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

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 targetted 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, Rhukshana 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 reasearch 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

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

^{m}EpiLab molecular epidemiology and public health laboratory. 47, 48, 50, 110, 112, 114, 116, 138, 165

aspA aspatase. 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

MANOVA multivariate analysis of variance. 79
MAV maximum acceptible value. 115, 120, 122
mCCDA modified charcoal cefoperazone deoxycholate 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

List of Figures

List of Tables