surface raw water . . . . .	189
A.16	Lithology for the second six study catchments supplying surface raw water . . . . .	190
A.17	Lithology for the last four study catchments supplying surface raw water . . . . .	191
A.18	Land cover for study campground catchments located in the North Island, New Zealand . . . . .	192
A.19	Land cover for study campground catchments located in the South Island, New Zealand . . . . .	193
A.20	Phred scores for 16S sequences . . . . .	205
A.21	Phred scores for WGS sequences . . . . .	205



## List of Tables

2.1	Bacterial pathogens associated with drinking water . . . . .	14
3.1	Description of the twenty study drinking water sources . . . . .	48
3.2	Description of variables used in both RF and regression analyses . . . . .	57
3.3	Percentage of positive samples from the twenty study drinking water sources . . . . .	59
3.4	Number of sampling occasions and positive samples for each drinking water source . . . . .	60
3.5	Random Forest predictions . . . . .	66
3.6	GLMM estimating the presence/absence of <i>Campylobacter</i> in raw water . . . . .	67
3.7	GLMM estimating the <i>E. coli</i> concentrations in raw water . . . . .	67
3.8	GLMM estimating the presence/absence of <i>Cryptosporidium</i> in raw water . . . . .	69
3.9	GLMM estimating the presence/absence of <i>Giardia</i> in raw water . . . . .	69
4.1	Gastrointestinal illness annual incidence rates per 100 000 population for New Zealand and countries of similar socioeconomic status, 2013. . . . .	75
4.2	Description of variables used in a principal correspondence analysis . . . . .	89
4.3	Median drinking water distribution zone populations and median annual cases reported . . . . .	92
4.4	Drinking water abstraction sites used in distributed lag non-linear modelling . . . . .	93
5.1	Description of study campgrounds operated by DOC . . . . .	121
5.2	Number of water samples collected from DOC campgrounds . . . . .	123
5.3	Number of faecal samples, stratified by animal source, collected from DOC campgrounds . . . . .	125
5.4	Multilocus sequence types for faecal and water <i>Campylobacter</i> isolates . . . . .	131
5.5	Multilocus sequence types for <i>Campylobacter</i> isolated from water . . . . .	132
5.6	GLMM estimating the presence of <i>Campylobacter</i> in campground faecal samples . . . . .	132
5.7	GLMM estimating the concentration of <i>E. coli</i> in campground tap water . . . . .	132
5.8	GLM estimating the concentration of <i>E. coli</i> in campground intake water . . . . .	133
6.1	Number of samples sequenced for 16S rRNA gene and whole genome shotgun . . . . .	147
6.2	Bacterial species deposited in the NCBI database matched with metagenome taxa . . . . .	151
7.1	Computer software used for data processing, data analysis and thesis compilation. . . . .	161
A.1	A description of shapefiles used for geospatial data and their sources. . . . .	182
A.2	Geospatial data for sixteen surface water sources monitored for microbial contamination . . . . .	183
A.3	Geospatial data for four groundwater sources monitored for microbial contamination. . . . .	184
A.4	Constituents of the <i>Campylobacter</i> and <i>Giardia</i> polymerase chain reaction master mixes . . . . .	202
A.5	PCR conditions for selected <i>Campylobacter</i> and <i>Giardia</i> . . . . .	203
A.6	Encoding for the four bases (A, C, T, G) and ambiguous DNA sequences. . . . .	204
A.7	The 1-proportional similarity index values used for <i>Campylobacteraceae</i> taxa divergence . . . . .	206
A.8	The 1-proportional similarity index values used for WHO-recognised pathogen taxa divergence . . . . .	206