

**MOLECULAR CHARACTERIZATION OF BACTERIAL
DIVERSITY IN NEW ZEALAND GROUNDWATER**

BY

KATUGAMPALAGE KOSALA AYANTHA SIRISENA

A thesis

Submitted to the Victoria University of Wellington

in fulfilment of the requirements for the degree of

Doctor of Philosophy

in Cell and Molecular Biosciences

Victoria University of Wellington

2014

Abstract

Groundwater is a globally important natural resource and an integral part of the water supply in New Zealand. Due to high demand, the quality and availability of groundwater are both extensively monitored in New Zealand and globally, under State-of-the-Environment (SOE) monitoring programmes. SOE groundwater monitoring in New Zealand mainly evaluates hydrochemistry and until this thesis has largely overlooked the biotic component. Microbes including bacteria play a crucial role in ecosystem functioning by mediating biogeochemical processes in subsurface environments. Therefore, analysis of microbiological content will enable better evaluation of the health of groundwater ecosystems that is not fully reflected by chemical data alone.

This project characterizes the bacterial diversity in New Zealand groundwater at national and regional scales using molecular methods and explores the underlying factors that shape the bacterial community structure. A simple molecular profiling tool, Terminal Restriction Fragment Length Polymorphism (T-RFLP) was used to determine community structure at local and national scales. The results revealed considerable diversity that was driven by groundwater chemistry. Roche 454-pyrosequencing was then used to obtain a deeper insight into New Zealand groundwater ecosystems, and showed that bacterial communities have many low abundance taxa and relatively few highly abundant species. In addition, microbial diversity is mainly related to the redox potential of the groundwater. But, despite this relationship, *Pseudomonas spp.* were the dominant genus at many sites even those with diverse chemistries and environmental factors. The final phase of the project set the platform to test whether these *Pseudomonas spp.* have acquired genetic material from other species via horizontal gene transfer

(HGT) enabling them to adapt into a diverse range of habitats. A whole-genome sequencing approach (Illumina MiSeq platform) was used to develop six metagenomic databases as a resource to test this hypothesis. Initial results show some evidence for HGT and further investigations are underway.

Overall, the knowledge generated across all phases of this project provides novel insights into New Zealand groundwater ecosystems and creates a scientific basis for the future inclusion of microbial status assessment criteria into regional and national groundwater monitoring programmes and related policies in New Zealand.

Acknowledgements

I would like to extend my sincere thanks to my supervisors: Dr Geoffrey Chambers, Dr Ken Ryan and Dr Chris Daughney. I consider myself very lucky to have such a wonderful team of supervisors. They have helped me to develop a very good research project and this task would not be possible without their support and motivation.

My PhD research was funded by GNS Science, New Zealand. I would like to take this opportunity to thank them for their valuable support. I would also like to acknowledge the special support given by Dr Magali Moreau (GNS Science) and Ms Sheree Tidswell (Greater Wellington Regional Council) for coordinating the groundwater sample collection and retrieving the hydrochemical data. Special thanks go to all the groundwater research staff at the 15 regional councils in New Zealand for their valuable support in sample collection.

I would like to thank Victoria University of Wellington for providing me scholarship support throughout my PhD studies and the School of Biological Sciences for providing me all the facilities that were required for my project. My thanks go to Patricia Stein, Sandra Taylor, Paul Marsden, Lesley Thompson and Mary Murray for their support and concern in administrative matters and to Cameron Jack, Craig Doney, Angela Fleming and Chris Thorn for their assistance in laboratory work.

I would like to thank Professor Craig Cary, Dr Charles Lee, Dr Craig Herbold and Sarah Kelly at the University of Waikato for introducing me to the next-generation sequencing world. I would also like to thank Dr Patrick Biggs and Ms Lorraine Berry at the Massey Genome Service for their valuable support in genomics work. My gratitude goes to Dr Dalice Sim and Dr David Eccles for helping me with data analyses.

Special thanks go to my colleagues: Leighton Thomas, Maheshini Mawalagedera, Edinur Atan, Phil Sirvid, Eileen Koh, Luke Thomas, Marie Fernandez, Aashish Morani, Arun Kanakkanthara, Peter Bosch, Cong Zeng, Heather Constable, Catarina Silva, Gandeep Jain, Sharada Paudel, Shamsar Lamichhane for motivating me to succeed in this task.

I acknowledge with thanks the moral support given by my teachers in Sri Lanka: Professor S. P. Samarakoon, Professor Athula Perera, Professor V. A. Sumanasinghe and Professor Deepthi Bandara.

Finally, I would like to thank my parents: Parakrama Sirisena and Leela Sirisena, my loving sister: Achala, my brothers: Suneth and Manjula, my wife: Chathuri, my little nephew: Sehas and brother-in-law: Srinath for their unconditional love and support. All I have achieved is yours.

List of related publications and presentations

A. Peer reviewed international Journals

1. **Sirisena KA**, Daughney CJ, Moreau-Fournier M, Ryan KG, Chambers GK 2013. National survey of molecular bacterial diversity of New Zealand groundwater: relationships between biodiversity, groundwater chemistry and aquifer characteristics. *FEMS Microbiology Ecology* 86: 490-504.
2. **Sirisena KA**, Daughney CJ, Moreau M, Ryan KG, Chambers GK 2014. Relationships between Molecular Bacterial Diversity and Chemistry of Groundwater in the Wairarapa Valley, New Zealand. *New Zealand Journal of Marine and Freshwater Research* *In press*.
3. **Sirisena KA**, Daughney CJ, Moreau M, Sim DA, Ryan KG, Chambers GK 2014. Pyrosequencing analysis of bacterial diversity in New Zealand groundwater. (*to be submitted to Molecular Ecology by June 2014*)

B. International conference presentation

1. **Sirisena KA**, Daughney CJ, Herbold CW, Lee CK, Ryan KG & Chambers GK 2012. Molecular Microbial Diversity in New Zealand Groundwater. 14th International Symposium on Microbial Ecology: 19 – 24 August 2012: Copenhagen, Denmark. *Poster presentation*.

C. Oral and poster presentations

1. **Sirisena KA** 2010. Molecular Characterization of Bacterial Diversity in New Zealand Groundwater. Ecology and Evolution Seminar Series 2010, New Zealand, 17 September 2010: Victoria University of Wellington, New Zealand. *Oral presentation.*
2. **Sirisena KA**, Chambers GK, Daughney CJ 2010. Molecular Characterization of Bacterial Diversity in New Zealand Groundwater. Conference of the New Zealand Microbiological Society and New Zealand Society for Biochemistry & Molecular Biology: 30 November – 03 December 2010: University of Auckland, New Zealand. *Poster presentation.*
3. **Sirisena KA**, Daughney CJ, Chambers GK 2010. Molecular Characterization of Bacterial Diversity in New Zealand Groundwater. Victoria University PGSA Conference 2010: 14 November 2010: Victoria University of Wellington, New Zealand. *Poster presentation.*
4. **Sirisena KA**, Moreau-Fournier M, Daughney CJ, Ryan KG, Chambers GK 2011. National Survey of Microbial Biodiversity in New Zealand Groundwater. 50th New Zealand Hydrological Society Conference: 5 – 9 December 2011; Te Papa, Wellington, New Zealand. *Oral presentation.*
5. **Sirisena KA**, Moreau-Fournier M, Daughney CJ, Ryan KG & Chambers GK 2012. Molecular Microbial Diversity of New Zealand Groundwater: Terminal Restriction Fragment Length Polymorphism (T-RFLP) Approach. 2nd Conference of the New Zealand Microbial

Ecology Consortium: 16 – 17 February 2012: University of Auckland, New Zealand. *Poster presentation.*

Table of Contents

Abstract.....	I
Acknowledgements.....	III
List of related publications and presentations.....	V
Table of Contents.....	VIII
List of Tables.....	XII
List of Figures.....	XV
Declaration.....	XX
Chapter 1: General Introduction.....	1
1.1 Groundwater in New Zealand.....	3
1.2 Groundwater monitoring.....	3
1.3 Microbial assessments in groundwater ecosystems.....	5
1.4 Molecular and culturing techniques in microbial ecology.....	7
1.4.1 Molecular profiling techniques.....	8
1.4.1.1 Terminal restriction Fragment Length Polymorphism (T-RFLP).....	8
1.4.1.2 Automated Ribosomal Intergenic Spacer Analysis (ARISA).....	10
1.4.1.3 Denaturing Gradient Gel Electrophoresis (DGGE).....	11
1.4.2 Metagenomics.....	11
1.4.2.1 Metagenomics with clone library construction and Sanger sequencing	12
1.4.2.2 Metagenomics using next-generation sequencing technologies.....	14
1.4.2.2.1 Roche 454 sequencing technology.....	14
1.4.2.2.2 Illumina sequencing technology.....	16
1.4.2.2.3 Ion Torrent sequencing technology.....	17

1.5 Broad objectives of the project.....	17
1.5.1 National scale assessment of groundwater bacterial diversity.....	18
1.5.2 Local scale assessment of groundwater bacterial diversity.....	18
1.5.3 Relationships between bacterial diversity and hydrochemistry.....	19
1.5.4 Horizontal gene transfer and bacterial diversity.....	19
1.6 Formal statement of main hypotheses.....	20
Reference.....	21
Chapter 2: Extended Information on Materials and Methods.....	27
2.1 Groundwater Sampling Strategy.....	27
2.2 General Laboratory Practices.....	28
2.3 DNA Quantification.....	28
2.4 Control Experiments.....	28
2.5 Visualization of PCR and Restriction Digestion Products.....	29
2.6 Direct DNA sequencing.....	29
2.7 Terminal Restriction Fragment Length Polymorphism (T-RFLP).....	30
2.8 Roche 454 Pyrosequencing.....	31
2.9 Illumina MiSeq™ High Throughput Sequencing.....	33
2.10 Data Analysis.....	33
2.10.1 Terminal Restriction Fragment Length Polymorphism (T-RFLP).....	33
2.10.2 Roche 454 Pyrosequencing Data Analysis.....	33
2.10.3 Illumina MiSeq™ High-Throughput Sequencing.....	34
References.....	35

Chapter 3.1: National Survey of Molecular Bacterial Diversity of New Zealand Groundwater: Relationships between Biodiversity, Groundwater Chemistry and Aquifer Characteristics.....	37
Abstract.....	38
Introduction.....	39
Materials and methods.....	42
Results and discussion.....	49
Concluding remarks.....	67
Acknowledgments.....	69
References.....	71
Supplementary information.....	78
Chapter 3.2: Relationships between molecular bacterial diversity and chemistry of groundwater in the Wairarapa Valley, New Zealand.....	97
Abstract.....	98
Introduction.....	99
Materials and methods.....	102
Results.....	107
Discussion.....	123
Acknowledgments.....	129
References.....	130
Supplementary information.....	137
Chapter 3.3: Pyrosequencing analysis of bacterial diversity in New Zealand groundwater....	149
Abstract.....	150
Introduction.....	151

Materials and methods.....	154
Results.....	161
Discussion.....	176
Acknowledgments.....	180
References.....	181
Supplementary information.....	186
Chapter 3.4: Preliminary overview of horizontal gene transfer in groundwater bacteria.....	201
Abstract.....	202
Introduction.....	203
Materials and methods.....	205
Results.....	214
Discussion.....	223
References.....	226
Chapter 4: General Discussion and Conclusion.....	231
References.....	239

List of Tables

Chapter 3.1

Table 1: Number of samples belonging to each classification level based on number of FAM and HEX peaks identified in T-RFLP electropherograms.....	58
Table 2: Number of samples belonging to each Biocluster at different clustering levels.....	58
Table 3: Typical chemical characteristics for Hydrochemical categories and subcategories defined by Daughney and Reeves (2005).....	59
Table 4: Summary of Shannon-Wiener diversity indices (H') in each Biocluster at 11-cluster threshold calculated using FAM and HEX T-RFs separately.....	60
Table 5: Summary of groundwater features in Bioclusters at 11-cluster threshold.....	61
Table S1: Summary of genomic DNA yields obtained from two litres of groundwater from each sample.....	78
Table S2: Median values of 15 chemical parameters and 4 physical parameters derived from the actual values measured quarterly from March 2008 to March 2012.....	79
Table S3: Summary of P values (95.0% confidence level) of Kruskal-Wallis test for each parameter at different threshold levels.....	83
Table S4: Summary of P values (95.0% confidence level) of Chi- square test for each parameter at different threshold levels.....	84
Table S5: Characteristics of groundwater sampling sites.....	85

Chapter 3.2

Table 1: Summary of the relative magnitudes of chemical parameters in each Biocluster at 3-cluster threshold.....	111
---	-----

Table 2: Summary of mean Shannon diversity indices (H') and standard deviations (SD) for each Biocluster, separately calculated for FAM and HEX T-RFs.....	112
Table 3: Summary of groundwater characteristics in each Biocluster.....	113
Table S1: Concentration values of 30 hydrochemical parameters at each groundwater monitoring site in the September 2009 sampling round.....	137
Table S2: Summary of geographical location (in Northing and Easting), aquifer confinement and usage of groundwater of the GWRC sampling sites.....	140
Table S3: Summary of P values (95.0% confidence level, $n=35$, d. f. = 34) of Kruskal-Wallis tests for each chemical parameter at the 3- and 2-Cluster thresholds.....	141
Table S4: Summary of groundwater chemistry at GWRC sampling sites included in both van Bekkum et al. (2006) study and present study.....	142

Chapter 3.3

Table 1: Summary of bacterial diversity and richness estimates based on 454-pyrosequencing operational taxonomic units (OTUs) defined at 0.03 cut-off level.....	166
Table S1: Median values of 19 hydrochemical parameters derived from the actual values measured quarterly from March 2008 to March 2012 in groundwater monitoring sites.....	186
Table S2: Summary of site-specific information: aquifer lithology, confinement, well depth (depth code), groundwater mean residence time (MRT class), land use activities in the aquifer recharge zone, geographical region and hydrochemical category to which each site belongs...	188
Table S3: Summary of richness and abundance of unique OTUs in each sample.....	190
Table S4: Summary of richness and abundance of shared OTUs in each sample.....	191
Table S5: Shannon diversity indices and number of OTUs based on 454 pyrosequencing data and T-RFLP data presented in Sirisena <i>et al</i> (2013).....	192

Table S6: Summary of the contribution of each bacterial species for the similarity within each hydrochemical category.....	193
--	-----

Chapter 3.4

Table 1: General characteristics of the hydrochemical categories at the three thresholds.....	206
Table 2: Summary of the site specific information of groundwater sampling sites.....	208
Table 3: Median hydrochemical values of 19 parameters derived from the actual values measured quarterly from March 2008 to March 2012.....	209
Table 4: Summary of the reference <i>Pseudomonas</i> genomes used in short read mapping.....	212
Table 5: Summary of the number of reads resulted from each metagenome.....	214
Table 6: Summary of the short read mapping into reference <i>Pseudomonas</i> genomes.....	217
Table 7: Summary of the maximum genome sizes for each sample.....	219
Table 8: Summary of the Glimmer gene predictions.....	220
Table 9: Summary of the BLAST hits for each genome.....	220

List of Figures

Chapter 1

Figure 1: Distribution of global water resources.....	1
Figure 2: Schematic cross-section of confined and unconfined aquifer system.....	2
Figure 3: Schematic overview of the Terminal Restriction Fragment Length Polymorphism (T-RFLP) technique.....	9
Figure 4: Schematic overview of the Automated Ribosomal Intergenic Spacer Analysis (ARISA) technique.....	10
Figure 5: Schematic overview of the Sanger-sequencing metagenomics approach based on clone library preparation and sequencing.....	13
Figure 6: Schematic overview of the Roche 454 sequencing technology.....	15
Figure 7: Overview of the Illumina sequencing workflow.....	16

Chapter 3.1

Figure 1: Groundwater sampling sites across New Zealand.....	62
Figure 2: Dendrogram produced by hierarchical cluster analysis conducted using FAM and HEX labelled terminal fragments.....	63
Figure 3: Box-and-Whisker Plot of median concentrations of NO ₃ -N (a), NH ₄ -N (b), Fe (c) and Mn (d) across Bioclusters defined at the 11-cluster threshold.....	64
Figure 4: Percentage frequency distribution of samples with hydrochemical categories a), MRT Classes (b), aquifer well depth (c), aquifer lithology (d), land use activities of aquifer recharge zone (e) and regional council (f).....	65
Figure 5: Summary of mean Shannon-Wiener diversity indices (H') values for each Biocluster using FAM and HEX T-RFs.....	66

Figure S1: Summary of genomic DNA yields obtained from two litres of groundwater from each sample.....	86
Figure S2: Summary of the number of samples detected with each (a) FAM and (b) HEX Operational Taxonomic Unit (OUT).....	87
Figure S3: Examples of T-RFLP profiles categorized as (a) simple, (b) moderately complex or (c) complex based on number of FAM or HEX peaks.....	88
Figure S4: Summary of the total number of FAM (a) and HEX (b) peaks over 200 RFU in each sample.....	89
Figure S5: The Box-and-Whisker Plot of median HCO_3 (a), Ca (b), Fe (c) and Mn (d) across Bioclusters defined at the 4-cluster threshold.....	90
Figure S6: The Box-and-Whisker Plot of median $\text{NO}_3\text{-N}$ (a), F (b) and Dissolved Oxygen (c) across Bioclusters defined at the 5-cluster threshold.....	91
Figure S7: The Box-and-Whisker Plot of median SiO_2 (a) and Mg (b) across Bioclusters defined at the 7-cluster threshold.....	92
Figure S8 (i): Box-and-Whisker Plot of median concentrations F (a), $\text{PO}_4\text{-P}$ (b), Dissolved Oxygen (c) and Br (d) across Bioclusters defined at the 11- cluster threshold.....	93
Figure S8 (ii): Box-and-Whisker Plot of median concentrations of SO_4 (a), HCO_3 (b), SiO_2 (c) and Mg (d) across Bioclusters defined at the 11-cluster threshold.....	94
Figure S8 (iii): Box-and-Whisker Plot of median concentrations of Na (a), K (b), Cl (c) and Ca (d) across Bioclusters defined at the 11-cluster threshold.....	95
Figure S8 (iv): Box-and-Whisker Plot of median concentrations of Electrical conductivity (a), Water temperature (b) and Acidity (h) across Bioclusters defined at the 11-cluster threshold....	96

Chapter 3.2

Figure 1: Groundwater sites sampled in the Wairarapa valley and the Riversdale area, New Zealand.....	114
Figure 2: Summary of the total number of FAM (Black) and HEX (Grey) T-RFs over 200 RFU in each sample.....	115
Figure 3: Summary of the frequency of each (A) FAM and (B) HEX T-RF.....	116
Figure 4: Summary of Shannon diversity index (H') values for each sample using FAM (Black) and HEX (Grey) OTUs.....	117
Figure 5: Dendrogram produced by hierarchical cluster analysis performed using Ward's linkage rule with FAM and HEX T-RFs standardized to the sum of all peaks in each profile and the Bray-Curtis similarity index.....	118
Figure 6: Box-and-Whisker Plot comparisons of concentrations of (A) Fe, (B) Mn, (C) $\text{NH}_4\text{-N}$, (D) $\text{NO}_2\text{-N}$, (E) $\text{NO}_3\text{-N}$ and (F) Dissolved Oxygen across bioclusters defined at the 3-cluster threshold.....	119
Figure 7: Percentage of samples in each biocluster defined at 3-cluster threshold as a function of (A) aquifer confinement and (B) groundwater bore usage.....	120
Figure 8: Summary of mean Shannon diversity index (H') values for each biocluster using FAM and HEX T-RFs.....	121
Figure 9: Hierarchical cluster analysis pattern for the five samples: Seymour; Trout Hatchery; Johnson; CDC South; and George revealed by van Bekkum et al. (2006).....	122
Figure S1: Dendrograms of the hierarchical cluster analyses performed with different combinations of of peak scaling methods and distance measures.....	144

Figure S2 (i): Box-and-Whisker Plot comparisons of concentrations of SO ₄ , Total Dissolved Solids, Total Oxidized Nitrogen, Na, K and Mg across bioclusters defined at the 3-cluster threshold.....	145
Figure S2 (ii): Box-and-Whisker Plot comparisons of concentrations of Ca, B, HCO ₃ , Cl, Br and F across bioclusters defined at the 3-cluster threshold.....	146
Figure S2 (iii): Box-and-Whisker Plot comparisons of concentrations of PO ₄ -P, SiO ₂ , Alkalinity, Total hardness, Total cations and Total anions across bioclusters defined at the 3-cluster threshold.....	147
Figure S2 (iv): Box-and-Whisker Plot comparisons of Electrical conductivity, Acidity, Free CO ₂ , Total organic carbon, concentrations of Pb and Zn across bioclusters defined at the 3-cluster threshold.....	148

Chapter 3.3

Figure 1: Groundwater sampling sites across New Zealand.....	168
Figure 2: Non-metric multidimensional scaling based on the relative abundances of all OTU..	169
Figure 3: Canonical correspondence analysis of the relative abundance of all OTUs with the 19 hydrochemical parameters.....	170
Figure 4A: Groundwater bacterial taxonomic diversity at phylum level.....	171
Figure 4B: Groundwater bacterial taxonomic diversity at class level.....	172
Figure 4C: Groundwater bacterial taxonomic diversity at order level.....	173
Figure 4D: Groundwater bacterial taxonomic diversity at family level.....	174
Figure 5: Groundwater bacterial taxonomic diversity at genus level.....	175
Figure S1 A: Rarefaction curves for oxidized groundwater samples with high human impact..	194
Figure S1 B: Rarefaction curves for oxidized groundwater samples with low human impact...	195

Figure S1 C: Rarefaction curves for moderately reduced groundwater samples.....	196
Figure S1 D: Rarefaction curves for highly reduced groundwater samples.....	197
Figure S2: Non-metric multidimensional scaling based on the relative abundances of: (a) all OTUs; (b) all OTUs except singletons; and (c) the 100 most abundant OTUs.....	198
Figure S3: Percentage dissimilarity between each pair of hydrochemical categories.....	199
Chapter 3.4	
Figure 1: Groundwater sampling sites across New Zealand.....	207
Figure 2: The T-RFLP profiles for the six selected samples. The experimental details are explained in chapter 3.1.....	215
Figure 3: Bacterial community compositions of three samples obtained by Roche 454 pyrosequencing approach.....	215
Figure 4: Word-cloud summary of taxonomic identities in the six metagenomes using the PAUDA: A) based on absolute read counts mapped; and B) based on normalized read counts mapped on the square root scale.....	218
Figure 5: Illustration of the contig coverage for contig 1026 from the sample GGW ID 11.....	221
Figure 6A: The BLAST results for the selected section of the contig 1026 from the sample GGW ID 11.....	222
Figure 6B: The BLAST results for the selected section of the contig 1034 from the sample GGW ID 11.....	222

Declaration

This PhD thesis includes a General Introduction, extended information on Materials and Methods, Results chapters and General Discussion and Conclusion. Three results chapters are in the form of research articles/manuscripts. Therefore, the formats of these chapters are in accord with the style of the journal in which they are published/submitted. These chapters involve contribution from the co-authors and all of them are my PhD supervisors and collaborators. The below section outlines the author contribution to each research article/manuscript included in this thesis.

Chapter 3.1:

KAS designed and conduct the research and prepared the manuscript; CJD helped to design the research, provided funding and contributed to the preparation of paper; MMF helped in sample collection, managed the NGMP database; KGR helped to design the research and contributed to the preparation of the paper; GKC helped to design the research and contributed to the preparation of the paper.

Chapter 3.2:

KAS designed and conduct the research and prepared the manuscript; CJD helped to design the research, provided funding and contributed to the preparation of paper; MM helped in sample collection and creating sampling site maps; KGR helped to design the research and contributed to the preparation of paper; GKC helped to design the research and contributed to the preparation of paper.

Chapter 3.3:

KAS designed and conduct the research and prepared the manuscript; CJD helped to design the research, provided funding and contributed to the preparation of paper; MM helped in sample collection and creating sampling site maps; DAS advised on statistics; KGR helped to design the research and contributed to the preparation of paper; GKC helped to design the research and contributed to the preparation of paper.

I confirm that I am the first author with the greatest contribution for these articles/manuscripts included in this thesis.

GENERAL INTRODUCTION

Life on earth and water are inseparable, as life without water is impossible. Water is found almost everywhere: 1) above the earth's surface as atmospheric vapours and clouds; 2) on the surface as oceans, rivers, lakes, glaciers and inside animals and plants; and 3) below the surface as groundwater. However, the majority (97.5%) of water present on our planet is saline water that cannot be directly used for human needs (Shiklomanov 2000). The remaining 2.5%, the freshwater, is crucial for mankind, but a large fraction of it (68.7%) is trapped as glaciers and ice caps (Fig. 1) and thus, unavailable for our usage (Carpenter et al. 2011). Groundwater is the largest portion of the remainder and accounts for nearly 99% of the total volume of liquid freshwater presently circulating on our planet (Younger 2007). Therefore, groundwater plays an essential role in the survival of human beings on earth. It is the world's major drinking water source, providing about 60% of drinking water in Europe and more than 80% in North Africa and the Middle East (Struckmeier et al. 2005; Steube et al. 2009). In addition, groundwater is a major water source for irrigation and industrial purposes in many parts of the world (Siebert et al. 2010). Therefore, research on groundwater is very important for the sustainable management of this valuable natural resource.

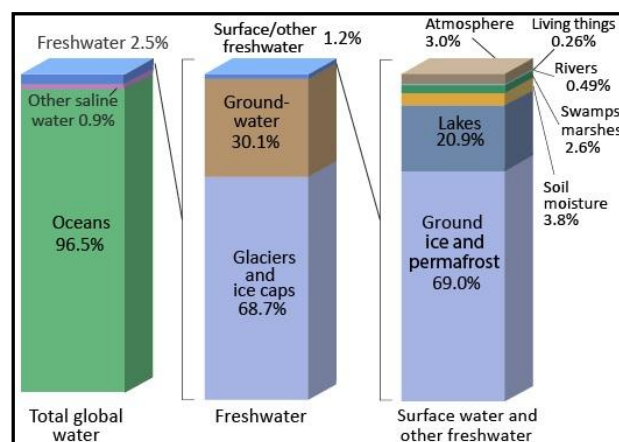


Figure 1 Distribution of global water resources.

Source: US Geological Survey (<http://ga.water.usgs.gov/edu/earthwherewater.html>)

Groundwater mainly originates from rainfall that slowly infiltrates through the soil particles and is trapped in the pores in soil and rocks. However, the rate and the amount of this infiltration are largely influenced by soil porosity and permeability. Porosity reflects the ability to store the water in pores between the individual soil particles. Permeability refers to the ability to transmit water stored in pores between them and it is determined by the degree of connectivity of the pores in between the soil particles. These factors vary from one soil type to another. Porous and permeable soils are ideal places to accumulate groundwater. As the infiltration process continues, the bottom soil layers become fully saturated with water while the upper layers remain unsaturated. The interface between the saturated and unsaturated zones is called the water table. The water present in the unsaturated zone is referred to as soil moisture, whereas that below the water table is groundwater. The saturated zone materials that transmit and store groundwater are called aquifers. There are two types of aquifers: 1) unconfined aquifers that occur when the upper limit is the water table and the lower margin is a low-permeability rock (confining unit); and 2) confined aquifers which have low-permeability rock on both upper and lower boundaries (Fig. 2).

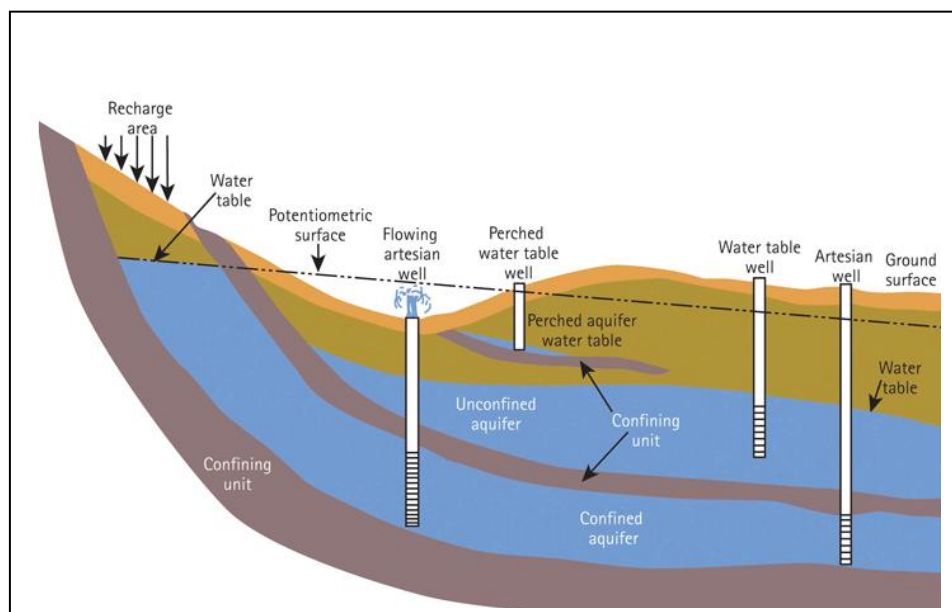


Figure 2 Schematic cross-section of confined and unconfined aquifer system. This figure was modified after National Groundwater Association (2010)

Because the basic accumulation process of groundwater only involves infiltration of rainwater through soil particles and storage in aquifers, this valuable natural resource can usually be found in any part of the globe. Therefore, the majority of the world's population has direct access to this resource and there is a large interest in gaining a full understanding of this resource. Despite the fact that there is a growing interest in this research area, still there are vast knowledge gaps especially on the biological activities in groundwater ecosystems (Griebler & Lueders 2009).

1.7 Groundwater in New Zealand

Groundwater is an important part of the national water supply in New Zealand. Nearly one quarter of the New Zealand population uses groundwater as its major drinking water source (Daughney & Reeves 2005). For example, approximately half of the Waikato region's rural population relies on groundwater for drinking (Waikato Regional Council 2014). Further, some cities such as Napier, Hastings, Wanganui, Lower Hutt and Christchurch are totally dependent on groundwater for all their water requirements. In addition to drinking purposes, a significant fraction of the water requirements for the agricultural and industrial sectors are also fulfilled by groundwater. Overall, nearly 34% of the total water use in New Zealand excluding hydro power generation is supplied from groundwater (Daughney & Reeves 2005; Rajanayaka et al. 2010).

1.8 Groundwater monitoring

As groundwater is such a valuable resource, its quality and availability are extensively monitored, both in New Zealand and globally under State-of-the-Environment (SOE) monitoring programmes. SOE monitoring is usually conducted at a regional or national scale and is also referred to as baseline, background, ambient or long-term monitoring. A typical

SOE monitoring scheme that includes groundwater assessment aims to: 1) characterize groundwater quality in terms of its current state and trends; 2) relate the observed state and trends to specific causes such as land use, pollution or natural processes; and 3) provide data to assess the effectiveness of groundwater management policies. SOE monitoring usually involves regular collection of groundwater samples from a fixed network of sites followed by analyses of these samples for a suite of physical and chemical parameters.

In New Zealand, SOE groundwater quality monitoring is mainly undertaken by the 15 regional authorities that evaluate the state of the groundwater chemistry within their own areas of jurisdiction. This provides an efficient framework to obtain useful information on regional groundwater quality. In addition to regional monitoring, the National Groundwater Monitoring Programme (NGMP) also plays an important role in SOE groundwater monitoring on a national scale (Daughney et al. 2012). The NGMP provides a useful network of sites across the country and was originally established in 1990 by the Groundwater Group of the Department of Scientific and Industrial Research, New Zealand. In the initial phase, only two regional councils (Tasman and Bay of Plenty) were involved with the NGMP. The other regional authorities collaborated with the network with the gradual development of the overall programme: Hawke's Bay and Taranaki joined in 1992; Waikato and Manawatu-Wanganui in 1994; Canterbury and Wellington in 1995; Otago, Northland, Gisborne and Auckland in 1996; and finally West Coast, Marlborough and Southland in 1998. Presently, the NGMP is run by GNS Science in collaboration with the above listed 15 regional authorities and includes 110 active monitoring sites throughout the country (Rosen 2001; Daughney & Reeves 2005, 2006; Morgenstern & Daughney 2012). These sites are located in discrete aquifers representing an array of environmental and geological factors and provide a highly representative picture of groundwater quality across New Zealand (Daughney et al. 2012). The NGMP conducts quarterly analyses (in March,

June, September and December) of groundwater quality in terms of the groundwater chemistry. The concentrations of major chemical constituents such as Na, K, Mg, Ca, HCO_3 , Cl, SO_4 , $\text{NO}_3\text{-N}$, $\text{NH}_4\text{-N}$, $\text{PO}_4\text{-P}$, Fe, Mn, Br, F and SiO_2 and site-specific data such as dissolved oxygen, electrical conductivity, pH and water temperature are measured. These hydrochemical data are stored in the GNS Science Geothermal and Groundwater (GGW) Database (<http://ggw.gns.cri.nz/ggwdata/mainPage.jsp>) and are readily available to interested parties, providing a useful framework for groundwater studies on a national scale. Therefore, the New Zealand groundwater monitoring activities in terms of the state of hydrochemistry is both efficient and actively growing.

1.9 Microbial assessments in groundwater ecosystems

Historically, groundwater studies have been conducted mainly to investigate the hydrological aspects of the resource without attempts to evaluate groundwater biology (Humphreys 2009). In these cases, groundwater monitoring is simply referred to as hydrochemical analysis. However, with recent advances, groundwater is now considered not only as a valuable resource for human use, but also as a dynamic ecosystem. Therefore, in some parts of Europe and Australia, criteria for assessments of ecological status have already been included in their national groundwater monitoring policies (Griebler et al. 2010; Stein et al. 2010; Korbel & Hose 2011). Microbiologists have taken the lead role in this transition thanks to rapidly developing modern techniques (Humphreys 2009). Microorganisms are the key driving force for biogeochemical processes taking place in the groundwater ecosystem as in many other subsurface ecosystems (Falkowski 2008). Groundwater microbial communities are selected and regulated by the chemical and physical nature of groundwater and conversely, they mediate redox reactions, thus controlling the dissolved concentrations of elements such as Fe, Mn, N, S and many others (Ghiorse 1997; Chapelle 2000; Bethke et al.

2008; Hedrich et al. 2011). It is further expected that any change in the chemical composition of groundwater or aquifer sediment will cause a corresponding shift in the subsurface microbial community structure (Haack et al. 2004). Therefore, the most crucial ecological aspect of the groundwater studies could be to understand the microbial component, and this will enable us to postulate trends in ecosystems that are not visible with hydrochemical data alone (Griebler & Lueders 2009; Larned 2012).

To date, in most parts of the world including New Zealand, SOE monitoring has almost completely overlooked the microbiological component of groundwater systems. Although the NGMP has developed over two decades, the importance of including criteria for the assessment of microbial state of the groundwater is yet to be fully recognized and adapted. The national and regional SOE programmes typically only assess the presence of coliform bacteria (mainly *Escherichia coli*) in groundwater as a biological factor, because it is an indicator species of faecal contamination that could cause serious human health problems (Ministry for the Environment 2010; Greater Wellington Regional Council 2013). However, during recent years, an increasing number of studies have been conducted to assess bacterial parameters in groundwater including bacterial diversity and its relationships with biogeographical and hydrochemical conditions across varying spatial and temporal scales (Griebler et al. 2010; Stein et al. 2010; Sinreich et al. 2011; Zhou et al. 2012; Korbel et al. 2013). This type of studies can also help to increase our understanding of biogeochemical processes related to human health, i.e. the redox cycling of toxic metals like arsenic, mercury, and uranium. In New Zealand, a preliminary evaluation of microbial biodiversity of groundwater was conducted by van Bekkum et al. (2006). In this pilot study, bacterial community structure was determined using 20 groundwater samples collected from bores around the Hutt Valley and Wairarapa regions. This work provided initial indications of relationships between bacterial community structure and groundwater chemistry. However, it

is expected that the recent advances in microbiological techniques will help to expand our understanding of the groundwater microbial communities in New Zealand and globally.

1.10 Molecular and culturing techniques in microbial ecology

There are number of techniques available to study the bacterial diversity in subsurface environments including groundwater. Microscopic examination is the oldest approach for bacteria (Maier et al. 2009). However, this method is time consuming and has largely been superseded by culturing techniques and newly developed DNA based methods (Kim & Byrne 2006).

Culturing techniques are widely used in bacteria analyses in subsurface environments (Zhou et al. 1997; Janssen et al. 2002; Neufeld & Mohn 2005; Lozupone & Knight 2007). However, these methods can also be very laborious. In addition, many of the bacterial species present in environmental samples cannot be easily cultured in artificial culture media (Janssen et al. 2002). This limitation could be due to inadequate knowledge of the culturing conditions or the length of time required for visible microbial growth. Further, one soil bacteria study has showed that the actual bacterial diversity in that particular soil was approximately 170 times higher than the diversity found in the bacterial cultures isolated from the same soil (Torsvik et al. 1996). Thus, culturing methods may not be the most effective way to evaluate the actual bacterial diversity in environmental samples, and more importantly, the actual potential for discovering new species from environmental samples in this way is low (Chen & Pachter 2005).

Culture independent molecular methods have become more prominent in exploring microbial diversity in environmental samples. With the recent advances in molecular techniques, an array of DNA-based approaches is now available to explore subsurface microbial diversity (Maier et al. 2009). However, among the many different molecular

techniques, the polymerase chain reaction (PCR) plays a central role in environmental sample analysis. PCR is used to amplify a target gene or region in the genome, resulting in a significant amount of a specific DNA product copied from a minute DNA sample collected from the environment. Although these molecular detection methods are becoming very popular, it is important to note that these methods alone may also not be able to identify all the bacterial species in subsurface environments (Donachie et al. 2007). This is because the most important requirement for analysing microbial composition in environmental samples with a molecular method is to extract all of the DNA from the sample. However, it is not possible to ensure this has happened as there could be some species that have thick cell walls and DNA cannot be easily extracted from such species. In addition, certain species might need specific PCR conditions of which investigators may only have limited knowledge.

1.10.1 Molecular profiling techniques

These techniques are usually simple molecular fingerprinting tools that reveal the microbial community structure in environmental samples. However, many of these approaches have so far failed to provide exact taxonomic information of the microbes present in the sample.

1.10.1.1 Terminal restriction Fragment Length Polymorphism (T-RFLP)

The T-RFLP technique was developed to compare the microbial community structure of environmental samples based on the sequence differences of the 16S rRNA gene, which codes for the small sub unit of bacterial ribosomal RNA (Liu et al. 1999). This gene is found in the genomes of all bacterial species and most of the archaeal species. Several regions of this gene are highly conserved among all bacteria, whereas some regions are conserved only among particular genera or species. Thus, universal primer sets can easily be designed and used to amplify a particular 16S rDNA target region lying between two such conserved sites.

In T-RFLP, part of the 16S rRNA gene is amplified using fluorescently labelled forward and/or reverse universal primers. It results in PCR products that are fluorescently labelled at one or both ends. Next, the PCR products are digested with a restriction endonuclease, resulting in fluorescently labelled restriction fragments. Then, the terminal restriction fragments (T-RFs) are subjected to automated capillary electrophoresis for size detection. The fluorescent peak profiles of T-RFs reveal the bacterial community structure (Fig. 3). Although the T-RFLP technique does not provide exact taxonomic information of the species present in samples it is a reliable, cost effective and simple technique that can be effectively used in basic environmental microbial analyses.

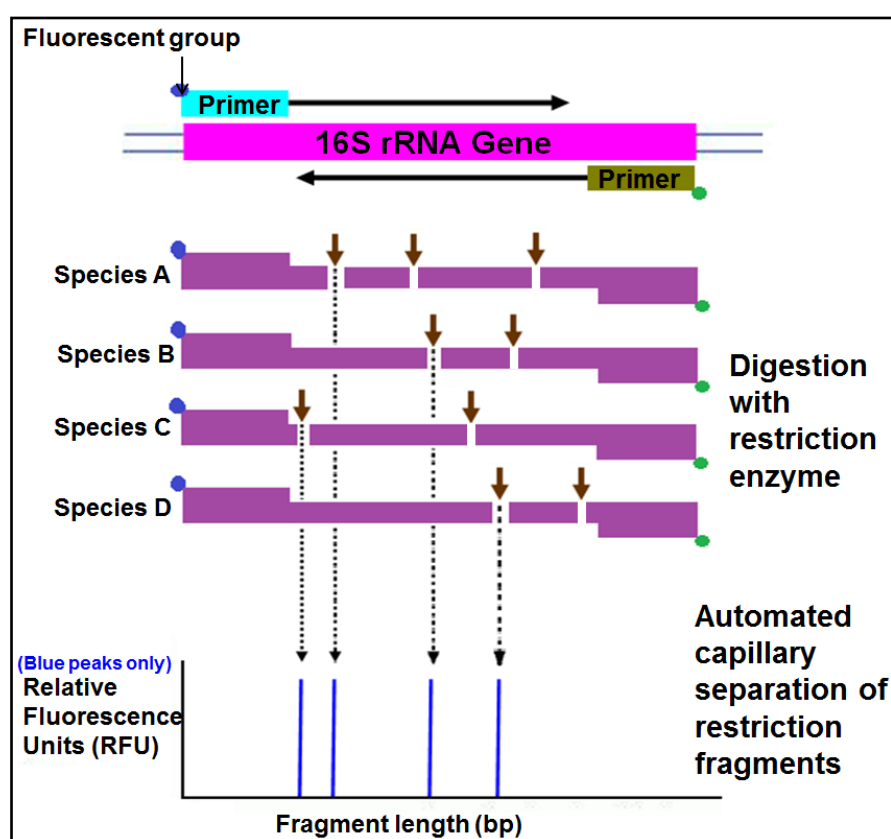


Figure 3 Schematic overview of the Terminal Restriction Fragment Length Polymorphism (T-RFLP) technique. The restriction sites for a particular restriction endonuclease are indicated with short brown arrows. Ideally, each blue line in the resulting electropherogram after automated capillary separation represents a particular taxon in the sample. Both primers can be labelled with different fluorescent labels (blue and green) to increase the resolution of the technique as the two terminal restriction fragments will provide two electropherograms for each fluorescent label.

1.10.1.2 Automated Ribosomal Intergenic Spacer Analysis (ARISA)

ARISA is an automated fingerprinting tool that targets the non-coding internal transcribed spacer (ITS) regions of the small and large subunit (SSU and LSU) rRNA gene (Ranjard et al. 2001). The highly variable nature of the ITS region and highly conserved nature of the flanking SSU/LSU genes provide the basis for this molecular tool. The detection of different taxa is based on the nucleotide sequence length of the ITS region amplified using two primers of which the forward primer is labelled with a fluorescent tag. As in T-RFLP, the amplified PCR products are subjected to automated capillary separation and the peaks in the resulting electropherogram correspond to the bacterial taxa present in the environmental sample (Fig. 4). However, one of the major drawbacks of this method is that it cannot differentiate between two different species that have similar nucleotide sequence lengths of ITS regions.

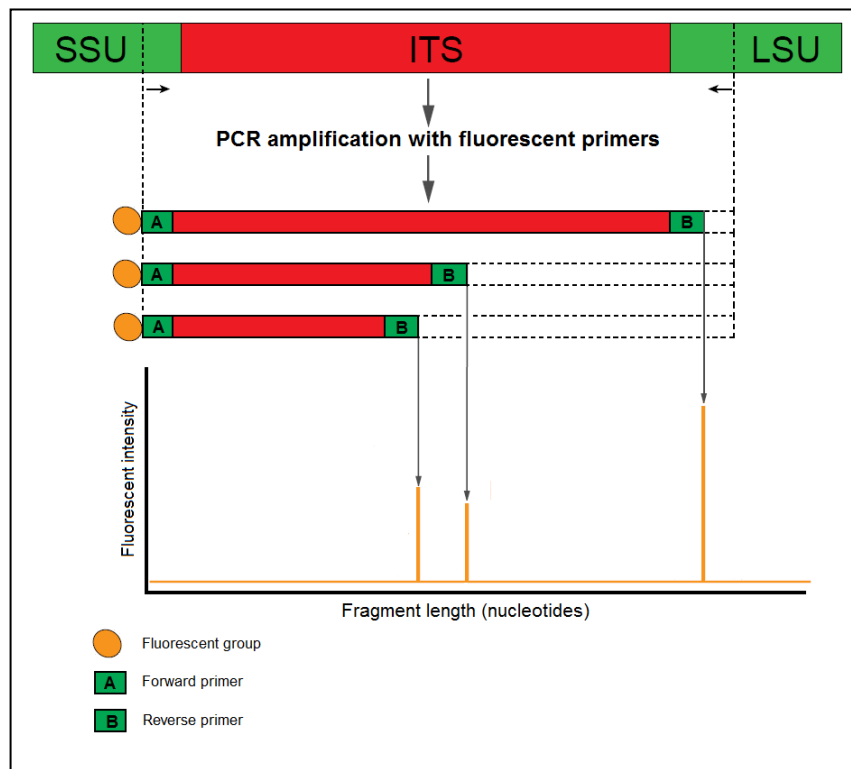


Figure 4 Schematic overview of the Automated Ribosomal Intergenic Spacer Analysis (ARISA) technique. This figure is modified after Wood et al. (2013).

1.10.1.3 Denaturing Gradient Gel Electrophoresis (DGGE)

DGGE is also another quite frequently used molecular profiling tool in microbial ecology that is based on the separation of multiple DNA sequences according to their mobility in increasingly denaturing conditions (Muyzer et al. 1993; Muyzer 1999). As in T-RFLP, the 16S rRNA gene is the most common target for DGGE. First, a variable region that is flanked by two conserved regions on the 16S rRNA gene is amplified by PCR using a universal primer set. Next, the PCR product is run on a polyacrylamide gel containing a linear concentration gradient of DNA denaturant such as urea or formamide. The mobility of the PCR product is dependent on the degree of denaturation of the double-stranded DNA molecule as fully dissociate PCR fragments stop moving along the gel. The degree of denaturation is related to the nucleotide sequence of the PCR product. Therefore, PCR bands that migrate to different positions on the gel can be identified as different taxa. One of the main drawbacks of this method is that it is hard to detect less abundant taxa in environmental samples.

1.10.2 Metagenomics

Metagenomics is a recently developed, powerful approach that provides a new way of examining the microbial world. In this methodology, the power of genomic analysis is applied to an entire microbial community, as opposed to classical microbiological approaches where the main focus was on single species in pure laboratory cultures. Therefore, metagenomics avoids the need to isolate and culture individual bacterial community members (Handelsman et al. 2007). Further, metagenomics is not limited to fingerprinting approaches, but it is capable of providing the taxonomic and functional composition of the sample, including detection of less abundant species. It thus has good potential to produce many exciting discoveries from environmental sources (Chen & Pachter 2005). To date, the

various different metagenomics approaches have rapidly evolved. However, in any metagenomics study, the first step is to directly extract DNA from all the microbes living in a particular environment. The mixed sample of DNA can then be analyzed directly, or cloned into vectors for subsequent genetic analyses.

1.10.2.1 Metagenomics with clone library construction and Sanger sequencing

In the early days of metagenomics, conventional Sanger DNA sequencing (Sanger et al. 1977) was used. Here, the first step was to construct a clone library from the amplified DNA sequences obtained from an environmental sample. These fragments were cloned into bacterial plasmids and transformed to host cells. The clones were then screened from the growth plates and subjected to Sanger sequencing that provides taxonomic information on the microbial community. In this sequence-based approach, clones are usually selected for sequencing based on the presence of phylogenetically informative genes, such as the 16S rRNA gene (Fig. 5).

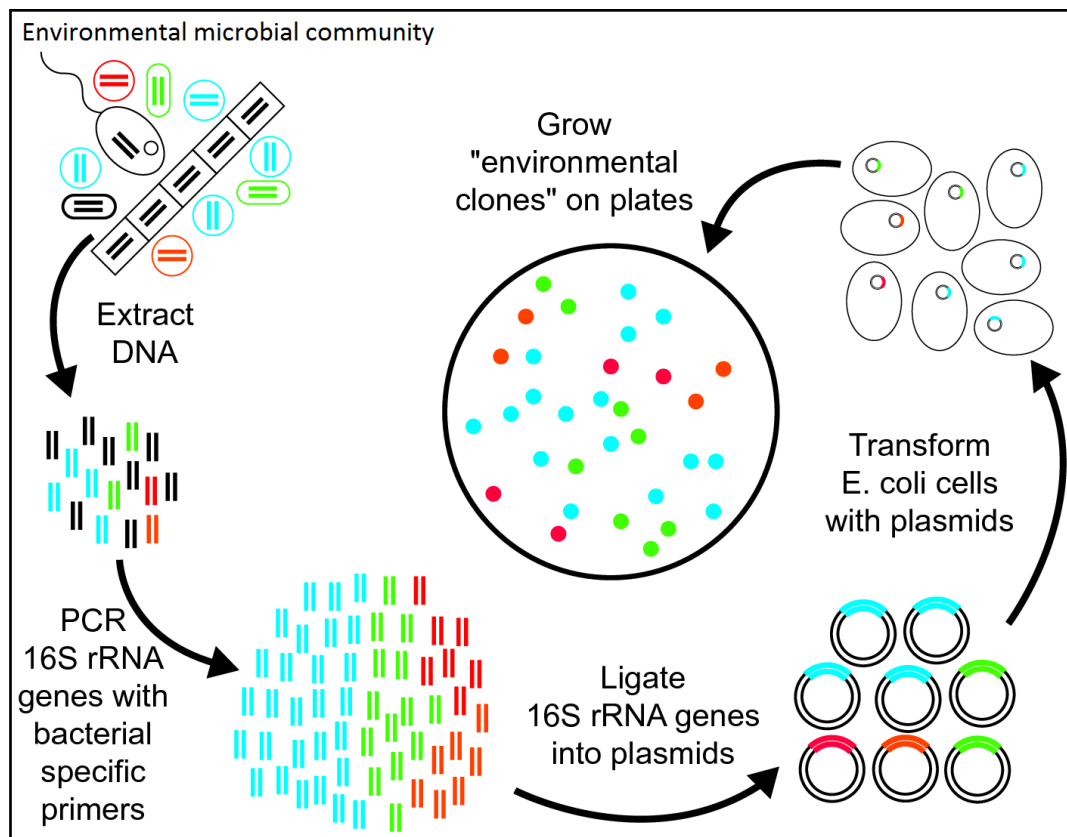


Figure 5 Schematic overview of the Sanger-sequencing metagenomics approach based on clone library preparation and sequencing.

In addition to sequencing, the DNA fragments that are cloned into vectors can be translated into proteins by the host bacteria under suitable laboratory conditions. These novel proteins can then be screened for various functions, such as vitamin production or antibiotic resistance. Therefore, clone library based metagenomic approaches can demonstrate the genetic diversity in the microbial community of environmental samples without having any prior knowledge on the DNA sequences or the origin of the microorganism. However, clone library construction is time consuming and recent advances in the development of DNA sequencing technologies are providing greater genetic analysis power.

1.10.2.2 Metagenomics using next-generation sequencing technologies

Next-generation sequencing (NGS) technologies have revolutionized methodological approaches in many scientific research areas including microbiology (Wood et al. 2013). NGS methods produce enormous number of DNA sequences relatively quickly and cheaply. This enables biologists to sequence even entire genomes of several microbial species present in different environments in a single experiment (Wrighton et al. 2012). In addition, NGS bypasses the requirement for clone library construction. Thus, NGS has become the central approach in modern environmental microbiological studies. To date, several NGS platforms have been commercialized (Glen 2011). Although different platforms employ unique chemistry and base incorporation/detection tools, all of them include library preparation (fragmentation or amplicon preparation), and detection of incorporated nucleotides (Wood et al. 2013). Presently, NGS platforms are referred to as 2nd generation sequencing technologies as another advance of sequencing techniques will soon emerge in the future as 3rd generation sequencing technologies that are capable of sequencing individual DNA/RNA molecules in real-time (Glen 2011; Wood et al. 2013). The three most commonly used 2nd generation NGS platforms are briefly discussed in the section below.

1.10.2.2.1 Roche 454 sequencing technology

The Roche 454 sequencing platform was first introduced in 2005 (Margulies et al. 2005). In this technique, nebulized DNA fragments or PCR amplicons are ligated into specific adaptor molecules and separated into single strands. These fragments are bound to micro-beads as one fragment per bead. Next, the immobilized DNA molecules are subjected to an emulsion-based PCR amplification that results in beads each carrying ten million copies of their original DNA templates. The beads are then loaded into a picotitre plate that has millions of wells, where each well accommodates only a single bead while serving as an individual

reactor vessel for enzymatic DNA sequencing (Fig. 6). Finally, all the beads are subjected to parallel sequencing by flowing pyrosequencing reagents across the picotitre plate. As each nucleotide is incorporated, the emission of a particular fluorescent signal is detected in each well using a charge-coupled device (CCD) camera (Rothberg & Leamon 2008). The 454 sequencing platform provides the longest sequence reads (i.e. 400-800 bp) compared to other NGS platforms (Wood et al. 2013).

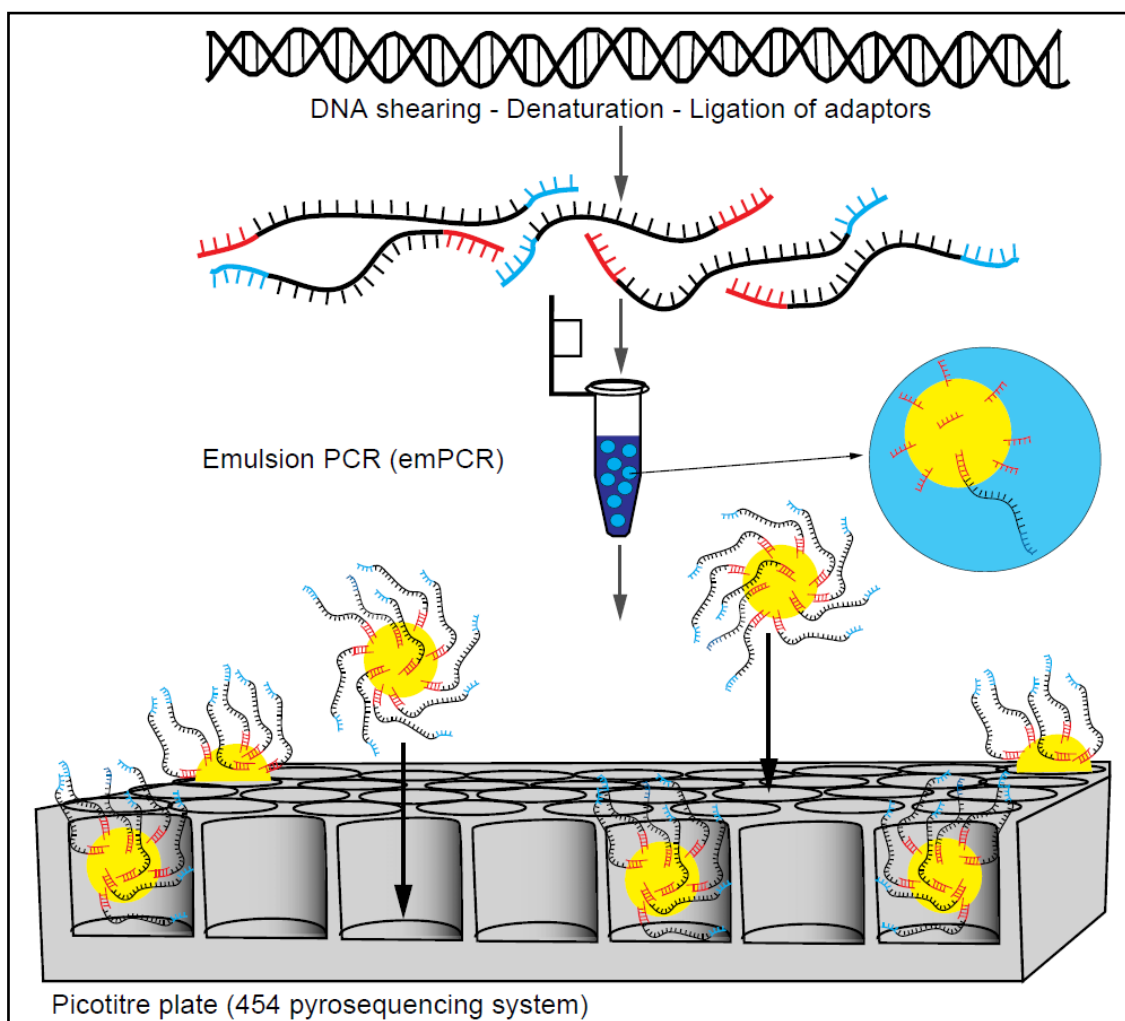


Figure 6 Schematic overview of the Roche 454 sequencing technology. This figure is reproduced after Wood et al. (2013).

1.10.2.2.2 Illumina sequencing technology

The Illumina sequencing platform performs massively parallel sequencing of millions of DNA/RNA fragments by the “sequencing by synthesis” method (Quail et al. 2008). First, DNA is fragmented into small size pieces and adapters are ligated to both ends of the fragmented DNA molecules (Fig. 7). Then, these fragments are size selected and purified. A solid glass surface is then used to generate clusters of DNA molecules destined to be sequenced. A dense amount of capture oligonucleotides are then attached to this surface to ligate with the library fragments. Single DNA molecules are hybridized to the immobilized oligonucleotides and isothermal bridge-PCR amplification results in millions of unique clusters. Finally, the prepared DNA templates are sequenced base by base in parallel using four fluorescently labelled nucleotides. After addition of each base, the clusters generate a fluorescent signal that can be used to call the added base.

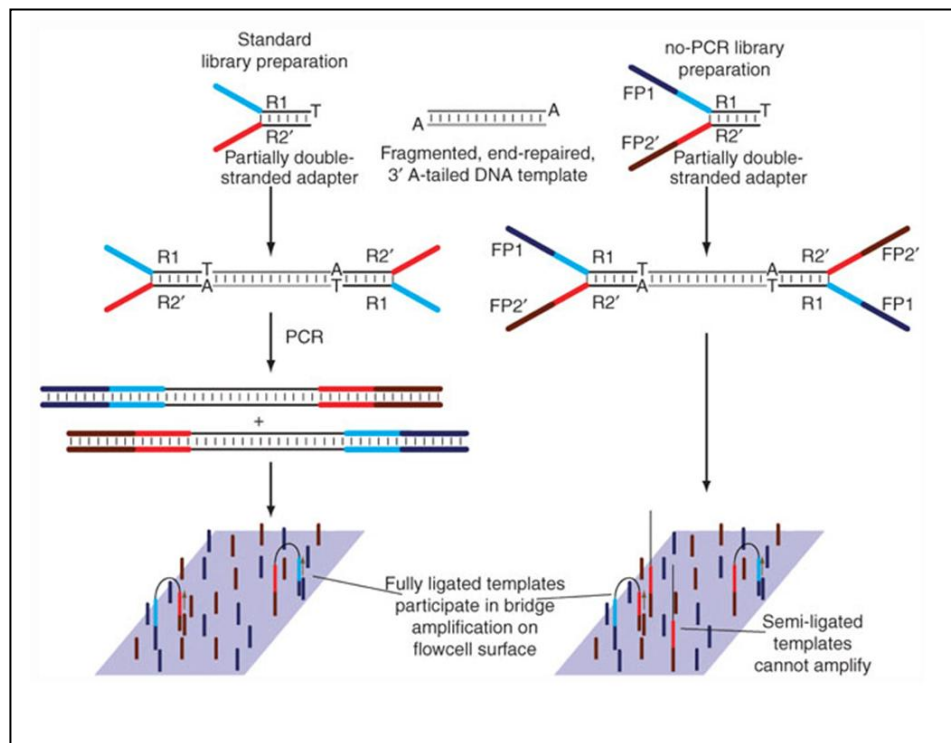


Figure 7 Overview of the Illumina sequencing workflow. This figure is reproduced from Kozarewa et al. (2009).

1.10.2.2.3 Ion Torrent sequencing technology

This semiconductor chip base sequencing platform is the newest and fastest NGS technology currently available (Wood et al. 2013). This chip has millions of wells that capture chemical information from DNA sequencing that is then translated into digital information in terms of nucleotide bases. First, the DNA sample is fragmented to small pieces. Each small fragment is attached to a single micro-bead and it is copied until the bead is covered with millions of copies of that particular DNA fragment. These beads are deposited in the wells of the semiconductor chip. Next, the chip is flooded with one of the four DNA nucleotides.

Whenever a nucleotide is incorporated to the single stranded DNA molecule, a hydrogen ion (H^+) is released and this changes the pH in the solution in the well. The ion sensitive layer below the well measures the pH change and converts it to a voltage reaction. The magnitude of voltage change indicates which nucleotide has been incorporated and the base is included in the sequence information. This process is repeated over every 15 seconds with a different nucleotide washing over the chip.

1.11 Broad objectives of the project

The central theme of my PhD project is to characterize the bacterial diversity in New Zealand groundwater at national and regional scales using molecular methods. The thesis will explore the relationships among microbial diversity, groundwater chemistry, environmental factors such as aquifer properties, and land use activities in the aquifer recharge zones. I have used several molecular approaches, including the simple molecular profiling tool, T-RFLP, as well as high-throughput NGS approaches (Roche 454 and Illumina). Due to the lack of initial information on microbiota in New Zealand groundwater ecosystems, the project began as an exploratory study and gradually expanded to test hypotheses developed based on the exploratory data obtained. The project was conducted as four main studies that are related to

each other and are briefly described below. The overall outcome of this project provides a solid platform to demonstrate to policy makers the significance of incorporating microbial assessment criteria into regional and national SOE monitoring programmes.

1.11.1 National scale assessment of groundwater bacterial diversity

Although the significance of studying groundwater microbiota is widely recognizing all over the world, it is surprising to note that the complete microbial biodiversity of groundwater has never been systematically surveyed in any country at the national scale. Therefore, one of the primary objectives of this study was to characterise the bacterial community structures of New Zealand groundwater systems at a national scale using a simple molecular fingerprinting technique: Terminal Restriction Fragment Length Polymorphism (T-RFLP). A secondary aim of this part of the study was to evaluate the relationships among bacterial diversity and geographical region, aquifer lithology, land use activities in aquifer recharge zones, well depth, groundwater chemistry and mean residence time (MRT).

1.11.2 Local scale assessment of groundwater bacterial diversity

The second main objective of this study was to explore whether the relationships between bacterial diversity and environmental factors that were observed at the national scale are consistent and stable at the local scale. For this purpose, the bacterial community structure in groundwater in the Wairarapa Region was determined and the relationships among microbial community structure and groundwater chemistry, aquifer confinement and groundwater bore usage were explored. This study was designed in a way that allows comparison of the contemporary bacterial communities in the Wairarapa Region groundwater with the results of van Bekkum et al. (2006) in an attempt to determine changes in community structure over time.

1.11.3 Relationships between bacterial diversity and hydrochemistry

In the first two studies, it is revealed that groundwater bacterial community structure is mainly related to the hydrochemistry (see results chapters 3.1 and 3.2). However, the molecular technique used in those studies (T-RFLP) does not provide very detailed or reliable taxonomic information about the populations. Therefore, my third main goal was to obtain more precise information on the species present in groundwater. For this purpose, bacterial diversity in 35 selected groundwater monitoring sites was explored using Roche 454 sequencing technology. I also tested the hypothesis that groundwater bacterial diversity is related to hydrochemistry and examined the effect of land use.

1.11.4 Horizontal gene transfer and bacterial diversity

Chapter 3.3 suggested that the bacterial diversity is shaped in a way that there are many taxa of low abundance with relatively a few highly abundant species. Further, it was found that on the basis of identifiable operational taxonomic units (OTUs), bacterial community structure is mainly related to groundwater chemistry. However, the 454 results indicated that *Pseudomonas spp.* were highly abundant and found across a range of different chemistries. Therefore, I proposed that *Pseudomonas spp.* may have acquired genetic materials from other species through horizontal gene transfer to survive and become a dominant species under various groundwater chemistries. The fourth main objective of this project was to set up a solid platform to test this hypothesis using a whole-genome sequencing approach on the Illumina MiSeq platform.

1.12 Formal statement of main hypotheses

In the following chapters, I tested these hypotheses:

Chapter 3.1: that a considerable bacterial diversity is present in New Zealand groundwater at national scale and there are identifiable relationships between bacterial diversity and environmental factors.

Chapter 3.2: that the relationships among bacterial diversity and environmental factors that are identified at a national scale are consistent and stable at a regional scale.

Chapter 3.3: that groundwater bacterial diversity is mainly related to the hydrochemistry in particular to the redox potential of groundwater.

Chapter 3.4: The Illumina MiSeq high throughput sequencing technology can be successfully used to develop a solid platform to explore whether the dominant *Pseudomonas spp.* have acquired genetic material from other species in the environment, via the process of horizontal gene transfer (HGT) which helps to maintain their dominance under different hydrochemical and environmental conditions.

References

- Bethke CM, Ding L, Jin Q, Sanford RA 2008. Origin of microbiological zoning in groundwater flows. *Geology* 36: 739–742.
- Carpenter SR, Stanley EH, Vander Zanden MJ 2011. State of the World's Freshwater Ecosystems: Physical, Chemical, and Biological Changes. *Annual Review of Environment and Resources* 36: 75–99.
- Chapelle FH 2000. The significance of microbial processes in hydrogeology and geochemistry. *Hydrogeology Journal* 8: 41–46.
- Chen K, Pachter L 2005. Bioinformatics for whole-genome shotgun sequencing of microbial communities: Review. *PLoS Computational Biology* 1: 106–112.
- Daughney CJ, Reeves RR 2005. Definition of hydrochemical facies in the New Zealand groundwater monitoring programme. *Journal of Hydrology (New Zealand)* 44: 105–130.
- Daughney CJ, Reeves RR 2006. Analysis of temporal trends in New Zealand's groundwater quality based on data from the National Groundwater Monitoring Programme. *Journal of Hydrology (New Zealand)* 45: 41–62.
- Daughney CJ, Raiber M, Moreau-Fournier M, Morgenstern U, van der Raaij R 2012. Use of hierarchical cluster analysis to assess the representativeness of a baseline groundwater quality monitoring network: Comparison of New Zealand's national and regional groundwater monitoring programs. *Hydrogeology Journal* 20: 185–200.
- Donachie SP, Foster JS, Brown MV 2007. Culture clash: challenging the dogma of microbial diversity. *The ISME Journal* 1:97–99.
- Falkowski PG, Fenchel T, Delong EF 2008. Microbial engines that derived earth's biogeochemical cycles. *Science* 320: 1034–1038.

- Ghiorse CW (1997) Subterranean life. *Science* 275: 789-790.
- Glenn TC 2011. Field guide to next-generation DNA sequencers. *Molecular Ecology Resources* 11: 759–69.
- Greater Wellington Regional Council 2013. Groundwater quality. Wellington, Greater Wellington Regional Council.
www.gw.govt.nz/groundwater-2/ [accessed 7 October 2013]
- Griebler C, Lueders T 2009. Microbial biodiversity in groundwater ecosystems. *Freshwater Biology* 54: 649–677.
- Griebler C, Stein H, Kellermann C, Berkhoff S, Brielmann H, Schmidt S, Selesi D, Steube C, Fuchs A, Hahn HJ 2010. Ecological assessment of groundwater ecosystems – vision or illusion? *Ecological Engineering* 36: 1174–1190.
- Haack SK, Fogarty LR, West TG, Alm EW, McGuire JT, Long DT, Hyndman DW, Forney LJ 2004. Spatial and temporal changes in microbial community structure associated with recharge-influenced chemical gradients in a contaminated aquifer. *Environmental Microbiology* 6: 438–448.
- Handelsman J, Tiedje JM, Alvarez-Cohen L, Ashburner M, Cann IKO, Delong EF et al. 2007. *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. Washington, DC, United States, The National Academies Press.
- Hedrich S, Schlömann M, Johnson DB 2011. The iron-oxidizing proteobacteria. *Microbiology* 157: 64.
- Humphreys WF 2009. Hydrogeology and groundwater ecology: Does each inform the other? *Hydrogeology Journal* 17: 5–21.
- Janssen PH, Yates PS, Grinton BE, Taylor PM, Sait M 2002. Improved culturability of soil bacteria and isolation in pure culture of novel members of the divisions Acidobacteria,

- Actinobacteria, Proteobacteria, and Verrucomicrobia. *Applied and Environmental Microbiology* 68: 2391–2396.
- Kim KC, Byrne LB 2006. Biodiversity loss and the taxonomic bottleneck: emerging biodiversity science. *Ecological Research* 2: 794–810.
- Korbel KL, Hose GC 2011. A tiered framework for assessing groundwater ecosystem health. *Hydrobiologia* 661:329–349.
- Korbel KL, Hancock PJ, Serov P, Lim RP, Hose GC 2013. Groundwater ecosystems vary with land use across a mixed agricultural landscape. *Journal of Environmental Quality* 42: 380–390.
- Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+ C)-biased genomes. *Nature methods*, 6: 291–295.
- Larned ST 2012. Phreatic groundwater ecosystems: research frontiers for freshwater ecology. *Freshwater Biology* 57: 885–906.
- Liu W, Marsh TL, Cheng H, LJ Forney 1997. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Applied and Environmental Microbiology* 63: 4516–4522.
- Lozupone CA, Knight R 2007. Global patterns in bacterial diversity. *Proceedings of the National Academy of Sciences USA* 104: 11436–11440.
- Maier MR, Pepper IL, Gerba CP 2009. *Environmental Microbiology* (2nd Edition). San Diego, CA, United States, Academic Press.
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bembien LA et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376–80.
- Ministry for the Environment 2010. *Faecal pollution (bacteria) in groundwater*. Wellington, New Zealand, Ministry for the Environment.

www.mfe.govt.nz/environmental-reporting/fresh-water/groundwater-quality-indicator/faecal-pollution.html [accessed 7 October 2013].

Morgenstern U, Daughney CJ 2012. Groundwater age for identification of baseline groundwater quality and the impacts of land-use intensification - The National Groundwater Monitoring Programme of New Zealand. *Journal of Hydrology* 456/457: 79–93.

Muyzer G 1999. DGGE/TGGE a method for identifying genes from natural ecosystems. *Current Opinion in Microbiology* 2: 317–322.

Muyzer G, De Wall EC, Uiterlinden AG 1993. Profiling of Complex Microbial Populations by Denaturing Gradient Gel Electrophoresis Analysis of Polymerase Chain Reaction-Amplified Genes Coding for 16S rRNA. *Applied and Environmental Microbiology* 59: 695–700.

National Groundwater Association 2010. Groundwater facts. United States, Ohio, National Groundwater association.

www.ngwa.org/Fundamentals/use/Documents/gwfactsheet.pdf [accessed 31 March 2014]

Neufeld JD, Mohn WW 2005. Unexpectedly high bacterial diversity in arctic tundra relative to boreal forest soils, revealed by serial analysis of ribosomal sequence tags. *Applied and Environmental Microbiology* 71: 5710–5718.

Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R et al. 2008. A large genome center's improvements to the Illumina sequencing system. *Nature Methods* 5: 1005–1010.

Rajanayaka C, Donaggio J, McEwan H 2010. Update of Water Allocation Data and Estimate of Actual Water Use of Consented Takes - 2009-10 Update of Water Allocation Data and Estimate of Actual Water Use of Consented Takes - 2009-10. Wellington, New Zealand, Ministry for the Environment. 2 p.

- Ranjard L, Poly F, Lata JC, Mougél C, Thioulouse J, Nazaret S 2001. Characterization of bacterial and fungal soil communities by automated ribosomal intergenic spacer analysis fingerprints: biological and methodological variability. *Applied and Environmental Microbiology* 67: 4479–4487.
- Rosen MR 2001. Hydrochemistry of New Zealand's aquifers. In: Rosen MR, White PA eds. *Groundwaters of New Zealand*. Wellington, The New Zealand Hydrological Society. Pp. 77–110.
- Rothberg JM, Leamon JH 2008. The development and impact of 454 sequencing. *Nature Biotechnology* 26: 1117–24.
- Sanger F, Nicklen S, Coulson AR 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences USA* 74: 5463–5467.
- Shiklomanov IA 2000. Appraisal and assessment of world water resources. *Water International* 25: 11–32.
- Siebert S, Burke J, Faures JM, Frenken K, Hoogeveen J, Döll P, Portmann FT 2010. Groundwater use for irrigation – a global inventory. *Hydrology and Earth System Sciences* 14: 1863–1880.
- Sinreich M, Pronk M & Kozel R (2011) Microbiological spring water quality monitoring across Switzerland. *Proc H2Karst, 9th Conference on Limestone Hydrogeology, Besançon (France) 1-3 Sep. 2011*: 447-450.
- Stein H, Kellermann C, Schmidt SI, Brielmann H, Steube C, Berkhoff SE, Fuchs A, Hahn HJ, Thulin B, Griebler C 2010. The potential use of fauna and bacteria as ecological indicators for the assessment of groundwater quality. *Journal of Environmental Monitoring* 12: 242–254.
- Struckmeier W, Rubin Y, Jones JAA 2005. Groundwater - reservoir for a thirsty planet?: *Earth Sciences for Society; a Prospectus for a Key Theme of the International Year of*

- Planet Earth. Trondheim, IUGS International Union of Geological Sciences Secretariat, Geological Survey of Norway.
- Torsvik V, Sørheim R, Goksøyr J, 1996. Total bacterial diversity in soil and sediment communities: a review. *Journal of Industrial Microbiology* 17: 170-178.
- Wood S, Smith K, Banks J et al. 2013. Molecular genetic tools for environmental monitoring of New Zealand's aquatic habitats, past, present and the future. *New Zealand Journal of Marine and Freshwater Research* 47: 90–119.
- van Bekkum M, Sainsbury JP, Daughney CJ, Chambers GK 2006. Molecular analysis of bacterial communities in groundwaters from selected wells in the Hutt Valley and the Wairarapa, New Zealand. *New Zealand Journal of Marine and Freshwater Research* 40: 91–106.
- Waikato Regional Council (2014) Groundwater. Waikato Regional Council, Hamilton, New Zealand.
- <http://www.waikatoregion.govt.nz/Environment/Natural-resources/Water/Groundwater/>
- Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC et al. 2012. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* 337: 1661-1665.
- Younger PL 2007. Groundwater in the environment: an introduction. Malden, MA, United States, Blackwell Publishing.
- Zhou Y, Kellermann C, Griebler C 2012. Spatio-temporal patterns of microbial communities in a hydrologically dynamic pristine aquifer. *FEMS Microbiology Ecology* 81: 230–242.
- Zhou J, Davey ME, Figueras JB, Rivkina E, Gilichinsky D, Tiedje JM 1997. Phylogenetic diversity of a bacterial community determined from Siberian tundra soil DNA. *Microbiology* 143: 3913-3919.

EXTENDED INFORMATION ON MATERIALS AND METHODS

The detailed experimental protocols are given in each results chapter (Chapter 3.2 – 3.5). Additional information pertaining to these experiments and general laboratory methods are provided in this section.

2.11 Groundwater Sampling Strategy

Groundwater sample collection for the entire project was conducted as two sets: 1) single aliquots from 100 sites were sampled across New Zealand in June 2010 with the collaboration of the National Groundwater Monitoring Programme (NGMP) operated by the GNS Science; and 2) single aliquots from 35 sites were sampled around the Wairarapa region, Wellington in September 2009 with the collaboration of the Greater Wellington Regional Council as a part of their quarterly groundwater monitoring practices. The national scale study (Chapter 3.1) was based on Set 1, whereas the local scale analysis of microbial diversity using the terminal restriction fragment length polymorphism (T-RFLP) tool (Chapter 3.2) utilized groundwater samples from Set 2. The Roche 454 pyrosequencing analysis of bacterial diversity (Chapter 3.3) and the Illumina high throughput sequencing analysis of groundwater bacterial metagenomes (Chapter 3.4) used representative samples from Set 1.

The 2-litre plastic bottles that were used for collection of water were sterilized prior to use with three washing steps: 1) rinsed with double distilled water (ddH₂O); 2) washed with 70% ethanol (EtOH); and 3) re-washed with double distilled water (ddH₂O) followed by an air-drying step. The groundwater sampling was performed according to the National Protocol for State of the Environment Groundwater Sampling (Daughney et al. 2006).

2.12 General Laboratory Practices

All laboratory practices used standard sterilized conditions. Glassware, plasticware, pipette tips, ddH₂O and Tris-Acetate-EDTA (TAE) buffer were sterilized by autoclaving at 20 psi (121 °C) for 20 minutes. A new pair of sterile disposable rubber gloves was used in each reaction block. The post-PCR and pre-PCR activities were conducted in designated areas to avoid any possible cross-contamination.

2.13 DNA Quantification

The DNA quantifications for method validation steps were conducted using NanoPhotometer™ Pearl (IMPLEN, Germany) because it provides results easily and quickly. However, accurate DNA quantification is crucial in T-RFLP and high throughput sequencing methodologies. Therefore, the quantity of DNA in extracts and PCR products used in these experimental applications was determined using Quant-iT™ High-Sensitivity DNA Assay kits (Invitrogen, United States) as per the manufacturer's instructions. A calibration curve was constructed using the standard DNA mixtures and absolute concentration values expressed in ng/μl were determined accordingly.

2.14 Control Experiments

Control experiments were conducted at all major steps: groundwater filtrations; DNA extractions; and PCR amplifications. A 2-litre aliquot of sterile ddH₂O was collected in the same type of plastic bottle that was sterilized as described above and filtered along with the groundwater filtrations. The standard DNA extraction protocol used for groundwater samples (Chapters 3.1 and 3.2) was applied to the ddH₂O filter to obtain a DNA extract. The PCR amplification was performed (as described in Chapters 3.1 and 3.2) using this extract as

the template DNA. The absence of PCR product on the agarose gel verified that the introduction of non-groundwater source bacteria into the samples during the sample collection and water filtration was minimal. A T-RFLP profile was obtained for the ddH₂O DNA extract as described in Chapter 3.2 and it was used as the negative control for the T-RFLP analyses. Further, each PCR reaction (in T-RFLP and 454 pyrosequencing studies) was accompanied with a negative control reaction using ddH₂O and a positive control reaction using *Escherichia coli* DH5 α genomic DNA as the DNA templates.

2.15 Visualization of PCR and Restriction Digestion Products

The PCR amplification and restriction digestions were confirmed by running aliquots of products on a 1% agarose gel in 1 X TAE buffer (Life Technologies, United States), stained with ethidium bromide and visualized by UV trans-illumination. For the gel purification steps outlined in Chapter 3.3, PCR products were run on a 2% agarose gel stained with SYBR[®] Safe in 1X TAE buffer (Life Technologies, United States) and visualized by a Safe Imager[™] 2.0 Blue Light Transilluminator (Life Technologies, United States). Here, blue light was used instead of UV to avoid any potential damages to the DNA as the purified PCR products were subjected for Roche 454 sequencing.

2.16 Direct DNA sequencing

Conventional Sanger DNA sequencing of *Escherichia coli* DH5 α and groundwater bacterial 16S rRNA gene (discussed in Chapters 3.1 and 3.2) was performed on the PCR product obtained using the two universal oligonucleotide primers: forward primer - F63 (5'-CAG GCC TAA CAC ATG CAA GTC-3') and reverse primer - R1389 (5'-ACG GGC GGT GTG TAC AAG-3') as described by van Bakkum et al. (2006). Briefly, the target 16S rRNA gene region was amplified using a Mastercycler Pro S PCR system (Eppendorf, Germany) in a 50

µl reaction mixture that contained approximately 20 ng (2 µl) of template DNA, 1 µl of each primer (10µM), 30 µl of 2X BioMix™ (BIOLINE, UK), 0.5 µl of MgCl₂ (3 mM) and approximately 15.5 µl of molecular biology grade water to adjust the final volume. The PCR cycling regime was as follows: initial denaturation at 94 °C for 3 min; 30 cycles at 94 °C for 20 s, 56 °C for 20 s and 72 °C for 1 min; and final extension at 72 °C for 10 min (Sirisena et al. 2013). The PCR products were purified using DNA Clean & Concentrator™ kits and quantified using NanoPhotometer™ Pearl. The DNA sequencing was conducted on the ABI3730 DNA Analyzer (Life Technologies, United States) at the Massey Genome Service, Palmerston North, New Zealand. Two DNA sequences were obtained for each sample using the forward and reverse primers. The sequencing chromatograms were visualized using the Sequence Scanner Software v1.0 (Life Technologies, United States). The MEGA 4 software (Tamura et al. 2007) was used to generate the complement of the reverse sequence and align it with the forward sequence to obtain a consensus contig sequence for each sample.

2.17 Terminal Restriction Fragment Length Polymorphism (T-RFLP)

The groundwater microbial community structure was determined at national (Chapter 3.1) and local (Chapter 3.2) scales using the terminal restriction fragment length polymorphism (T-RFLP) technique. This methodology was chosen as it was successfully employed in a pilot study assessing the bacterial diversity of New Zealand groundwater at local scale by van Bekkum et al. (2006). Their protocol was validated and used in this study with slight modifications as described in Chapter 3.1. The genotyping was performed on the ABI3730XL DNA Analyzer (Life Technologies, United States) at Macrogen, Inc., Republic of Korea.

2.18 Roche 454 Pyrosequencing

Amplicon library preparation for pyrosequencing was conducted according to the protocol developed by the Craig Cary Lab, at the University of Waikato, and a detailed description of the methodology is given in Chapter 3.3. Extreme care was taken to avoid the possible contamination of PCR mix from foreign DNA and to ensure the PCR products were solely derived from the groundwater bacterial DNA sources. In addition to the standard sterilization practices outlined above, the entire PCR clean room with pipette tips, plasticware, molecular biology grade water and PCR buffer was exposed to UV light for 15 minutes prior to the reaction preparations to ensure extreme DNA-free conditions.

The PCR master-mix that includes all the components except template DNA was treated with ethidium monoazide bromide (EMA) to remove any potential contaminating double-stranded DNA (ds DNA) that might arise through a contaminated reagent or DNA polymerase enzyme (Rueckert & Morgan 2007). The EMA is a fluorescent nucleic acid stain that can be covalently intercalated with ds DNA to produce a stable complex at the absence of light. The EMA-bound DNA is unable to participate in PCR amplification. The free EMA in the solution is photolyzed by light and is no longer capable of making covalent attachments when the template ds DNA is added (Nogva et al 2003).

EMA stock solution was prepared by adding 500 μ l of absolute ethanol into 5 mg of EMA according to the manufacturer's protocol (Biotium, United States). The working solution of EMA was prepared by diluting the stock solution 1:200: 1 μ l of EMA stock solution was added into 199 μ l of molecular biology grade water. In each PCR reaction setup, 1 % (V/V) of working EMA solution was added to the PCR master-mix (e. g. 5 μ l of EMA for 500 μ l PCR master-mix) and vortexed for 2-3 seconds. The mixture was incubated on ice for 1 min allowing EMA to intercalate with any contaminating ds DNA. All these steps were conducted in complete darkness. Finally, the master-mix was exposed to a strong

incandescent light for 1 min to degrade the unreacted EMA. After the EMA treatment, the PCR master-mix was processed as described in Chapter 3.3. A new EMA working solution was prepared for each PCR setup.

Previous validations conducted by Craig Cary's research group suggested that a single 454 parallel sequencing run on the Roche GS Junior system resulted in approximately 90,000 reads. Based on this information, it was decided to multiplex 8 groundwater DNA samples in a single sequencing run as a trade-off between the yield and the resolution of the technique. Therefore, the forward primers for each sample were labelled with unique Multiplex Identifiers (MIDs), which are short DNA sequences. The primers were designed according to the instructions given for the fusion primer design in the Roche GS Junior System Guidelines for Amplicon Experimental Design Manual (454 Life Sciences Corporation 2010). Detailed description of the MIDs that were used in this study was given in Table 1.

MID ID	MID Sequence
MID 1	CGAGCT
MID 2	ATCGTC
MID 3	ACGATACG
MID 4	ACTGAC
MID 5	ATATCGTAC
MID 6	TGTCACGT
MID 7	TCCTGACG
MID 8	CGTCTAGTA
MID 13	CTATAGCTG

Table 1 DNA sequences of the MIDs used in this study

2.19 Illumina MiSeq[™] High Throughput Sequencing

A detailed description of Illumina MiSeq[™] High Throughput Sequencing methodology is outlined in Chapter 3.5. All the library preparations, sequencing and bioinformatics analyses were conducted at Massey Genome Service, Palmerston North, New Zealand.

2.20 Data Analysis

2.20.1 Terminal Restriction Fragment Length Polymorphism (T-RFLP)

The electropherogram representations of T-RFLP profiles shown in this thesis were obtained from the Peak Scanner[™] Software v1.0 (Life Technologies, United States) using standard parameters whereas the numerical data related to the profiles (i. e. absolute peak heights and fragment lengths) were obtained using the GeneMapper[®] v3.1 software (Life Technologies, United States). Further, a Microsoft Excel macro programme was used to: 1) filter the peaks with heights above the threshold; 2) bin the terminal restriction fragments into correct fragment sizes; and 3) standardize the peak heights to the highest peak and the sum of all peaks. The Shannon diversity indices (H') for the T-RFLP data (in Chapters 3.2 and 3.3) were manually calculated using Microsoft Excel 2010 package. Kruskal-Wallis and chi-square tests were performed using SPSS version 19 (SPSS IBM, New York, USA). The calculation of distance matrices (Euclidean and Bray-Curtis) and constructions of dendrograms and box-and-whisker plots were performed using the R (version 2.15.0) statistical programme (R Core Team 2013). All the graphs were constructed using Microsoft Excel 2010.

2.20.2 Roche 454 Pyrosequencing Data Analysis

A substantial amount of pyrosequencing data analysis was conducted at the University of Waikato DNA Sequencing Facility. This included the separation of sequences according to

MIDs, quality filtration of reads, identification of operational taxonomic units (OTUs) and taxonomic identification of OTUs. I have performed all the other statistical analyses and detailed explanations are given in Chapter 3.4.

2.20.3 Illumina MiSeq[™] High-Throughput Sequencing

The Illumina high-throughput sequencing of six groundwater DNA samples generated an enormous amount of sequence data. Therefore, the complete analysis of the whole data set is beyond the scope of my PhD project as it is highly time consuming. All the sequence data analyses conducted to date were performed at Massey Genome Service, Palmerston North, New Zealand. I have done the data interpretations with relevance to the groundwater microbial ecology and a detailed description is outlined in Chapter 3.5. The remainder of the genomic data analyses will be conducted after the submission of my PhD thesis with the collaboration of the Massey Genome Service.

References

- 454 Life Sciences Corporation 2010. Roche GS Junior System Guidelines for Amplicon Experimental Design Manual, Connecticut, United States, 454 Life Sciences Corporation.
- Daughney CJ, Jones A, Baker T, Hanson C, Davidson P, Zemansky GM, Reeves RR, Thompson M 2006. A national protocol for state of the environment groundwater sampling in New Zealand: ME 781. Wellington, New Zealand, Ministry for the Environment.
- Nogva HK, Dromtorp SM, Nissen H, Rudi K 2003. Ethidium monoazide for DNA-based differentiation of viable and dead bacteria by 5'-nuclease PCR. *Biotechniques*, 34: 804-813.
- R Core Team 2013. R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing.
<http://www.R-project.org/>
- Rueckert A, Morgan HW 2007. Removal of contaminating DNA from polymerase chain reaction using ethidium monoazide. *Journal of Microbiological Methods* 68: 596-600.
- Sirisena KA, Daughney CJ, Moreau-Fournier M, Ryan KG, Chambers GK 2013. National survey of molecular bacterial diversity of New Zealand groundwater: relationships between biodiversity, groundwater chemistry and aquifer characteristics. *FEMS Microbiology Ecology* 86: 490–504.
- Tamura K, Dudley J, Nei M, Kumar S 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* 24: 1596–1599.

van Bekkum M, Sainsbury JP, Daughney CJ, Chambers GK 2006. Molecular analysis of bacterial communities in groundwaters from selected wells in the Hutt Valley and the Wairarapa, New Zealand. *New Zealand Journal of Marine and Freshwater Research* 40: 91–106.

**National Survey of Molecular Bacterial Diversity of New Zealand
Groundwater: Relationships between Biodiversity, Groundwater
Chemistry and Aquifer Characteristics**

Status:

Sirisena KA, Daughney CJ, Moreau-Fournier M, Ryan KG, Chambers GK 2013. National survey of molecular bacterial diversity of New Zealand groundwater: relationships between biodiversity, groundwater chemistry and aquifer characteristics. *FEMS Microbiology Ecology* 86: 490-504.

Keywords

Bacteria; T-RFLP; New Zealand National Groundwater Monitoring Programme; NGMP;
DNA Analysis

Running title: Bacterial diversity in New Zealand groundwater

Abstract

Groundwater is a vital component of rural and urban water supplies in New Zealand.

Although extensive monitoring of chemical and physical properties is conducted due to the high demand for this valuable resource, current information on its bacterial content is limited.

However, bacteria provide an immense contribution to drive the biogeochemical processes in the groundwater ecosystem as in any other ecosystem. Therefore, a proper understanding of bacterial diversity is crucial to assess the effectiveness of groundwater management policies.

In this study, we investigated the bacterial community structure in NZ groundwater at a national scale using the terminal restriction fragment length polymorphism (T-RFLP) molecular profiling tool and determined the relationships among bacterial diversity and groundwater chemistry, geological parameters and human impact. Considerable bacterial diversity was present and the community structures were strongly related to groundwater chemistry, and in particular to redox potential and human impact, reflecting their potential influence on determination of bacterial diversity. Further, the mean residence time of groundwater also showed relationships with bacterial community structure. These novel findings pertaining to community composition and its relationships with environmental parameters will provide a strong foundation for qualitative exploration of the bacterial diversity in NZ groundwater in relation to sustainable management of this valuable resource.

Introduction

Groundwater is a valuable natural resource that accounts for nearly 99% of the total volume of fresh water presently circulating on our planet (Younger, 2007). It is the world's major drinking water source, providing about 60% of drinking water in Europe with an even greater percentage in individual countries and more than 80% in North Africa and the Middle East (Struckmeier *et al.*, 2005; Steube *et al.*, 2009). In New Zealand, groundwater is also an integral part of urban and rural water supplies. Nearly one quarter of the New Zealand population uses groundwater as its major drinking water source; groundwater also supplies a significant fraction of the water requirements for the agricultural and industrial sectors (Daughney & Reeves, 2005).

Due to the importance of groundwater, its quality and availability are extensively monitored, both in New Zealand and globally. State-of-the-environment (SOE) monitoring is typically conducted at a regional or national scale and is also referred to as baseline, background, ambient or long-term monitoring. In general any SOE monitoring scheme aims to: 1) characterize groundwater quality in terms of its current state and trends; 2) associate the observed state and trends with specific causes such as land use, pollution or natural processes; and 3) provide data to assess the effectiveness of groundwater management policies. SOE monitoring typically involves regular collection of groundwater samples from a fixed network of sites followed by analyses of these samples for a suite of physical and chemical parameters. In New Zealand, SOE monitoring is undertaken through the National Groundwater Monitoring Programme (NGMP) and through the regional networks operated by 15 regional authorities (Daughney *et al.*, 2012).

However, with the recent advances of policy planning, groundwater is now considered not only as a valuable resource for human consumption, but also as a dynamic

ecosystem. In some parts of Europe and Australia, assessments of ecological status have already been included into their national groundwater monitoring policies (Steube *et al.*, 2009; Griebler *et al.*, 2010; Stein *et al.*, 2010; Korbel & Hose, 2011). Microorganisms are the key driving force in this ecosystem. Therefore, it is crucial to understand the reference status of the microbial components in differing aquifer systems, which will enable us to test hypotheses related to trends in relationships between microorganisms and the physicochemical environment (Larned, 2012).

To date, in most parts of the world including New Zealand, SOE monitoring has almost completely overlooked the microbiological component of groundwater systems. Only the presence of *E. coli* is regularly monitored in New Zealand groundwater, because it is an indicator species of faecal contamination (Ministry for the Environment, 2010). During recent years, an increasing number of studies have been conducted in some parts of Europe and Australia to fulfil this necessity: assessing bacterial parameters in groundwater including bacterial diversity and its relationships with biogeographical and hydrochemical conditions across varying spatial and temporal scales (Griebler *et al.*, 2010; Stein *et al.*, 2010; Sinreich *et al.*, 2011; Zhou *et al.*, 2012; Korbel *et al.*, 2013). However, to our knowledge, the complete microbial biodiversity of groundwater, including all indigenous species, has never been systematically surveyed in any country at the national scale. This is surprising given the fact that most biogeochemical cycles on the planet are strongly influenced by subsurface microbial communities (Falkowski, 2008). The microbial communities are selected and regulated by the chemical and physical nature of groundwater and, conversely, the groundwater microbial communities mediate redox reactions while obtaining energy for survival, thus controlling the dissolved concentrations of elements such as iron, manganese, nitrogen, sulphur and many others (Ghiorse, 1997; Chapelle, 2000; Bethke *et al.*, 2008; Hedrich *et al.*, 2011). It is further expected that any change in the chemical composition of

groundwater or aquifer sediment will cause a corresponding shift in the subsurface microbial community structure (Haack *et al.*, 2004).

One primary aim of this study was to characterise the bacterial community structures of New Zealand groundwater systems at a national scale. The second primary aim of this study was to evaluate the relationships between bacterial diversity and geographical region, aquifer lithology, land use activities in aquifer recharge zones, well depth, groundwater chemistry and mean residence time (MRT). For these two purposes, the NGMP provided a useful platform because of its national coverage and the range of conditions represented for each of the above-mentioned variables.

There are number of techniques available to study bacterial diversity in subsurface environments including groundwater. Culturing techniques have been frequently used for this type of survey, but it is now widely believed that many bacterial species present in environmental samples cannot be easily grown in artificial culture media (Zhou *et al.*, 1997; Janssen *et al.*, 2002; Neufeld & Mohn, 2005; Lozupone & Knight, 2007). With the recent advances of molecular tools, an array of DNA-based methods is available to explore subsurface microbial diversity (Maier *et al.*, 2009). Due to the lack of pre-existing knowledge on New Zealand groundwater bacterial diversity, we chose Terminal Restriction Fragment Length Polymorphism (T-RFLP), which is a relatively simple, rapid and cost-effective molecular profiling tool (Edlund *et al.*, 2006; van Bekkum *et al.*, 2006) and it provides highly valid comparable results even in the era of next generation sequencing technologies (Camarinha-Silva *et al.*, 2012; Pilloni *et al.*, 2012). However, T-RFLP does not provide taxonomic information on diversity in an absolute sense, but it does provide a quantitative insight into bacterial community structure. These results helped to provide a baseline for the bacteria community structure and were used in combination with various statistical techniques to evaluate relationships between groundwater chemistry and aquifer

properties. The overall outcome of this study has been to generate the first picture in depth of bacterial diversity and its significance in New Zealand groundwater ecosystems and enables us to begin to emphasise to policy makers the significance of incorporating microbial assessment criteria into SOE monitoring programmes. In addition, the study laid a solid foundation for two follow up studies, which are designed to test hypotheses on relationships between bacterial diversity and hydrochemical and environmental parameters, using high throughput sequencing technologies.

Materials and methods

Study area and groundwater sampling

Groundwater samples were available from 100 of the 110 different sites comprising the NGMP (Fig. 1). The NGMP is a long-term research and monitoring programme that aims to identify spatial patterns and temporal trends in groundwater quality at the national scale and relate them to specific causes (Rosen, 2001; Daughney & Reeves, 2005, 2006; Morgenstern & Daughney, 2012). The NGMP sites are located in discrete aquifers (or on discrete flow lines in larger aquifer systems) and provide a highly representative picture of groundwater quality across New Zealand (Daughney *et al.*, 2012). Site-specific details pertaining to the NGMP are available in the Electronic Supplementary Material in Daughney *et al.* (2010) and from the GNS Science Geothermal and Groundwater (GGW) Database (<http://ggw.gns.cri.nz/ggwdata/mainPage.jsp>).

Groundwater samples (2 litres from each site) were collected in June 2010 into individual sterilized plastic bottles according to the National Protocol for State of the Environment Groundwater Sampling (Daughney *et al.*, 2006). All samples were kept at

4°C during transportation and refrigerated until they were used. Additional samples were collected by the regional council staffs at the same time as a part of routine operations in the NGMP; these samples were analysed at GNS Science for a suite of compounds (Na, K, Mg, Ca, HCO₃, Cl, SO₄, NO₃-N, NH₄-N, PO₄-P, Fe, Mn, Br, F and SiO₂) in accordance with standard procedures (Daughney & Reeves, 2005; Daughney *et al.*, 2010). Dissolved oxygen, electrical conductivity, pH and temperature were also measured in the field at the time of sampling. The hydrochemical data are available from the GNS Science Geothermal and Groundwater (GGW) Database (<http://ggw.gns.cri.nz/ggwdata/mainPage.jsp>).

Groundwater filtration and DNA extraction

Each two litre sample was filtered through a sterile 0.22 µm nitrocellulose membrane filter (Millipore, Australia) using a vacuum system. Long sample storage periods were avoided by conducting the filtrations immediately after they were received as delay can (and in our experience, does) lead to apparent alterations in community composition in stored samples. The filters with the retained bacteria were frozen at -20°C in sterilized 50 mL plastic tubes until use. For DNA extraction, 10 mL of sterile double distilled water were added to each tube and kept for 5 min in a slant position to soak the membrane filters. Each filter was abraded with a sterile plastic inoculation loop to transfer bacterial cells from the filter into the water. The tubes were centrifuged at 3000 g for 20 min to recover bacterial cells as a pellet which was then used as the starting material for the DNA extraction. Bacterial genomic DNA extractions were performed using ZR Fungal/Bacterial DNA kits (Zymo Research, United States) as directed by the manufacturer. The concentrations of DNA extracts were quantified using Quant-iT™ High-Sensitivity DNA Assay kits (Invitrogen, United States). Extracted

DNA was dissolved in 100 µl of molecular biology grade water (MO BIO Laboratories, Inc., United States) and stored at -20°C until further use.

PCR amplification of bacterial 16S rRNA gene

The PCR was performed on bacterial genomic DNA to amplify an approximately 1400 bp region of bacterial 16S rRNA gene using two fluorescently labelled bacterial specific oligonucleotide primers: F63 (5'-[6-FAM]CAG GCC TAA CAC ATG CAA GTC-3') and R1389 (5'-[6-HEX]ACG GGC GGT GTG TAC AAG-3'). (van Bekkum *et al.*, 2006; Parkinson, 2009). The reaction mixture contained 20 ng of bacterial genomic DNA extract, 1.5 µL of each primer (final concentration of 0.3 µM), 25 µl of BioMix™ PCR mix containing dNTPs and *Taq* DNA polymerase (BIOLINE, United Kingdom), 0.5 µl of MgCl₂ (final concentration of 2.5 mM) and 18.5 µL of molecular biology grade water in a total volume of 50 µL. For each sample, reactions were performed in triplicate and the PCR products were pooled for downstream application. The amplifications were performed in a Mastercycler® Pro S PCR system (Eppendorf, Germany). The PCR regime consisted of an initial denaturation step of 3 min at 94°C followed by 30 cycles at 94°C for 20 s, 56°C for 20 s and 72°C for 1 min. The reaction was completed with a final extension at 72°C for 10 min and then held at 4°C. The PCR products were purified using DNA Clean & Concentrator™ kits (Zymo Research, United States) and recovered dsDNA products were eluted in 25 µL of sterile double distilled water. The purified PCR products were quantified using Quant-iT™ High-Sensitivity DNA Assay kits and stored at -20°C until further use.

Restriction enzyme digestion

For T-RFLP analysis, purified PCR product (500 ng) was digested with 10 U of AluI (Roche, United States) in a total volume of 25 μ L. We used AluI for this study because it worked reliably in our experimental systems and is one of the most commonly used enzymes for T-RFLP investigations (Osborn *et al.*, 2006; Parkinson, 2004; van Bekkum *et al.*, 2006). The reaction mixture was incubated at 37°C for 5 h and the enzyme was inactivated by heating at 65°C for 20 min. Complete restriction digestion was confirmed by running aliquots of products on a 2 % agarose gel.

T-RFLP genotyping by automated capillary electrophoresis

The restriction products were run on an ABI 3730XL DNA Analyzer (Applied Biosystems Inc., United States) to separate fluorescently labelled terminal restriction fragments (T-RFs), resulting in a unique genotype profile for each sample. The T-RFLP profile is a graphical representation of the bacterial diversity of the sample. Ideally, a single bacterial taxon should be represented by two peaks (one FAM and one HEX) in the profile but there could be instances where a single peak (FAM or HEX) represents one taxon. The peak height that is given as fluorescence intensity indicates the relative abundance of the corresponding taxon. The samples were run along with a GeneScan™-400HD ROX™ internal size standard (Applied Biosystems Inc., United States), which permits precise determination of the fragment sizes up to 400 bp in length. The T-RFLP electropherograms were converted to numerical barcodes using GeneMapper® v 3.1 software (Applied Biosystems Inc., United States). The binary presence (1) or absence (0), fragment sizes (bp) and heights corresponding to each peak were tabulated using a bin size of 1 bp.

Quality controls

All DNA extractions were carried out with extract control to ensure the reagent solutions used were free of DNA contamination, e.g. from foreign microorganisms from the surrounding environment, or carry over contamination. As the negative control, 2 litres of sterile double distilled water were filtered and DNA was extracted and analysed as described above. Contamination-free DNA extraction was confirmed by performing PCR amplification using this extract as template DNA and checking the product on a 1% agarose gel. The PCR product obtained from the extract control was digested with AluI restriction endonuclease to obtain a blank genotype profile, which was used as the negative control in the T-RFLP data analysis and used to determine the baseline relative fluorescence units (RFU) threshold for true peaks.

For the validation of correct PCR amplification, a single PCR product from *E. coli* DH5 α strain was sequenced with unlabelled F63 and R1389 primers by direct DNA sequencing and checked against a databank of microbial DNA sequences at NCBI (<http://blast.ncbi.nlm.nih.gov/>). A virtual restriction map was constructed for AluI, TaqI, MspI, AvaII, and MaeIII restriction endonucleases using this DNA sequence and hypothetical 16S T-RFLP profiles were constructed with each enzyme for the *E. coli* DH5 α strain. Actual T-RFLP profiles were also obtained with the same enzymes and compared to the hypothetical profiles.

Data analysis

A number of approaches have previously been used to prepare T-RFLP data for quantitative analysis, depending on the scale and objective of the study. Some studies have counted only

the binary presence (1) or absence (0) of peaks (Anderson *et al.*, 2010), whereas other studies take account of the peak heights too. In some studies, peak heights were standardised relative to the highest peak in the profile (Parkinson, 2004, 2009; van Bekkum *et al.*, 2006) while in other studies the peak heights were standardised relative to the sum of all peaks in each profile (Culman *et al.*, 2008). This shows that there is no commonly accepted best practice for standardizing T-RFs and we have analysed our data using both approaches, separately for FAM and HEX signals. In our data analysis, a fixed value of 200 RFU was used to separate true peaks from the background noise, because no peaks over this size were observed in the negative control T-RFLP profile. In this approach, we used an experimental value to determine the RFU threshold whereas most previous studies used an arbitrary value or hypothetical value as the RFU threshold. Note that most of the resulting fragment lengths in this dataset were referenced with decimal values. Therefore, these original decimal fragment sizes were rounded to the nearest integer value using ± 0.5 bp as the binning threshold (i.e. to the nearest 1 bp). If two or more decimal fragment sizes were assigned to a single bin size after rounding, heights of the peaks were summed as if they were a single peak. Finally, peaks over 400 bp were eliminated from the analysis because they are outside the range of calibration based on the internal size standard described above. Additionally, FAM peaks below 21 bp and HEX peaks below 18 bp were also excluded because these values correspond to the lengths of the primers.

A number of graphical and statistical techniques were applied to evaluate the T-RFLP data, after it has been prepared following the above-listed steps. Electropherograms resulting from T-RFLP were first compared visually to evaluate bacterial diversity in the groundwater samples. We used Hierarchical cluster analysis (HCA) approach to evaluate similarity between T-RFLP profiles obtained from different groundwater samples. However, again, there is no commonly accepted distance measure to perform HCA with T-RFLP data and the

method selection should be guided by the complexity of the dataset (Culman *et al.*, 2008).

Some studies have used the Euclidean distance as a measure of sample dissimilarity (Dollhopf *et al.*, 2001; Blackwood *et al.*, 2003), whereas other studies have used different distance metrics such as the Common Area Index (van Bekkum *et al.*, 2006), Dice coefficient (Costa *et al.*, 2009; Nordentoft *et al.*, 2011), Sorenson's similarity index (Anderson *et al.*, 2010) or Bray-Curtis distance (Griebler *et al.*, 2010; Stein *et al.*, 2010; Baho *et al.*, 2012, Ibekwe *et al.*, 2012). We have analysed our data using two of the most commonly used distance metrics: Euclidean distance and Bray-Curtis distance. The distance values between each pair of samples were calculated based on standardized FAM and HEX peak heights. A dendrogram was constructed using Ward's method to display similarities between the samples (van Bekkum *et al.*, 2006). The dendrogram was used to identify groups of samples having similar T-RFLP profiles; the sample groups are hereafter referred to as "Bioclusters".

The Bioclusters were compared to groundwater chemistry (quantitative variables) using Box-and-Whisker plots and the Kruskal-Wallis test. The Bioclusters were also compared to categorical parameters: geographical region, groundwater mean residence time (MRT) class (Daughney *et al.*, 2010), well depth code, aquifer lithology, land use activities in the aquifer recharge zone and Hydrochemical categories described by Daughney and Reeves (2005), using cross-tabulation. The Chi-Square test was also performed to evaluate the statistical significance of these relationships. Further, permutational multivariate analysis of variance (PERMANOVA) and redundancy analysis were also performed to evaluate the relationships between bacterial community structures and groundwater chemistry and environmental factors (Korbel *et al.*, 2013). However, HCA in combination with Box-and-Whisker plots and the Kruskal-Wallis test, and cross-tabulation and the Chi-Square test were found to be most useful for interpretation of results from this study, and were therefore discussed in the remainder of this manuscript. Permutational analysis of variance

(PERMANOVA) and redundancy analysis did not provide additional insight and so were not discussed further. All statistical analyses were performed using the R (version 2.15.0) statistical programme.

Shannon-Wiener diversity indices (H') were calculated as $H' = - \sum \pi \ln \pi$, whereas π is the relative abundance of single T-RF in a given fingerprint (Hill *et al.*, 2003; Griebler *et al.*, 2010; Stein *et al.*, 2010). For our analysis, H' indices were calculated separately for FAM and HEX T-RFs which had been standardized relative to the sum of all peaks in a given profile. Mean H' indices were calculated with standard deviation for each “Biocluster”.

Results and discussion

Quality controls

Initial quality control checks provided satisfactory proof of principle for the experimental methods. The positive control, performed using *E. coli* DH5 α strain DNA, showed that the predicted T-RFLP results can be produced by the protocols used in this study. The DNA sequence obtained from *E. coli* DH5 α strain showed 100 % identity with the target region of *E. coli* 16S rRNA gene stored in the NCBI microbial databank. Actual T-RFLP profiles obtained for AluI, TaqI, MspI, AvaII, and MaeIII using the *E. coli* DH5 α strain PCR product were in general accord with the hypothetical fragment lengths determined by using the virtual restriction map constructed for the *E. coli* 16S reference sequence. Although all the above restriction endonucleases generally provided satisfactory results, AluI alone was used for this study because it is one of the most commonly used enzymes for T-RFLP investigations (Osborn *et al.*, 2006; Parkinson, 2004; van Bekkum *et al.*, 2006). The negative control

involved DNA extraction from sterile double distilled filtered water. The T-RFLP profile obtained with FAM and HEX fluorescent labels for the negative control contained no peaks greater than 200 Relative Fluorescent Unites (RFU). Thus, for this study, peaks with height less than 200 RFU were considered to be noise and were removed from all subsequent electropherograms regardless of their fragment sizes, instead of randomly assigning a RFU threshold to separate true peaks from noise (Osborn *et al.*, 2006). However, HCA was also performed using threshold values of 50 RFU and 100 RFU to check for significant alterations compared to the HCA performed with the 200 RFU threshold. This test showed that the results of HCA, including the basic clustering pattern and cluster composition, remained essentially unchanged with all three RFU thresholds (results not shown). Thus the 200 RFU threshold was applied to separate true peaks from noise for all subsequent analyses of the data.

DNA extraction and T-RFLP

Genomic DNA extractions from groundwater samples from 100 NGMP sites resulted in relatively low yields: 60 samples yielded less than 200 ng of DNA; 19 samples yielded between 200 and 500 ng of DNA; and only 7 samples yielded more than 1 µg of DNA (Table S1 and Fig. S1). This is an initial indication of the presence of low bacterial biomass in New Zealand groundwater compared with other environments such as soils (Parkinson, 2004) and acidic hydrothermal stream waters (Donachie *et al.*, 2002). The PCR amplifications resulted in a single fragment approximately 1400 bp in length from all groundwater DNA samples. Restriction digestion of representative products resulted in fragments shorter than 1400 bp, confirming effective enzymatic activity. The T-RFLP electropherograms showed a variety of peaks up to 400 bp by comparison with the internal size standard (Fig. S2).

Distribution and frequency of Bacterial Operational Taxonomic Units

The T-RFLP profiles obtained from the groundwater samples had different levels of complexity, as indicated by the number of peaks detected. The total number of FAM peaks in each profile ranged from 1 to 26 whereas the total number of HEX peaks ranged from 2 to 22. Here, profiles having one or two peaks, three to five peaks, or more than five peaks for each dye were separately classified as “simple”, “moderately complex” or “complex”, respectively (Fig. S3). The majority of groundwater samples produced complex T-RFLP profiles for each dye: 91 profiles for FAM and 83 profiles for HEX (Table 1). In addition, out of these complex profiles, 64 contained 10 or more FAM peaks and 44 profiles showed 10 or more HEX peaks (Fig. S4). However, restriction recognition sites must be present within the 400 bp region from both ends of a double labelled 1400 bp 16S PCR product for both a FAM and a HEX band to be produced by digestion. Otherwise only one peak will appear. This explains the difference in the number of samples belonging to each complexity level for FAM and HEX dyes (Table 1), and emphasizes the importance of considering FAM and HEX data together in numerical data analysis for a higher resolution of the technique. Nonetheless, these results indicate the presence of considerably high bacterial biodiversity in New Zealand groundwater compared with other highly diverse environments such as geothermal soils (Stott *et al.*, 2008) and acidic hydrothermal stream waters (Donachie *et al.*, 2002). However, as stated above, despite this biodiversity New Zealand groundwater appears to have relatively low biomass compared to other environments. This may imply the presence of many rare taxa with low relative abundance and highlights the importance of more powerful molecular techniques such as 454 pyrosequencing to explore these rare microbial communities.

Although FAM or HEX T-RFs may not accurately define a single taxon (see above), they will still provide a good representation of species richness in this sort of large scale

study, and hence they can be termed Operational Taxonomic Units (OTUs). A total of 148 unique FAM OTUs and 106 unique HEX OTUs were detected across all groundwater samples. However, only 38 FAM OTUs (25.7 % of total) were present in 10 or more groundwater samples. The FAM OTUs corresponding to fragment sizes of 198 and 28 bp occurred with the highest frequencies and were detected in 64 and 52 profiles, respectively (Fig. S2a). Similarly, only 31 HEX OTUs (29.2 % of total) were found in 10 or more profiles. The two HEX OTUs corresponding to fragment sizes 129 and 339 bp were detected in 78 and 48 profiles, respectively (Fig. S2b). Similar findings were obtained by Stein *et al.* 2010. These results again suggested that there is a considerable bacterial diversity across the country as majority of samples did not contain large number of common taxa. Bacterial diversity as expressed by Shannon-Wiener index (H') showed mean values of $H' = 1.81 \pm 0.59$ for FAM and $H' = 1.42 \pm 0.61$ for HEX. Although these values do not reflect extremely high diversity (Griebler *et al.*, 2010; Stein *et al.*, 2010; Zhou *et al.*, 2012), they still do provide a valuable insight in to bacterial diversity in New Zealand groundwater with respect to the scale of the study and limitations of the technique.

Relationships between bacterial communities and groundwater chemistry

The HCA was performed with Ward's method using two commonly used approaches: 1) peak heights were standardised relative to the highest peak and the Euclidean distance was used as the dissimilarity measure; 2) peak heights were standardized relative to the sum of all peaks in a given profile and the Bray-Curtis dissimilarity measure was used. The results revealed 84% similarity between the two approaches and suggested that the effect of standardisation method and similarity index is minimal on the clustering pattern for the

dataset collected in this study (Results not shown). Therefore, we used the above first approach for the purpose of presenting our results of this study.

The HCA demonstrated the relationship among groundwater bacterial communities as a cumulative measure of the presence and relative abundance of FAM and HEX OTUs (Fig. 2). At a high linkage distance threshold, the dendrogram revealed three main groups of samples (denoted as Bioclusters 1, 2 and 3; here and below all names are arbitrary). Increasing numbers of clusters were identified at lower linkage thresholds; eleven Bioclusters were identified at the lowest linkage distance threshold considered in this study. The number of samples belonging to each Biocluster at each threshold level is summarized in Table 2.

The Bioclusters at each threshold level were compared with median concentration values (mg L^{-1}) of 19 chemical compounds in groundwater (Table S2). The median values were derived from the actual values measured quarterly from March 2008 to March 2012 for each NGMP site. The chemical ions considered were: Na, K, Mg, Ca, HCO_3 , Cl, SO_4 , NO_3 -N, NH_4 -N, PO_4 -P, Fe, Mn, Br, F and SiO_2 . In addition, Bioclusters were compared with site-specific median values for Dissolved Oxygen (DO) in (mg L^{-1}), electrical conductivity (EC) in ($\mu\text{S cm}^{-1}$ at 25°C) and water temperature ($^\circ\text{C}$). To determine statistical significance, Kruskal-Wallis tests were performed with each of the above quantitative parameters at the five different linkage threshold levels. The results suggested that Bioclusters at the 11-cluster threshold were significantly associated (p value < 0.05) with dissolved oxygen, EC and all of the chemical parameters except K, SO_4 , NO_3 -N, PO_4 -P and Br. In addition, Bioclusters at the 4-, 5- and 7-cluster thresholds also showed clear relationships with Ca and HCO_3 . Further, at the 11-cluster threshold, SiO_2 , NH_4 -N and EC showed the strongest relationship (p value < 0.01) with Bioclusters, whereas Ca and HCO_3 reflected a similar feature at the 4- and 5-cluster thresholds (Table S3).

The Box-and-Whisker plots allowed us to distinguish qualitative aspects of the relationships between the Bioclusters and the hydrochemical variables. At the 3-cluster threshold, there is no significant association between Bioclusters and any of the parameters listed above. At this threshold, there are large differences in the parameter values for sites within a single Biocluster, such that systematic between-cluster parameter differences are not detectable. At the 4-cluster threshold, sites assigned to Bioclusters 2a vs. 2b are differentiated by the median concentrations of HCO_3 , Ca, Fe and Mn. It was particularly notable that Biocluster 2b contained sites with significantly higher Fe concentration than found in any of the other Bioclusters defined at this threshold (Fig. S5). At the 5-cluster threshold, Biocluster 2b1 includes sites with higher $\text{NO}_3\text{-N}$ and DO and lower F concentrations compared to Biocluster 2b2 (Fig. S6). At the 7-cluster threshold, Biocluster 3a includes sites with relatively low SiO_2 and Mg compared to Biocluster 3b (Fig. S7).

Relationships between Bioclusters and groundwater chemistry are highly visible at the 11-cluster threshold (Fig. 3 and Fig. S8 (i-iv)). For example, Biocluster 2b2b2 was associated with low $\text{NH}_4\text{-N}$, Fe and Mn and high $\text{NO}_3\text{-N}$ concentrations compared with Biocluster 2b2b1. This result suggests that Biocluster 2b2b2 was related to oxidized groundwater whereas Biocluster 2b2b1 was associated with reduced water (Daughney & Reeves, 2005), implying that the bacterial communities in these two clusters might be largely governed by groundwater redox chemistry. Further, high $\text{NO}_3\text{-N}$ concentrations in groundwater reflect the impacts of land use activities in aquifer recharge zones (Daughney and Reeves, 2005). Biocluster 3a showed considerable association with high $\text{NO}_3\text{-N}$ compared with Biocluster 3b, suggesting that the bacterial communities in the former cluster might be influenced by human activities in aquifer recharge zones.

In this study, we defined hydrochemical categories based on the hydrochemistry at each site using the method of Daughney and Reeves (2005), which provides a convenient

summary of the chemistry of groundwater in terms of redox potential and extent of human impact (Table 3). A highly significant relationship (p value < 0.01) between Bioclusters and the hydrochemical categories at almost all threshold levels was revealed by the Chi-Square test (Table S4). Cross-tabular representations between the Bioclusters and the hydrochemical categories indicated that Biocluster 2b2b1 was mainly comprised of samples having reduced groundwater whereas all or the majority of samples assigned to Bioclusters 2b1, 2b2b2, 2b2a, 2a3, 2a2, 2a1 and 1b were oxidized water. The degree of human impact is readily detectable only in oxidised groundwater because under reducing conditions $\text{NO}_3\text{-N}$, the predominant indicator of human impact, is removed via denitrification. Hence the ratio of samples assigned to hydrochemical categories 1A vs. 1B is an indicator of the proportion that are potentially influenced by human activities (Daughney & Reeves, 2005). On this basis Fig. 4a shows that most of the Bioclusters include at least some samples for which the hydrochemistry is indicative of human impact. The exceptions are Biocluster 1b, which contains a predominance of samples having oxidised water, none of which have evidence of human influence, and Bioclusters 2a1 and 2a2, for which a minority of samples showed hydrochemical indications of human impact. As discussed in the following section, these observed relationships between Bioclusters and hydrochemistry may indicate that human impact has altered the bacterial biodiversity of some New Zealand aquifers.

The Bioclusters were also related to mean residence time (MRT) of groundwater. The MRT at each NGMP site has been previously characterized using time series measurements of the age tracers tritium, chlorofluorocarbons and sulphur hexafluoride. For the purpose of this paper, we employ the four MRT classes defined by Daughney *et al.* (2010): <10 years; 11 to 40 years; 41 to 100 years; >100 years. Comparison between MRT classes and Bioclusters revealed that Biocluster 2b2b1 was mainly comprised of old groundwater whereas Bioclusters 1b and 2a1 contained mostly relatively young water (Fig. 4b).

Relationship between bacterial communities and geographical parameters

To evaluate the statistical significance of the relationships between Bioclusters and geographical parameters such as geographical region, aquifer lithology, land use activities in aquifer recharge zones and well depth, Chi-Square tests were performed with each parameter at the five different linkage thresholds. The geographical regions of New Zealand are displayed in Fig. 1 and Table S5 provides a summary of information on aquifer lithology and land use activities in the aquifer recharge zone for all NGMP sites. Median well depth across all NGMP sites is 26 m below ground level (b.g.l.), and the minimum, lower quartile, upper quartile and maximum well depths are 3, 10, 55 and 337 m b.g.l., respectively. For the purpose of this study, well depths were categorized by depth codes as follows: shallow (< 10 m), mid-depth (10-50 m) and deep (> 50 m). Bioclusters showed no significant association (p value > 0.05) to well depth, land use activities or aquifer lithology at any threshold.

Cross-tabular representations between Bioclusters at the 11-cluster threshold and the above-listed categorical parameters revealed some interesting qualitative aspects of some of the relationships. In relation to well depth code, Biocluster 2a2 did not occur in any shallow wells, whereas Bioclusters 1b and 2a1 did not occur in deep wells (Fig. 4c). Most of the Bioclusters included samples from a wide variety of aquifer lithologies. However, Bioclusters 1b and 2a1 were comprised only of samples from gravel and sand lithologies, Biocluster 2a3 included samples only from gravel lithologies, and Biocluster 2b1 contained samples only from basalt and limestone aquifers (Fig. 4d). Agricultural and horticultural land uses were associated with the majority of Bioclusters. However, Biocluster 2b2a contained no sites in agricultural settings but instead included sites with urban and horticultural land uses. Likewise, for Biocluster 2a2, half of the sites were situated in urban and industrial lands (Fig. 4e). Although the Chi-Square test showed that there is a significant relationship between

geographical region and the Bioclusters, graphical representation revealed that most of the Bioclusters consisted of samples from several regions, implying that there was no regional bias in classification of samples into the various Bioclusters (Fig. 4f). The only exception was found for Biocluster 2a1, for which the majority of samples were from the Wellington region and its remaining samples were from the Waikato region.

The results from this study may indicate that New Zealand's un-impacted aquifers have diagnostic natural (i.e. "baseline") bacterial communities, and that hydrochemical alteration due to human influence causes shifts in the bacterial composition of these groundwater ecosystems. It is notable that the majority of samples assigned to Bioclusters 1b and 2a1 were sourced from sand and gravel aquifers respectively, whereas some of the samples assigned to Biocluster 2a2 were sourced from volcanic lithologies such as ignimbrite and pumice. This may indicate that there are specific baseline bacterial communities that can be expected for certain aquifer lithologies in New Zealand, although additional work is required to test this hypothesis.

Bacterial diversity within each Biocluster at 11-Cluster threshold as expressed by Shannon-Wiener index (H') showed that there is considerable variation in diversity among clusters (Table 4). For FAM T-RFs, Biocluster 2a1 showed the highest mean value of $H' = 2.42 \pm 0.11$ and Biocluster 1a showed the lowest $H' = 1.12 \pm 0.62$. For HEX T-RFs, Bioclusters 2b1 reflected the highest $H' = 2.10 \pm 0.12$ and 1a showed the lowest $H' = 0.79 \pm 0.46$ (see Fig. 5). These index values may reflect another property of the Bioclusters, namely their species richness.

Table 1. Number of samples belonging to each classification level based on number of FAM and HEX peaks identified in T-RFLP electropherograms

	FAM	HEX
Simple	2	4
Moderately complex	7	13
Complex	91	83
Total	100	100

Table 2. Number of samples belonging to each Biocluster at different clustering levels

3-Cluster Level	Cluster	1	3			2						
	No. of Samples	21	26			53						
4-Cluster Level	Cluster	1	3		2a		2b					
	No. of Samples	21	26		19		34					
5-Cluster Level	Cluster	1	3		2a	2b1	2b2					
	No. of Samples	21	26		19	3	31					
7-Cluster Level	Cluster	1	3a	3b	2a	2b1	2b2a	2b2b				
	No. of Samples	21	12	14	19	3	6	25				
11-Cluster Level	Cluster	1a	1b	3a	3b	2a1	2a2	2a3	2b1	2b2a	2b2b1	2b2b2
	No. of Samples	16	5	12	14	5	6	8	3	6	10	15

Table 3. Typical chemical characteristics for Hydrochemical categories and subcategories defined by Daughney and Reeves (2005)

Hydrochemical Categories		Hydrochemical Subcategories	
1	Oxidized groundwater [NO ₃ -N] above DL [*]	1A	Impacted by human activity, [NO ₃ -N] above 3.5 mg/L
	[NH ₄ -N], [Fe], [Mn] near or below DL [*]	1B	Little impacted by human activity, [NO ₃ -N] below 3.5 mg/L
2	Reduced groundwater [NO ₃ -N] near or below DL [*]	2A	Moderately reduced, [SO ₄] above DL [*]
	[NH ₄ -N], [Fe], [Mn] above DL [*]	2B	Highly reduced, [SO ₄] near or below DL [*]

^{*} DL refers to the analytical detection limit.

Table 4. Summary of Shannon-Wiener diversity indices (H') in each Biocluster at 11-cluster threshold calculated using FAM and HEX T-RFs separately.

Cluster	FAM		HEX	
	Mean H'	SD	Mean H'	SD
1a	1.12	0.62	0.79	0.46
1b	2.18	0.21	1.86	0.32
2a1	2.42	0.11	1.92	0.42
2a2	1.84	0.50	0.80	0.39
2a3	2.18	0.41	1.76	0.47
2b1	2.07	0.06	2.10	0.12
2b2a	1.84	0.49	1.71	0.46
2b2b1	1.74	0.69	1.59	0.58
2b2b2	1.83	0.68	1.23	0.63
3a	1.86	0.34	1.45	0.44
3b	1.91	0.32	1.68	0.45

Table 5. Summary of groundwater features in Bioclusters at 11-cluster threshold

Bioclusters at 11-cluster threshold	Groundwater characteristics
1a	Oxidized human impacted water, shallow, mid-depth and deep wells
1b	Mainly oxidized water with less human impact, relatively young groundwater, shallow and mid-depth wells, only gravel and sand aquifers
2a1	Mainly oxidized water, majority from Wellington region, relatively young groundwater, shallow and mid-depth wells, only gravel and sand aquifers, low NO ₃ -N, low [DO]
2a2	Only oxidized water, mid-depth and deep wells, majority urban and industrial land use, moderate NO ₃ -N, high [DO]
2a3	Mainly oxidized water, only gravel aquifers, high NO ₃ -N, moderate [DO]
2b1	Only oxidized water, only basalt and limestone aquifers
2b2a	Mainly oxidized water, no agricultural land use
2b2b1	Mainly reduced, old groundwater
2b2b2	Mainly oxidized water
3a	High NO ₃ -N, low SO ₄ , low SiO ₂ , forestry land use
3b	High SO ₄ , high SiO ₂ , low NO ₃ -N

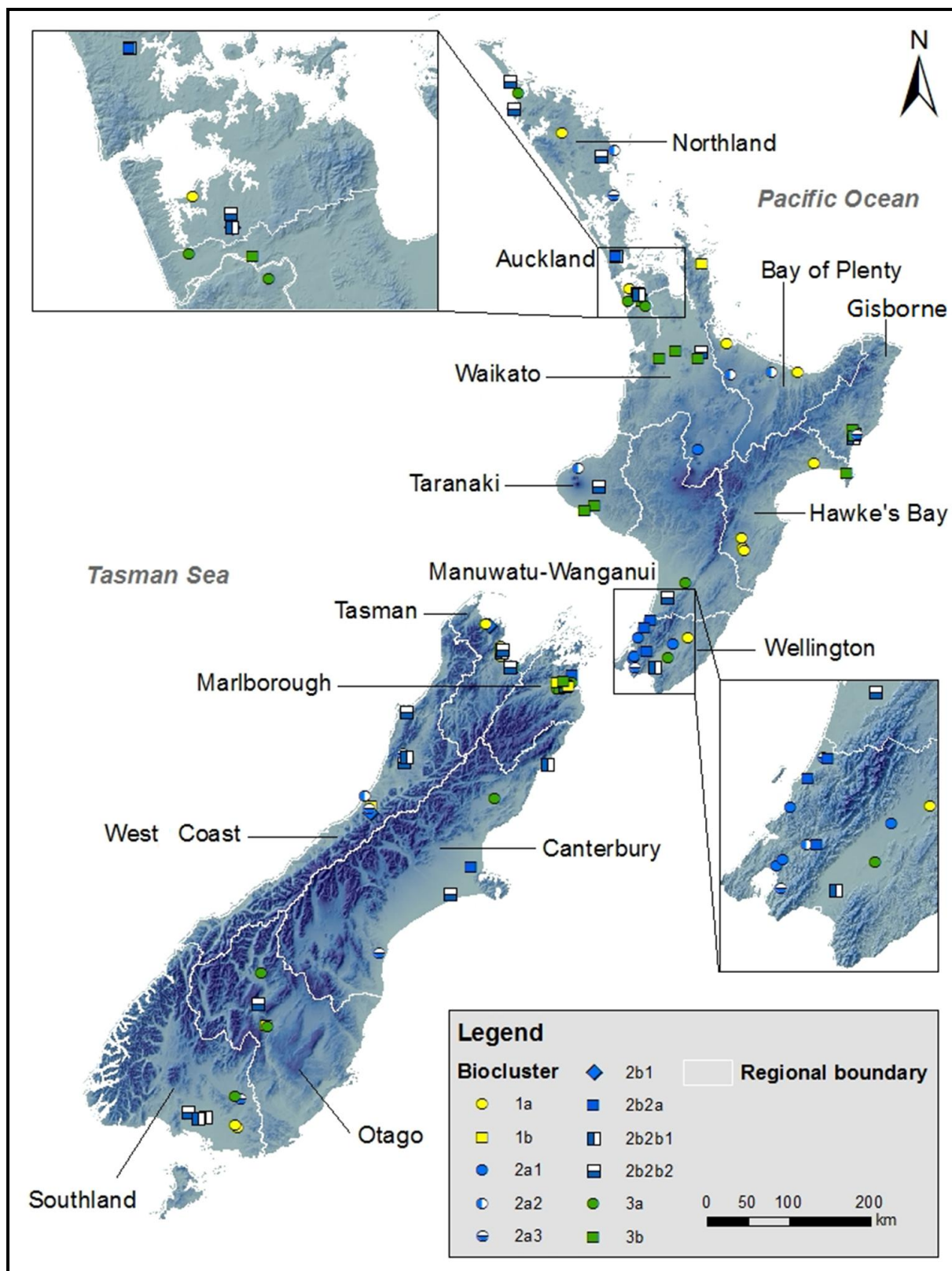


Fig. 1. Groundwater sampling sites across New Zealand. The boundaries of the 15 regional authorities are also shown.

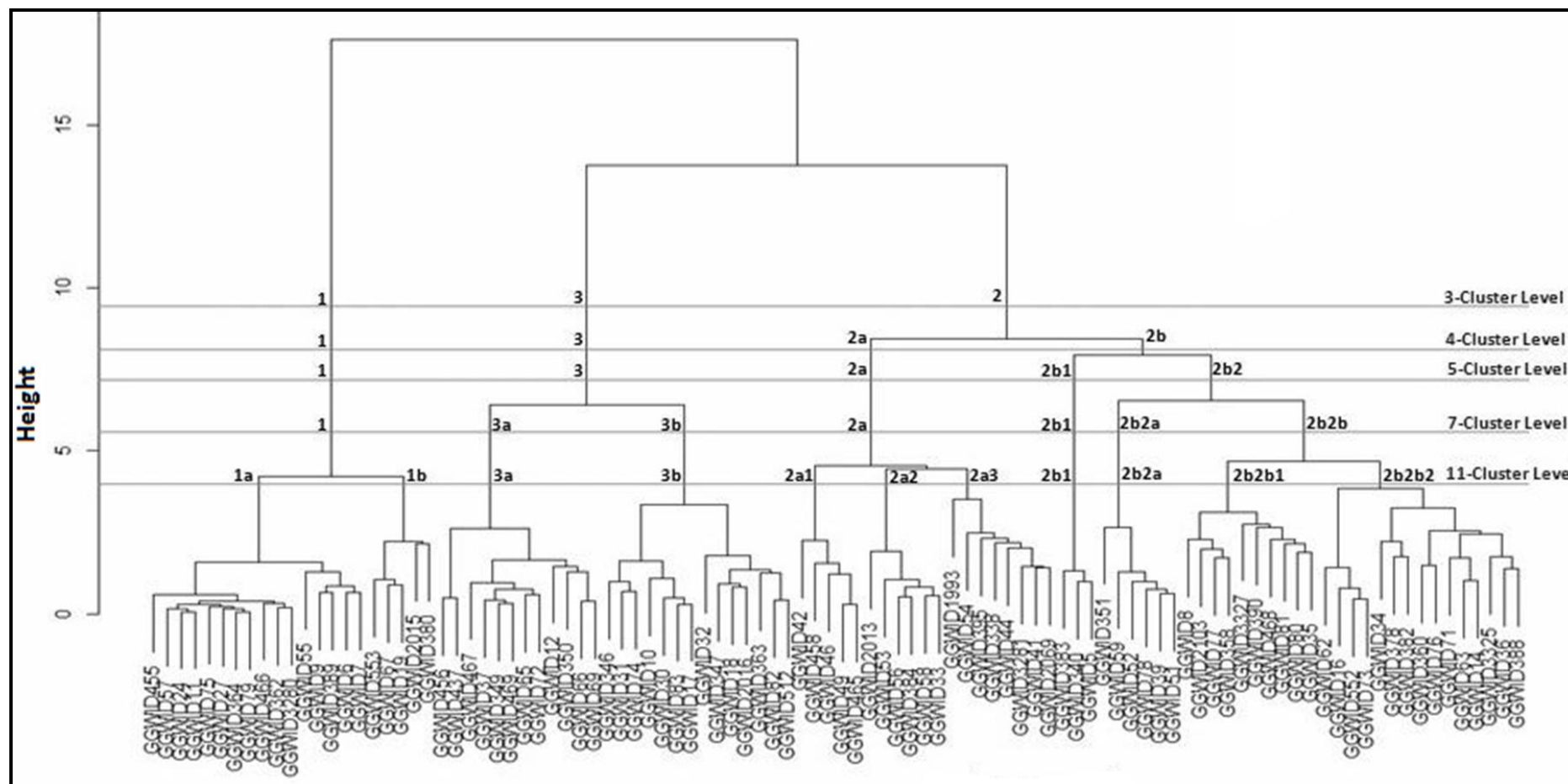


Fig. 2. Dendrogram produced by hierarchical cluster analysis conducted using FAM and HEX labelled terminal fragments. Clustering was performed using Ward's linkage rule and the square of the Euclidean distance as the separation measure.

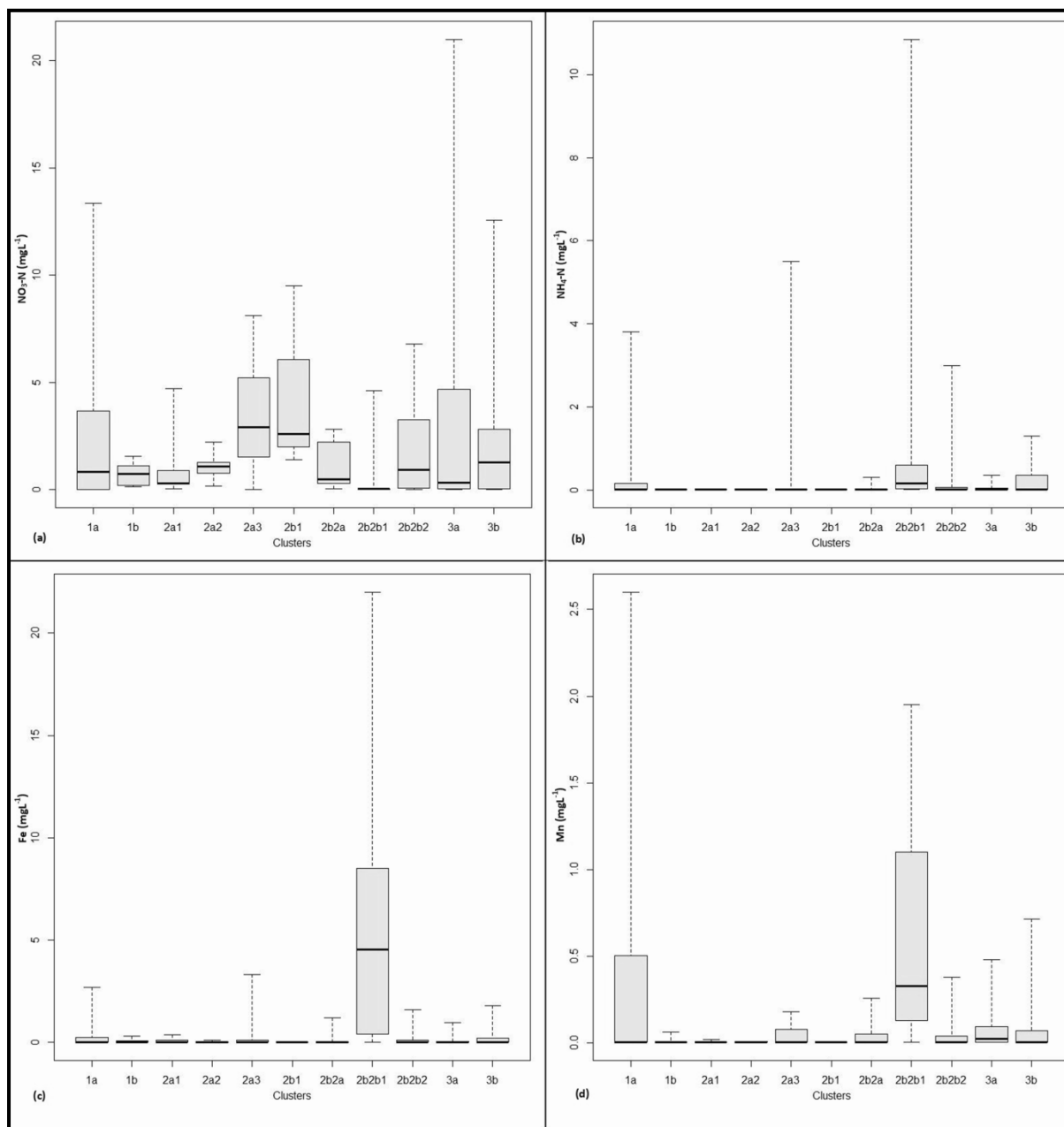


Fig. 3. Box-and-Whisker Plot of median concentrations of $\text{NO}_3\text{-N}$ (a), $\text{NH}_4\text{-N}$ (b), Fe (c) and Mn (d) across Bioclusters defined at the 11-cluster threshold.

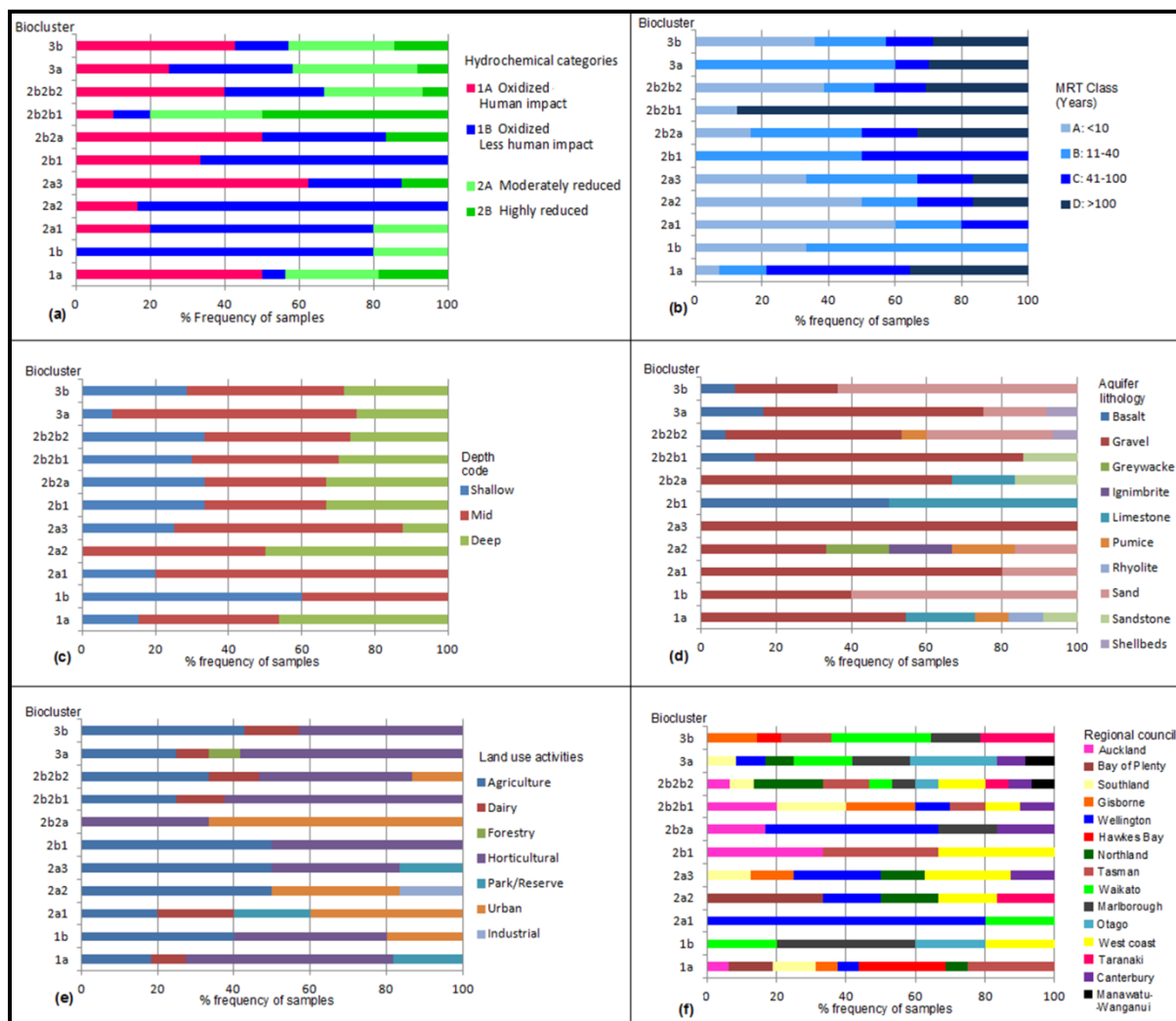


Fig. 4. Percentage frequency distribution of samples with hydrochemical categories (a), MRT Classes (b), aquifer well depth (c), aquifer lithology (d), land use activities of aquifer recharge zone (e) and regional council (f).

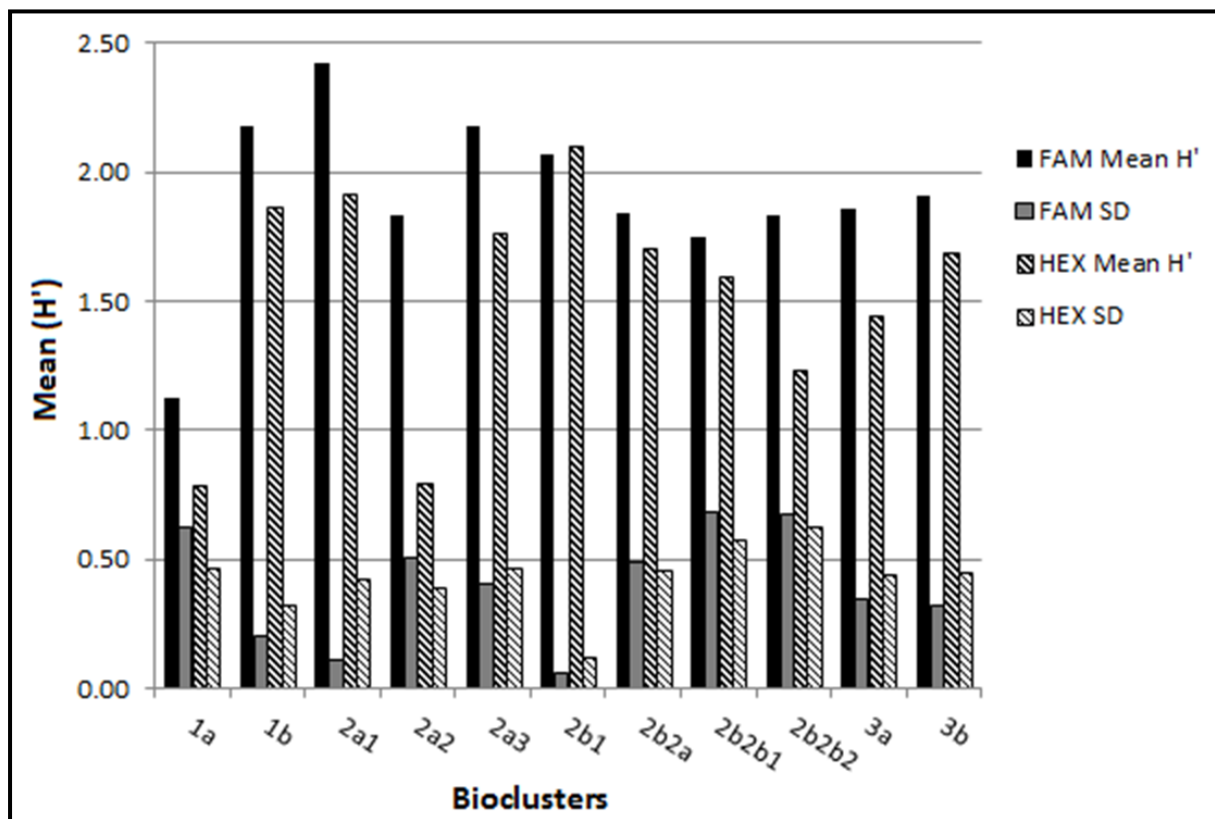


Fig. 5. Summary of mean Shannon-Wiener diversity indices (H') values for each Biocluster using FAM and HEX T-RFs.

Concluding remarks

The results of this study demonstrated that groundwater bacterial diversity was related to hydrochemistry, with geological factors and human activities as important secondary controls. Table 5 summarises the groundwater features related to the different Bioclusters at the 11-cluster threshold.

Previous studies have shown that the bacterial community structure of the liquid groundwater can be different from that of the aquifer itself, and that the latter may influence groundwater chemistry (Alfreider *et al.*, 1997; Flynn *et al.*, 2008; Griebler & Lueders, 2009). However, the main focus of this study was state-of-the-environment monitoring of groundwater quality. Therefore, we did not analyse aquifer materials directly, but instead focussed on the groundwater itself. Still, the identifiable relationships between the Bioclusters and groundwater chemistry implied that groundwater bacterial diversity can be comparable to that of the aquifer materials. However, further studies are needed to evaluate the actual relationships between these two bacterial communities in New Zealand aquifers.

Aquifer confinement could also influence the bacterial diversity by altering the groundwater chemistry. However, in our study, the Bioclusters were not compared with aquifer confinement categories, which could be used as a secondary indicator of groundwater chemistry as the direct chemical data were readily available for the analysis. Further, it is evident that seasonality may also strongly influence bacterial diversity (Zhou *et al.*, 2012), but we did not analyse this sort of variation.

The T-RFLP technique was highly effective in this study, as in previous investigations where the main objective was to understand the bacterial community structure quickly and cost-effectively (Flynn *et al.*, 2008; Luna *et al.*, 2009; Flynn *et al.*, 2012). Although, the technique is considered to be comparable with even high throughput

sequencing technologies, T-RFLP also has its own drawbacks as with any other molecular tool (Nordentoft *et al.*, 2011; Pilloni *et al.*, 2012). The DNA based fingerprinting methods including T-RFLP only assess the potential bacterial diversity, but not the viable community structure. However, our results do indicate that the T-RFLP technique more or less reflects the viable bacterial communities in groundwater because the Bioclusters showed strong relationships with chemistry (Sheridan *et al.*, 1998). This work provides a basic framework for the direction of future studies to understand the viable bacterial community structures with mRNA and protein based approaches. Although culture-independent molecular techniques are highly regarded as a superior approach to capture total microbial diversity in environmental samples during the recent past, this approach is also encountered with invisible challenges such as extracting total DNA from all species in samples, providing optimal experimental conditions suitable for diverse range of taxa and identifying novel microorganisms from databases which may not contain information on all the species (Donachie *et al.*, 2007). Therefore, we may not be able to identify the total bacterial diversity in groundwater even with molecular approaches including T-RFLP, and culture-dependant approaches might be able to detect this undiscovered diversity up to a certain extent. In addition, the resolution of the technique might not be powerful enough to capture very low abundance bacterial components in environmental samples (Pilloni *et al.*, 2012). Therefore, the actual bacterial diversity could be greater than the findings of the current method. Further, the T-RFLP technique does not provide names or any functional information about the microorganisms detected and there is a possibility that the same T-RF may be returned by closely related, yet different, taxa with divergent metabolic activities. Therefore, it is crucial to take into account these limitations when interpreting the results of the study.

In microbial ecology studies, it is desirable to assess the variability contributed to the results by replicate sampling appropriate for the objective and scale of study (Prosser, 2010).

As our aim was to provide a comprehensive overview of the bacterial community structure in groundwater across the country, we did not replicate sampling at local scale. A pilot study conducted by van Bekkum *et al.* (2006) using T-RFLP showed that the temporal variation of groundwater bacterial diversity was minimal. Therefore, we analysed a single groundwater sample from each location assuming that our sampling design provides strong replication of environmental factors – several samples were collected from sites with similar chemistries and geological factors, but which were distinct from each other. Accordingly, the results showed that the sampling design was highly effective for our objective because it showed relationships between Bioclusters and hydrochemical categories, which comprised distinct sites with similar chemistries. To the best of our knowledge, this is the first study to survey the bacterial diversity in groundwater in New Zealand using molecular techniques and is probably the first in the world to evaluate the groundwater bacterial diversity across an entire country. The results of this study provided a strong platform for the current metagenomics and genomic studies aiming to explore the unseen rare microbial fraction and to test hypotheses related to bacterial diversity and other chemical, physical and environmental factors of groundwater using advanced molecular tools such as high throughput DNA sequencing (Chapter 3.3).

Acknowledgments

The authors would like to thank all the groundwater research staff members of 15 regional councils for their valuable support in sample collection, and would also like to thank Professor Craig Cary, Waikato University, Hamilton, New Zealand and Dr. Els Mass, National Institute of Water and Atmospheric Research, Wellington, New Zealand for their useful advice regarding T-RFLP data analyses. Further, authors would like to thank the editor and the two anonymous reviewers for their constructive critiques which helped to improve

the quality of the manuscript and Dr. Dalice Sim, Victoria University of Wellington, New Zealand for her valuable advice to improve the quality of data analysis. This project was financially supported by public research funding from the Government of New Zealand.

The authors have no conflicts of interest.

References

- Alfreider A, Krossbacher M & Psenner R (1997) Groundwater samples do not reflect bacterial densities and activity in subsurface systems. *Water Res* **31**: 832-840.
- Anderson SA, Northcote PT & Page MJ (2010) Spatial and temporal variability of the bacterial community in different chemotypes of the New Zealand marine sponge *Mycale hentscheli*. *FEMS Microbiol Ecol* **72**: 328–342.
- Baho DL, Peter H & Tranvik LJ (2012) Resistance and resilience of microbial communities--temporal and spatial insurance against perturbations. *Environ Microbiol* **14**: 2283–2292.
- Bethke CM, Ding L, Jin Q & Sanford RA (2008) Origin of microbiological zoning in groundwater flows. *Geology* **36**: 739–742.
- Blackwood CB, Marsh T, Kim SH & Paul EA (2003) Terminal restriction fragment length polymorphism data analysis for quantitative comparison of microbial communities. *Appl Environ Microbiol* **69**: 926–932.
- Camarinha-Silva A, Wos-Oxley ML, Jáuregui R, Becker K & Pieper DH (2012) Validating T-RFLP as a sensitive and high-throughput approach to assess bacterial diversity patterns in human anterior nares. *FEMS Microbiol Ecol* **79**: 98–108.
- Chapelle FH (2000) The significance of microbial processes in hydrogeology and geochemistry. *Hydrogeol J* **8**: 41–46.
- Costa E, Puhl NJ, Selinger LB & Inglis GD (2009) Characterization of mucosa-associated bacterial communities of the mouse intestine by terminal restriction fragment length polymorphism : Utility of sampling strategies and methods to reduce single-stranded DNA artifacts. *J Microbiol Methods* **78**: 175–180.

- Culman SW, Gauch HG, Blackwood CB & Thies JE (2008) Analysis of T-RFLP data using analysis of variance and ordination methods: a comparative study. *J Microbiol Methods* **75**: 55–63.
- Daughney CJ & Reeves RR (2005) Definition of hydrochemical facies in the New Zealand groundwater monitoring programme. *J Hydrol NZ* **44**: 105-130.
- Daughney CJ & Reeves RR (2006) Analysis of temporal trends in New Zealand's groundwater quality based on data from the National Groundwater Monitoring Programme. *J Hydrol NZ* **45**: 41-62.
- Daughney CJ, Morgenstern U, van der Raaij R & Reeves RR (2010) Discriminant analysis for estimation of groundwater age from hydrochemistry and well construction: application to New Zealand aquifers. *Hydrogeol J* **18**: 417-428.
- Daughney CJ, Raiber M, Moreau-Fournier M, Morgenstern U & van der Raaij R (2012) Use of hierarchical cluster analysis to assess the representativeness of a baseline groundwater quality monitoring network: Comparison of New Zealand's national and regional groundwater monitoring programs. *Hydrogeol J* **20**: 185-200.
- Daughney CJ, Jones A, Baker T, Hanson C, Davidson P, Zemansky GM, Reeves RR & Thompson M (2006) *A national protocol for state of the environment groundwater sampling in New Zealand: ME 781*. Ministry for the Environment, Wellington, New Zealand.
- Dollhopf SL, Hashsham SA & Tiedje TM (2001) Interpreting 16S rDNA T-RFLP data: Application of self-organizing maps and principal component analysis to describe community dynamics and convergence. *Microb Ecol* **42**: 495–505.
- Donachie SP, Foster JS & Brown MV (2007) Culture clash: challenging the dogma of microbial diversity. *ISME J* **7**: 97-102.

- Donachie SP, Christenson BW, Kunkel DD, Malahoff A & Alam M (2002) Microbial community in acidic hydrothermal waters of volcanically active White Island, New Zealand. *Extremophiles* **6**: 419-25.
- Edlund A, Soule T, Sjöling S & Jansson JK (2006) Microbial community structure in polluted Baltic Sea sediments. *Environ Microbiol* **8**: 223-232.
- Falkowski PG, Fenchel T & Delong EF (2008) Microbial engines that derived earth's biogeochemical cycles. *Science* **320**: 1034-1038.
- Flynn TM, Sanford RA & Bethke CM (2008) Attached and suspended microbial communities in a pristine confined aquifer. *Water Resour Res* **44**: W07425.
- Flynn TM, Sanford RA, Santo Domingo JW, Ashbolt NJ, Levine AD & Bethke CM (2012) The active bacterial community in a pristine confined aquifer. *Water Resour Res* **48**: W09510.
- Ghiorse CW (1997) Subterranean life. *Science* **275**: 789-790.
- Griebler C & Lueders T (2009) Microbial biodiversity in groundwater ecosystems. *Freshw Biol* **54**: 649–677.
- Griebler C, Stein H, Kellermann C, Berkhoff S, Brielmann H, Schmidt S, Selesi D, Steube C, Fuchs A & Hahn HJ (2010) Ecological assessment of groundwater ecosystems – vision or illusion? *Ecol Eng* **36**: 1174–1190.
- Haack SK, Fogarty LR, West TG, Alm EW, McGuire JT, Long DT, Hyndman DW & Forney LJ (2004) Spatial and temporal changes in microbial community structure associated with recharge-influenced chemical gradients in a contaminated aquifer. *Environ Microbiol* **6**: 438-48.
- Hedrich S, Schlömann M & Johnson DB (2011) The iron-oxidizing proteobacteria. *Microbiology+* **157**: 64.

- Hill TCJ, Walsh KA, Harris JA & Moffett BF (2003) Using ecological diversity measures with bacterial communities. *FEMS Microbiol Ecol* **43**: 1–11.
- Ibekwe AM, Leddy MB, Bold RM & Graves AK (2012) Bacterial community composition in low-flowing river water with different sources of pollutants. *FEMS Microbiol Ecol* **79**: 155–66.
- Janssen PH, Yates PS, Grinton BE, Taylor PM & Sait M (2002) Improved culturability of soil bacteria and isolation in pure culture of novel members of the divisions Acidobacteria, Actinobacteria, Proteobacteria, and Verrucomicrobia. *Appl Environ Microbiol* **68**: 2391–2396.
- Korbel KL & Hose GC (2011) A tiered framework for assessing groundwater ecosystem health. *Hydrobiologia* 661:329–349.
- Korbel KL, Hancock PJ, Serov P, Lim RP & Hose GC (2013) Groundwater ecosystems vary with land use across a mixed agricultural landscape. *J Environ Qual* **42**: 380-390.
- Larned ST (2012) Phreatic groundwater ecosystems: research frontiers for freshwater ecology. *Freshw Biol* **57**: 885–906.
- Lozupone CA & Knight R (2007) Global patterns in bacterial diversity. *P Nat Acad Sci USA* **104**: 11436-11440.
- Luna GM, Anno AD, Corinaldesi C, Armeni M & Danovaro R (2009) Diversity and spatial distribution of metal-reducing bacterial assemblages in groundwaters of different redox conditions. *Int Microbiol* **12**:153-159.
- Maier MR, Pepper IL & Gerba CP (2009) *Environmental Microbiology (2nd Edition)*. Academic Press, San Diego, California.
- Ministry for the Environment (2010) *Faecal pollution (bacteria) in groundwater*. Ministry for the Environment, Wellington, New Zealand.

www.mfe.govt.nz/publications/water/national-protocol-groundwater-dec06/index.html Accessed March 2013.

- Morgenstern U & Daughney CJ (2012) Groundwater age for identification of baseline groundwater quality and the impacts of land-use intensification - The National Groundwater Monitoring Programme of New Zealand. *J Hydrol* **456/457**: 79-93.
- Neufeld JD & Mohn WW (2005) Unexpectedly high bacterial diversity in arctic tundra relative to boreal forest soils, revealed by serial analysis of ribosomal sequence tags. *Appl Environ Microbiol* **71**: 5710–5718.
- Nordentoft S, Mølbak L, Bjerrum L, De Vyllder J, Van Immerseel F & Pedersen K (2011) The influence of the cage system and colonisation of *Salmonella Enteritidis* on the microbial gut flora of laying hens studied by T-RFLP and 454 pyrosequencing. *BMC Microbiol* **11**: 187
- Osborn AM, Moore ER & Timmis KN (2000) An evaluation of terminal-restriction fragment length polymorphism (T-RFLP) analysis for the study of microbial community structure and dynamics. *Environ Microbiol* **2**: 39-50.
- Parkinson R (2004) Forensic DNA profiling of bacterial communities in soil. MSc thesis, Victoria University of Wellington, Wellington, New Zealand.
- Parkinson R (2009) Bacterial communities associated with human decomposition. PhD thesis, Victoria University of Wellington, Wellington, New Zealand.
- Pilloni G, Granitsiotis MS, Engel M & Lueders T (2012) Testing the limits of 454 pyrotag sequencing: reproducibility, quantitative assessment and comparison to T-RFLP fingerprinting of aquifer microbes. *PLoS One* **7**: e40467.
- Prosser JI (2010) Replicate or lie. *Environ Microbiol* **12**: 1806-1810.

- Rosen MR (2001) Hydrochemistry of New Zealand's aquifers. *Groundwaters of New Zealand*, (Rosen MR & White PA, eds), pp. 77-110. New Zealand Hydrological Society, Wellington, New Zealand.
- Sheridan GEC, Masters CI, Shallcross JA & Mackey BM (1998) Detection of mRNA by reverse transcription-PCR as an indicator of viability in *Escherichia coli* cells. *Appl Environ Microbiol* **64**: 1313–1318.
- Sinreich M, Pronk M & Kozel R (2011) Microbiological spring water quality monitoring across Switzerland. Proc H2Karst, 9th Conference on Limestone Hydrogeology, Besançon (France) 1-3 Sep. 2011: 447-450.
- Stein H, Kellermann C, Schmidt SI, Brielmann H, Steube C, Berkhoff SE, Fuchs A, Hahn HJ, Thulin B & Griebler C (2010) The potential use of fauna and bacteria as ecological indicators for the assessment of groundwater quality. *J Environ Monitor* **12**: 242–254.
- Steube C, Richter S & Griebler C (2009) First attempts towards an integrative concept for the ecological assessment of groundwater ecosystems. *Hydrogeol J* **17**: 23–35.
- Stott MB, Crowe MA, Mountain BW, Smirnova AV, Hou S, Alam M & Dunfield PF (2008) Isolation of novel bacteria, including a candidate division, from geothermal soils in New Zealand. *Environ Microbiol* **10**: 2030-2041.
- Struckmeier W, Rubin Y & Jones JAA (2005) *Groundwater - reservoir for a thirsty planet?: Earth Sciences for Society; a Prospectus for a Key Theme of the International Year of Planet Earth*. IUGS International Union of Geological Sciences Secretariat, Geological Survey of Norway, Trondheim, Norway.
- van Bekkum M, Sainsbury JP, Daughney CJ & Chambers GK (2006) Molecular analysis of bacterial communities in groundwaters from selected wells in the Hutt Valley and the Wairarapa, New Zealand. *New Zeal J Mar Fresh* **40**: 91–106.

Younger PL (2007) *Groundwater in the environment: an introduction*. Blackwell Publishing, Malden, Massachusetts, United States.

Zhou Y, Kellermann C & Griebler C (2012) Spatio-temporal patterns of microbial communities in a hydrologically dynamic pristine aquifer. *FEMS Microbiol Ecol* **81**: 230–242.

Zhou J, Davey ME, Figueras JB, Rivkina E, Gilichinsky D & Tiedje JM (1997) Phylogenetic diversity of a bacterial community determined from Siberian tundra soil DNA. *Microbiology* **143**: 3913-3919.

Supplementary information

Table S1. Summary of genomic DNA yields obtained from two litres of groundwater from each sample. GGW ID refers to the unique site identification numbers used in the Geothermal and Groundwater Database (<http://ggw.gns.cri.nz/ggwdata/mainPage.jsp>).

GGW ID	gDNA (ng)	GGW ID	gDNA (ng)	GGW ID	gDNA (ng)
GGW ID 10	78.73	GGW ID 358	109.16	GGW ID 512	170.19
GGW ID 11	50.77	GGW ID 36	83.48	GGW ID 52	111.57
GGW ID 12	174.94	GGW ID 360	815.32	GGW ID 53	131.84
GGW ID 14	529.85	GGW ID 362	27.84	GGW ID 54	423.65
GGW ID 16	1885.64	GGW ID 363	374.80	GGW ID 55	145.23
GGW ID 17	599.58	GGW ID 364	56.58	GGW ID 552	27.74
GGW ID 18	934.26	GGW ID 37	118.14	GGW ID 553	29.72
GGW ID 19	73.85	GGW ID 378	65.04	GGW ID 56	182.33
GGW ID 1993	82.89	GGW ID 380	122.07	GGW ID 57	502.07
GGW ID 2013	305.41	GGW ID 382	87.41	GGW ID 58	238.57
GGW ID 2015	29.24	GGW ID 383	60.12	GGW ID 59	101.93
GGW ID 2016	438.67	GGW ID 384	289.48	GGW ID 6	62.49
GGW ID 2069	85.12	GGW ID 388	85.09	GGW ID 62	490.39
GGW ID 2103	150.02	GGW ID 389	159.02	GGW ID 63	165.86
GGW ID 24	271.69	GGW ID 39	324.25	GGW ID 65	475.70
GGW ID 27	91.81	GGW ID 390	59.91	GGW ID 66	152.29
GGW ID 30	7560.48	GGW ID 395	152.14	GGW ID 67	65.88
GGW ID 31	1664.38	GGW ID 41	64.94	GGW ID 69	62.80
GGW ID 32	175.82	GGW ID 42	168.45	GGW ID 7	105.81
GGW ID 3280	763.79	GGW ID 437	615.56	GGW ID 71	221.49
GGW ID 3281	112.68	GGW ID 44	1138.77	GGW ID 72	78.34
GGW ID 33	71.10	GGW ID 455	93.12	GGW ID 73	236.76
GGW ID 3325/4	74.96	GGW ID 456	73.72	GGW ID 74	62.29
GGW ID 3327	128.20	GGW ID 458	244.18	GGW ID 75	457.83
GGW ID 338	63.08	GGW ID 46	117.56	GGW ID 76	566.87
GGW ID 34	1349.54	GGW ID 465	25.66	GGW ID 77	93.91
GGW ID 340	88.77	GGW ID 466	605.97	GGW ID 78	216.98
GGW ID 346	1315.75	GGW ID 467	223.84	GGW ID 79	886.75
GGW ID 347	524.46	GGW ID 468	95.29	GGW ID 8	65.44
GGW ID 349	696.90	GGW ID 469	141.50	GGW ID 80	26.21
GGW ID 35	1425.22	GGW ID 49	75.10	GGW ID 81	89.83
GGW ID 350	943.64	GGW ID 5	254.40	GGW ID 82	88.94
GGW ID 351	592.08	GGW ID 51	400.77	GGW ID 83	474.22
				GGW ID 9	99.17

CHAPTER 3.1

Table S2. Median values of 15 chemical parameters and 4 physical parameters derived from the actual values measured quarterly from March 2008 to March 2012 across the NGMP sites. Units in mg L⁻¹ for all variables except pH which is in pH units, Electrical conductivity (EC) in (μS cm⁻¹ at 25 °C) and Temperature in (°C). ND indicates that the parameter value was not determined.

GGW ID	Na	K	Mg	Mn	Fe	Ca	SO ₄	SiO ₃	PO ₄	NO ₃ -N	F	Cl	Br	HCO ₃	NH ₄ -N	pH	DO	EC	Temp.
5	4.80	0.69	3.10	<0.005	<0.02	47.00	4.60	10.00	0.01	2.60	0.04	6.20	0.01	149.00	<0.01	7.89	9.06	286.00	18.00
6	24.00	0.53	8.10	0.02	0.02	33.00	1.80	26.00	0.03	0.02	0.20	5.60	<0.04	196.00	0.02	8.20	1.53	311.50	19.00
7	66.00	4.65	9.10	<0.005	<0.02	67.00	18.70	6.70	0.00	0.44	0.03	112.00	0.23	217.00	<0.01	7.91	6.19	685.00	18.00
8	19.70	0.96	6.60	0.32	3.10	16.20	2.70	65.00	0.11	<0.03	0.38	4.70	<0.04	128.00	0.04	7.38	4.27	211.00	20.00
9	9.50	0.63	42.00	<0.005	<0.02	9.10	33.00	36.00	0.04	13.35	0.06	16.10	0.00	133.00	0.02	7.74	7.59	421.00	20.00
10	10.10	0.61	30.00	<0.005	<0.02	19.50	31.50	30.00	0.03	12.55	0.03	16.70	0.06	112.00	0.02	7.78	6.70	386.00	18.00
11	6.10	0.98	13.85	2.60	1.20	15.60	26.00	25.00	0.01	0.00	0.06	8.80	0.05	93.00	0.07	6.85	3.59	225.50	18.00
12	13.20	2.20	6.50	0.04	0.04	35.00	9.10	44.00	0.10	<0.03	0.15	16.80	0.04	138.00	0.35	8.13	0.34	285.00	17.50
14	23.00	7.25	8.45	0.38	1.60	24.00	0.05	51.00	1.00	0.00	0.25	15.70	0.05	172.00	3.00	7.64	4.66	318.00	17.00
16	11.20	4.50	3.00	<0.005	<0.02	5.90	7.80	81.00	0.07	3.60	0.04	7.40	0.04	36.00	<0.01	6.75	7.93	127.00	18.50
17	15.90	1.80	10.60	<0.005	<0.02	12.20	2.60	52.00	0.08	2.60	0.06	21.00	0.05	87.00	<0.01	7.58	5.77	225.50	18.50
18	14.80	3.40	5.40	<0.005	<0.02	10.80	13.50	33.00	0.02	7.00	0.07	11.80	0.04	46.00	<0.01	6.96	8.80	196.50	18.50
19	6.35	1.15	1.30	<0.005	<0.02	26.00	6.90	10.45	0.05	1.55	0.07	6.20	<0.1	79.00	<0.01	8.21	7.66	190.00	17.50
24	13.45	1.20	7.60	0.61	<0.02	11.90	7.30	23.50	0.01	4.40	0.23	12.10	0.00	66.00	<0.01	6.90	0.16	198.00	19.00
27	33.00	1.50	4.40	0.22	0.03	15.00	1.50	20.00	0.11	0.00	0.26	9.50	0.04	149.00	0.29	7.96	0.43	257.00	18.50
30	23.00	2.40	12.00	0.01	0.03	31.00	5.15	26.00	0.07	0.01	0.07	23.00	0.01	170.00	0.49	8.24	ND	335.50	19.00
31	30.00	6.70	7.30	0.72	1.80	17.30	0.03	53.00	0.38	<0.03	0.34	26.00	0.10	142.50	1.30	7.72	ND	309.00	19.00
32	51.00	5.00	11.15	<0.005	<0.02	16.70	49.00	51.00	0.05	2.50	0.14	69.50	0.25	60.00	0.01	6.90	ND	486.00	18.00
33	8.65	2.45	2.85	<0.005	<0.02	7.40	6.75	21.00	0.02	0.92	0.02	10.40	0.05	36.00	<0.01	6.34	ND	121.00	18.00
34	21.00	1.90	9.20	0.01	0.10	28.50	0.04	19.80	0.02	<0.03	0.17	9.30	0.04	182.50	0.72	8.10	ND	304.00	18.50
35	26.00	0.95	9.25	0.57	1.80	36.00	0.04	33.00	0.98	0.00	0.32	4.35	<0.1	220.00	0.28	7.60	1.00	338.00	19.00
36	9.10	1.10	4.50	<0.005	<0.02	26.00	11.40	13.05	<0.004	5.10	0.05	9.50	<0.04	75.00	<0.01	7.04	8.65	216.50	17.50

CHAPTER 3.1

37	10.20	1.10	3.60	<0.005	<0.02	37.00	10.40	15.20	0.01	6.60	0.09	4.00	<0.1	111.50	<0.01	8.01	8.90	256.00	18.00
39	8.60	0.73	3.10	<0.005	<0.02	14.20	3.00	18.50	0.01	0.28	0.12	5.10	<0.04	70.00	<0.01	7.94	6.00	136.00	18.00
41	21.00	1.60	9.00	<0.005	<0.02	60.00	19.90	16.10	0.01	3.90	0.12	30.00	0.02	189.50	<0.01	7.38	4.40	445.00	18.00
42	23.00	2.80	6.10	0.02	0.36	16.10	19.00	22.00	0.02	0.04	0.14	36.00	0.10	57.00	0.03	6.89	0.15	268.00	17.00
44	22.00	1.30	6.20	<0.005	<0.02	10.50	16.30	22.00	0.01	8.10	0.17	23.00	0.07	29.00	<0.01	6.34	6.13	240.50	19.00
46	12.20	0.97	3.20	<0.005	<0.02	8.35	7.50	15.80	0.01	0.89	0.06	15.80	<0.04	37.00	<0.01	6.42	1.25	135.00	17.00
49	9.90	0.78	2.00	<0.005	<0.02	5.40	4.25	10.90	0.00	0.27	0.06	13.20	0.04	26.00	<0.01	6.68	3.23	97.00	17.00
51	11.60	1.65	3.40	<0.005	<0.02	6.40	9.60	14.50	0.01	2.20	0.09	15.50	0.06	24.00	<0.01	6.26	4.69	125.00	16.00
52	26.00	1.40	6.30	<0.005	<0.02	9.70	11.85	19.30	0.01	2.80	0.17	25.00	0.13	65.00	<0.01	6.70	1.73	228.00	17.00
53	10.40	1.15	2.50	0.01	0.12	6.90	6.50	13.85	<0.004	1.20	0.05	12.30	<0.04	32.00	<0.01	6.33	2.78	109.00	18.00
54	18.30	2.00	4.00	<0.005	0.01	6.20	9.10	17.30	0.01	1.65	0.07	25.50	0.06	33.00	<0.01	6.15	3.17	175.50	18.00
55	25.00	2.70	8.00	<0.005	<0.02	13.30	6.00	36.00	0.13	1.30	0.15	30.50	0.08	82.00	<0.01	8.51	4.48	257.00	18.00
56	11.50	2.30	1.50	<0.005	<0.02	2.80	3.00	86.00	0.12	0.17	0.27	6.50	<0.04	35.00	<0.01	7.24	8.70	87.50	16.50
57	139.00	6.40	7.20	0.50	0.63	9.70	0.15	81.50	0.15	0.05	0.50	186.00	0.62	134.00	0.03	7.54	4.69	807.00	18.00
58	9.80	1.45	2.00	<0.005	<0.02	3.30	1.40	68.00	0.09	1.25	0.04	5.10	<0.04	34.00	<0.01	6.87	9.05	83.70	18.50
59	21.00	0.67	5.70	0.26	<0.02	8.40	2.40	25.00	0.04	0.46	0.18	12.30	0.04	88.00	<0.01	6.94	0.30	177.00	18.00
62	17.00	1.30	6.25	<0.005	0.03	43.50	11.90	25.00	0.01	0.90	0.08	14.50	0.04	162.50	<0.01	7.69	ND	329.00	17.00
63	34.00	1.30	5.70	0.07	1.40	7.30	6.30	41.00	0.01	<0.03	0.09	50.00	0.24	50.00	0.01	6.86	ND	266.00	17.50
65	41.00	2.60	5.50	0.11	0.06	33.00	8.30	41.00	0.09	<0.03	0.11	50.00	0.17	146.00	0.07	8.24	ND	400.00	17.00
66	2.30	0.73	2.60	<0.005	<0.02	45.00	4.10	7.10	<0.004	0.67	0.03	1.30	<0.04	150.00	<0.01	7.40	6.34	233.00	17.50
67	5.40	1.15	5.40	<0.005	<0.02	22.00	5.50	11.30	<0.004	0.72	0.07	2.10	<0.04	99.00	<0.01	7.12	8.25	173.50	18.00
69	7.35	0.87	4.40	<0.005	<0.02	32.00	6.00	13.50	0.01	0.49	0.12	2.80	<0.04	124.00	<0.01	7.40	7.25	230.00	18.00
71	10.00	1.15	3.80	<0.005	<0.02	27.00	5.00	13.30	<0.004	0.76	0.09	3.80	0.03	114.00	<0.01	7.89	7.08	198.50	18.00
72	2.00	0.34	1.80	<0.005	<0.02	15.20	3.80	7.50	0.01	0.14	0.11	0.49	<0.04	56.00	<0.01	8.09	7.12	94.50	18.00
73	8.40	0.54	12.70	<0.005	<0.02	18.50	13.70	16.00	0.01	2.90	0.04	12.40	0.06	96.00	<0.01	7.17	6.36	214.00	20.00
74	4.60	0.90	7.35	<0.005	<0.02	23.00	14.20	14.40	0.01	2.25	<0.03	6.30	<0.04	85.00	<0.01	7.10	7.42	200.00	20.00

CHAPTER 3.1

75	91.00	3.90	0.87	<0.005	<0.02	2.30	1.80	10.80	0.05	<0.03	0.11	34.00	0.00	193.00	0.25	8.99	0.09	424.00	18.00
76	21.00	3.10	7.60	0.06	0.12	32.00	0.11	46.00	0.14	<0.03	0.09	22.00	0.06	160.50	0.19	8.18	0.14	310.00	18.50
77	22.00	1.95	9.80	0.13	6.00	21.00	0.09	61.00	0.05	<0.03	0.11	24.00	0.09	134.00	0.15	7.02	0.27	278.00	18.50
78	36.00	1.60	30.00	0.05	1.20	78.00	0.03	23.00	0.01	<0.03	0.04	72.00	0.22	364.00	0.30	7.87	0.12	749.00	20.00
79	65.50	5.05	11.40	0.71	0.46	103.00	0.96	27.00	0.05	0.00	0.37	34.50	0.14	503.50	0.57	7.62	0.23	864.00	19.50
80	105.00	9.30	25.00	0.34	8.50	158.00	0.12	37.00	0.05	0.01	0.34	115.00	0.48	732.50	4.10	7.62	1.61	1358.50	19.00
81	248.00	23.00	45.50	1.95	22.00	284.00	39.00	39.00	0.01	0.00	0.19	747.00	2.15	612.00	0.60	7.51	0.30	2615.00	20.00
82	30.50	5.00	8.95	0.68	1.30	102.00	5.25	34.00	0.12	0.00	0.35	25.00	0.14	392.00	0.60	7.76	0.19	682.50	19.00
83	34.50	4.50	12.20	0.36	0.22	120.00	86.50	31.50	0.01	0.27	0.26	71.00	0.33	286.00	0.02	7.76	0.30	829.50	20.00
338	88.00	7.60	22.00	0.14	3.30	111.00	0.00	14.90	0.06	0.00	0.09	47.00	0.18	657.00	5.50	7.95	1.80	1092.00	19.00
340	17.50	0.85	9.30	<0.005	<0.02	10.90	1.90	47.00	0.04	9.50	0.02	16.60	0.08	52.00	<0.01	6.57	7.47	219.50	19.00
346	8.15	2.20	4.80	<0.005	<0.02	6.80	8.60	32.00	0.03	2.80	<0.03	14.90	0.04	24.00	<0.01	6.55	7.95	139.00	18.00
347	15.30	5.60	6.30	<0.005	<0.02	17.30	31.00	75.50	0.04	6.50	<0.03	17.60	0.01	38.00	<0.01	6.68	9.70	211.00	17.50
349	24.00	2.15	25.00	<0.005	<0.02	26.00	1.80	34.50	0.03	21.00	0.02	79.00	0.21	43.00	<0.01	6.73	8.19	462.00	18.50
350	23.00	1.20	8.10	<0.005	<0.02	7.70	2.85	30.00	0.01	9.90	0.02	31.00	0.11	31.00	<0.01	6.61	8.20	244.00	18.00
351	14.60	1.20	6.00	<0.005	<0.02	32.50	10.25	15.60	0.01	0.48	0.08	18.50	<0.04	115.50	<0.01	7.14	2.36	278.00	17.00
358	26.00	4.00	8.00	0.02	0.03	16.20	0.81	34.00	0.21	<0.03	0.12	23.00	0.05	123.00	0.18	8.50	0.19	274.00	18.00
360	25.00	1.60	3.90	0.05	1.55	4.90	6.20	36.00	0.01	0.31	0.02	32.00	0.05	38.00	0.11	6.70	ND	197.00	18.00
362	8.40	0.87	3.20	<0.005	<0.02	9.05	3.75	35.00	0.04	1.50	0.11	6.65	<0.1	46.00	<0.01	7.15	6.17	118.50	18.00
363	66.00	6.35	11.95	0.06	1.40	108.00	13.35	40.00	0.05	0.03	0.15	86.00	0.15	409.00	0.22	7.65	0.27	864.00	19.00
364	85.50	6.20	8.90	0.51	2.70	83.50	0.12	61.50	1.30	0.00	0.34	74.00	0.40	420.50	3.80	7.61	0.21	876.00	19.00
378	5.50	1.00	1.30	0.03	0.07	8.70	4.70	11.90	0.01	0.12	0.04	7.15	<0.04	32.00	<0.01	6.29	0.82	89.00	19.00
380	2.40	2.00	1.00	<0.005	0.08	11.10	5.80	8.30	0.01	1.10	0.05	3.20	<0.04	32.00	<0.01	6.16	7.35	89.00	18.00
382	7.50	0.69	2.70	0.01	<0.02	10.60	7.30	19.90	0.02	3.95	0.04	7.05	0.05	29.00	<0.01	6.26	6.77	122.00	19.00
383	2.90	2.60	1.40	<0.005	<0.02	9.80	4.80	13.05	0.01	1.40	0.04	3.30	<0.04	34.00	<0.01	6.09	5.10	84.40	17.00
384	8.20	1.60	2.80	<0.005	<0.02	6.20	4.90	16.70	0.03	0.75	0.09	9.90	0.04	32.00	<0.01	6.22	5.28	101.20	17.10

CHAPTER 3.1

388	15.80	0.65	11.20	<0.005	<0.02	27.00	14.10	26.00	0.01	6.80	0.05	21.00	<0.04	94.00	<0.01	6.83	8.36	301.00	18.00
389	13.60	0.79	4.80	0.01	<0.02	13.40	10.20	15.00	0.02	8.00	0.06	24.00	0.14	14.95	<0.01	5.79	8.53	205.00	17.00
390	17.80	0.68	8.20	1.10	0.41	14.00	22.00	26.00	<0.004	0.41	0.11	21.00	0.06	69.00	0.02	6.65	4.25	230.00	18.00
395	14.20	2.10	5.20	<0.005	<0.02	16.20	19.65	12.80	0.02	5.70	0.05	22.00	0.05	30.00	<0.01	6.34	7.20	225.50	19.50
437	26.00	1.00	9.20	0.23	0.02	17.60	4.20	22.00	0.02	0.06	0.16	10.90	0.01	148.00	<0.01	7.78	3.48	266.00	18.00
455	18.70	0.60	5.20	<0.005	<0.02	7.50	4.70	30.00	0.03	1.20	0.11	24.00	0.05	50.00	<0.01	6.57	7.40	180.00	17.50
456	17.30	2.40	4.70	0.06	0.62	4.85	3.70	11.80	0.03	0.00	0.09	6.85	<0.04	74.00	0.02	8.03	0.15	141.50	17.50
458	6.00	2.20	2.85	<0.005	0.12	5.30	5.05	55.50	0.05	0.26	0.04	4.05	<0.04	36.50	<0.01	6.85	6.14	88.00	16.50
465	16.00	0.91	4.20	<0.005	<0.02	8.80	9.50	25.00	0.02	4.70	0.17	11.90	0.04	41.00	<0.01	6.69	1.03	167.00	18.00
466	19.15	1.40	4.80	<0.005	<0.02	28.00	11.90	30.00	0.01	11.00	0.07	23.00	0.05	60.00	<0.01	6.60	2.20	282.50	18.00
467	32.00	1.40	7.45	0.48	0.97	17.00	10.80	23.00	0.04	<0.03	0.16	54.00	0.18	74.00	0.09	6.80	0.06	331.00	18.00
468	162.50	7.55	22.50	1.40	15.70	45.50	<0.1	39.00	0.19	<0.03	0.36	306.00	1.30	246.00	10.85	7.06	0.19	1200.00	18.00
469	14.00	0.84	7.10	0.08	0.05	12.80	8.70	18.70	0.01	2.75	0.07	18.20	0.05	60.00	<0.01	6.81	3.01	200.50	18.50
512	75.00	0.60	3.10	0.07	0.02	33.00	15.95	14.90	0.21	0.03	0.05	50.00	0.13	217.00	0.35	8.32	0.16	516.00	18.00
552	6.95	0.81	3.20	<0.005	<0.02	12.60	10.80	13.40	0.01	1.30	0.08	4.40	<0.04	49.00	<0.01	6.88	7.20	131.50	18.00
553	3.70	0.55	1.45	<0.005	<0.02	6.35	2.50	9.10	0.01	0.20	0.04	3.10	<0.04	29.00	<0.01	6.70	8.72	64.30	17.00
1993	3.90	3.00	2.00	0.18	0.04	12.40	9.40	4.60	<0.004	4.70	<0.03	10.00	0.18	16.00	<0.01	5.70	3.30	140.00	18.00
2013	112.50	3.45	13.45	0.01	<0.02	6.60	25.50	32.00	0.05	2.20	0.08	180.00	0.69	46.00	<0.01	6.25	ND	746.00	19.00
2015	12.10	1.45	7.40	0.06	0.30	13.90	5.50	16.80	0.01	0.13	0.16	8.45	0.03	95.00	<0.01	7.47	0.16	183.00	18.00
2016	3.80	0.51	1.75	<0.005	<0.02	7.10	3.90	9.00	0.02	0.26	0.06	3.25	<0.04	30.50	<0.01	6.93	8.09	68.45	18.00
2069	2.50	2.25	1.70	<0.005	0.02	14.90	6.75	9.00	0.01	1.40	0.04	3.00	<0.04	46.00	<0.01	6.25	3.62	119.00	16.00
2103	5.75	1.60	1.90	<0.005	<0.02	11.40	6.60	11.70	0.01	4.60	0.07	5.60	0.04	23.00	<0.01	5.85	4.68	116.00	18.00
3280	14.45	0.43	8.10	<0.005	<0.02	12.20	1.20	37.00	0.03	2.90	<0.03	15.40	<0.1	77.50	<0.01	7.47	ND	201.00	17.00
3281	13.60	1.50	4.40	0.02	0.16	5.90	2.70	26.00	0.05	1.90	0.06	18.10	<0.1	37.00	<0.01	6.56	ND	143.00	18.00
3325	6.40	1.80	6.75	<0.005	<0.02	19.80	16.00	15.10	0.03	1.70	0.04	8.45	0.05	66.00	<0.01	6.85	2.61	192.00	18.00
3327	23.00	0.94	9.60	0.14	7.80	15.30	4.55	37.00	<0.004	<0.03	0.21	25.00	0.12	125.00	0.03	7.00	1.96	271.50	18.50

Table S3. Summary of P values (95.0% confidence level) of Kruskal-Wallis test for each parameter at different threshold levels. **Bold** values show statistical significance ($P < 0.05$).

	3-Cluster	4-Cluster	5-Cluster	7-Cluster	11-Cluster
Na	0.7999	0.7012	0.3351	0.5197	0.0158
Mg	0.3683	0.0496	0.0569	0.0834	0.0106
K	0.6568	0.4499	0.5517	0.3078	0.3803
Ca	0.1131	0.0024	0.0060	0.0191	0.0108
Mn	0.9484	0.3094	0.1386	0.1814	0.0209
Fe	0.6185	0.2760	0.1118	0.0552	0.0130
SiO ₂	0.7195	0.5315	0.5339	0.1769	0.0054
NO ₃ -N	0.8961	0.4765	0.1955	0.3708	0.2004
SO ₄	0.2718	0.2093	0.2805	0.2423	0.3451
HCO ₃	0.1756	0.0017	0.0036	0.0142	0.0138
NH ₄ -N	0.7132	0.1271	0.0852	0.0644	0.0047
P	0.1620	0.2968	0.4381	0.3336	0.1334
Cl	0.7813	0.9165	0.6864	0.6039	0.0429
Br	0.9487	0.9811	0.8289	0.9524	0.4161
F	0.4912	0.4927	0.0917	0.2247	0.0173
EC	0.1324	0.0325	0.0536	0.1218	0.0057
Temperature	0.3466	0.5229	0.6543	0.3916	0.3685
DO	0.5931	0.7608	0.2809	0.4420	0.0496

Table S4. Summary of P values (95.0% confidence level) of Chi- square test for each parameter at different threshold levels. **Bold** values show statistical significance ($P < 0.05$).

	3-Cluster	4-Cluster	5-Cluster	7-Cluster	11-Cluster
Region	0.0069	0.0077	0.0462	0.0175	0.0064
Depth Code	0.9549	0.6870	0.8607	0.9193	0.7227
MRT Class	0.2494	0.1187	0.1032	0.0818	0.0617
Hydrochemical Clusters	0.0367	0.0006	0.0003	0.0016	0.000
Land use	0.3025	0.1156	0.3233	0.1006	0.1364
Aquifer Lithology	0.3621	0.3104	0.1229	0.1939	0.1621

CHAPTER 3.1

Table S5. Characteristics of groundwater sampling sites

		Auckland	Bay of Plenty	Canterbury	Southland	Waikato	Gisborne	Wellington	Hawke's Bay	Manuwatu-Wanganui	Marlborough	Northland	Otago	Taranaki	Tasman	West Coast	Total
Total sites		6	4	5	7	9	6	13	5	2	8	7	5	5	10	8	100
Surrounding land use	Agriculture	-	2	4	6	2	-	-	4	-	1	-	2	5	1	4	31
	Dairy	2	-	-	-	3	-	3	-	-	-	-	-	-	-	-	8
	Forestry	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	1
	Horticulture	4	-	-	-	2	6	3	-	2	5	3	3	-	9	-	37
	Industrial	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	1
	Park/Reserve	-	2	-	-	-	-	2	-	-	-	-	-	-	-	-	4
	Urban	-	-	1	-	1	-	4	1	-	2	2	-	-	-	1	12
	Unknown	-	-	-	1	-	-	-	-	-	-	2	-	-	-	3	6
Aquifer lithology	Basalt	2	-	-	-	3	-	-	-	-	-	1	-	-	-	-	6
	Gravel	-	-	5	6	1	2	13	2	2	-	-	5	-	8	4	48
	Greywacke	-	-	-	-	-	-	-	1	-	-	1	-	-	-	-	2
	Ignimbrite	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	1
	Lignite	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
	Limestone	-	1	-	-	-	-	-	-	-	1	-	-	-	2	-	4
	Pumice	-	1	-	-	1	1	-	-	-	-	-	-	-	-	-	3
	Rhyolite	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	1
	Sand	-	-	-	-	4	-	-	-	-	7	2	-	5	-	1	19
	Sandstone	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3
	Shellbeds	1	-	-	-	-	-	-	-	-	-	1	-	-	-	-	2
	Unknown	-	-	-	1	-	3	-	2	-	-	2	-	-	-	3	11

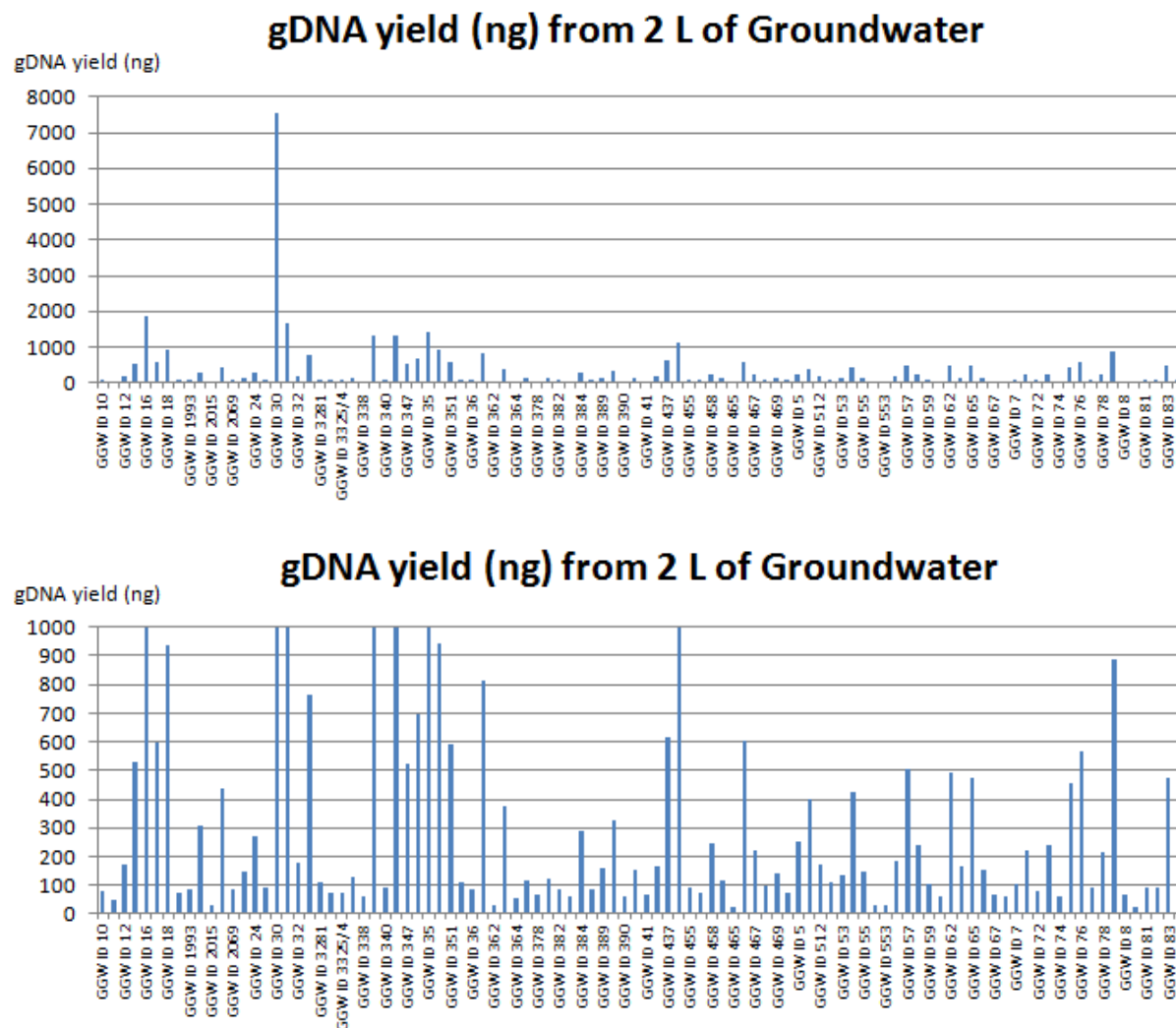


Fig. S1. Summary of genomic DNA yields obtained from two litres of groundwater from each sample. X-axis labels denote the unique site identification numbers used in the Geothermal and Groundwater Database (<http://ggw.gns.cri.nz/ggwdata/mainPage.jsp>).

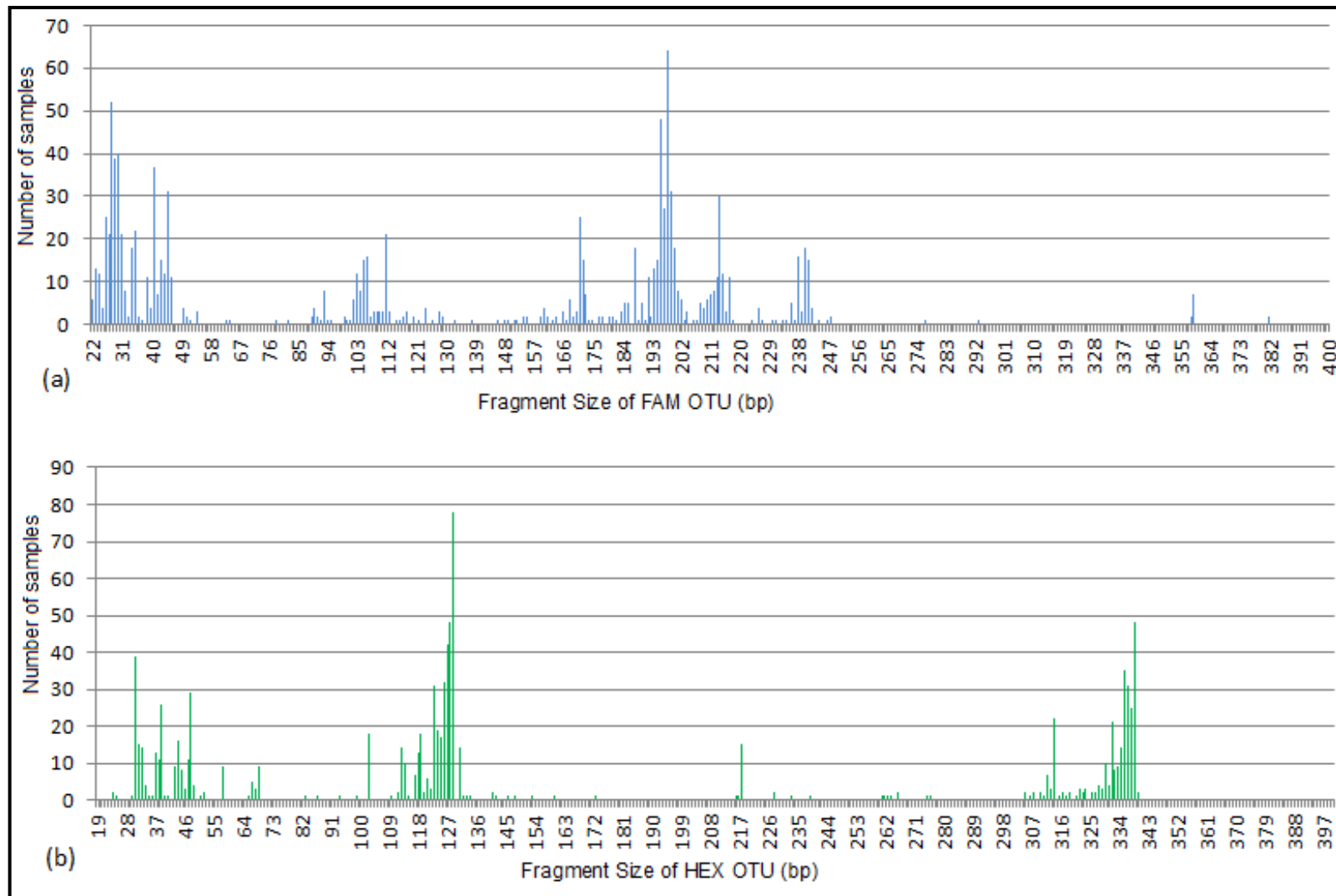


Fig. S2. Summary of the number of samples detected with each (a) FAM and (b) HEX Operational Taxonomic Unit (OTU).

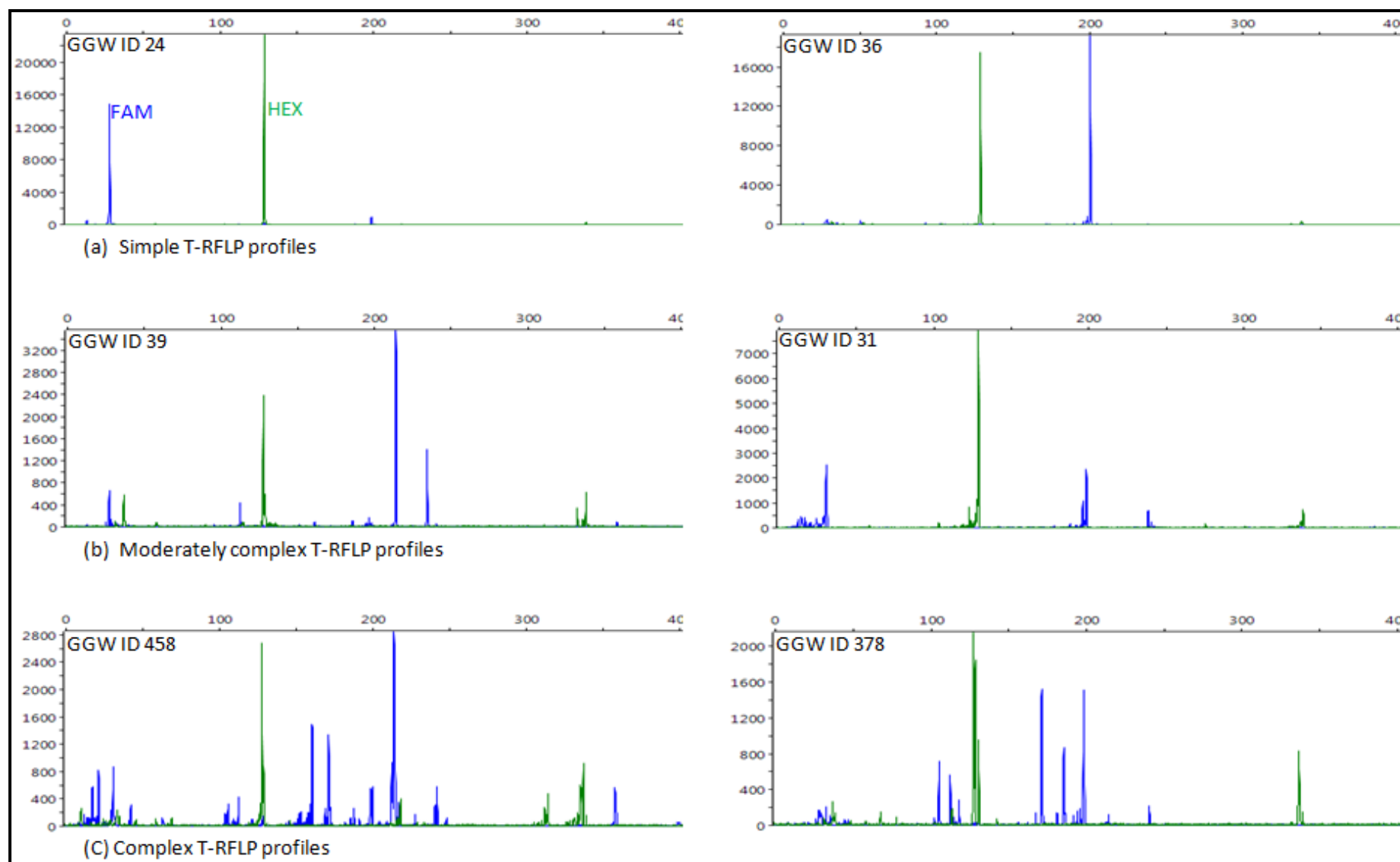


Fig. S3. Examples of T-RFLP profiles categorized as (a) simple, (b) moderately complex or (c) complex based on number of FAM or HEX peaks.

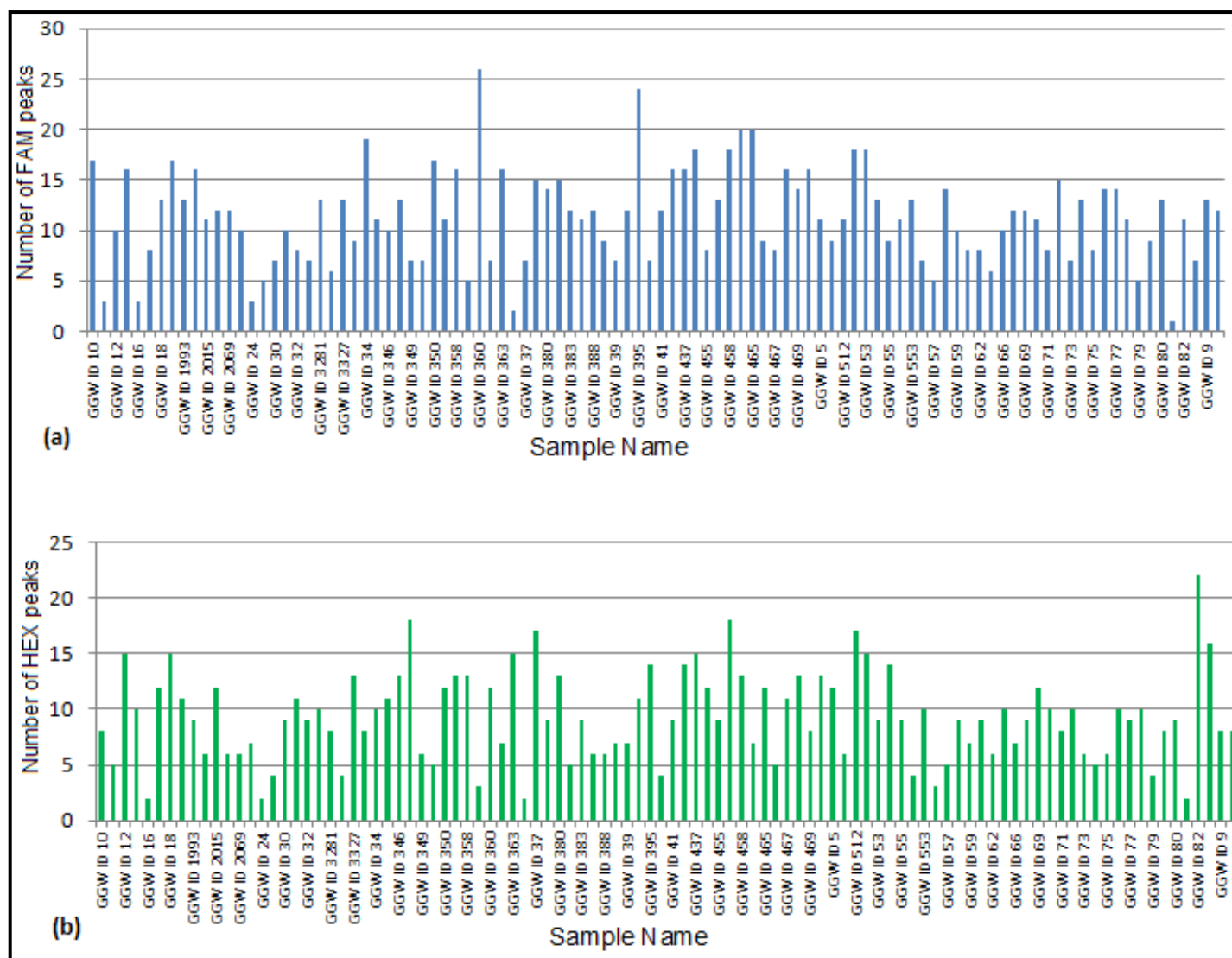


Fig. S4. Summary of the total number of FAM (a) and HEX (b) peaks over 200 RFU in each sample

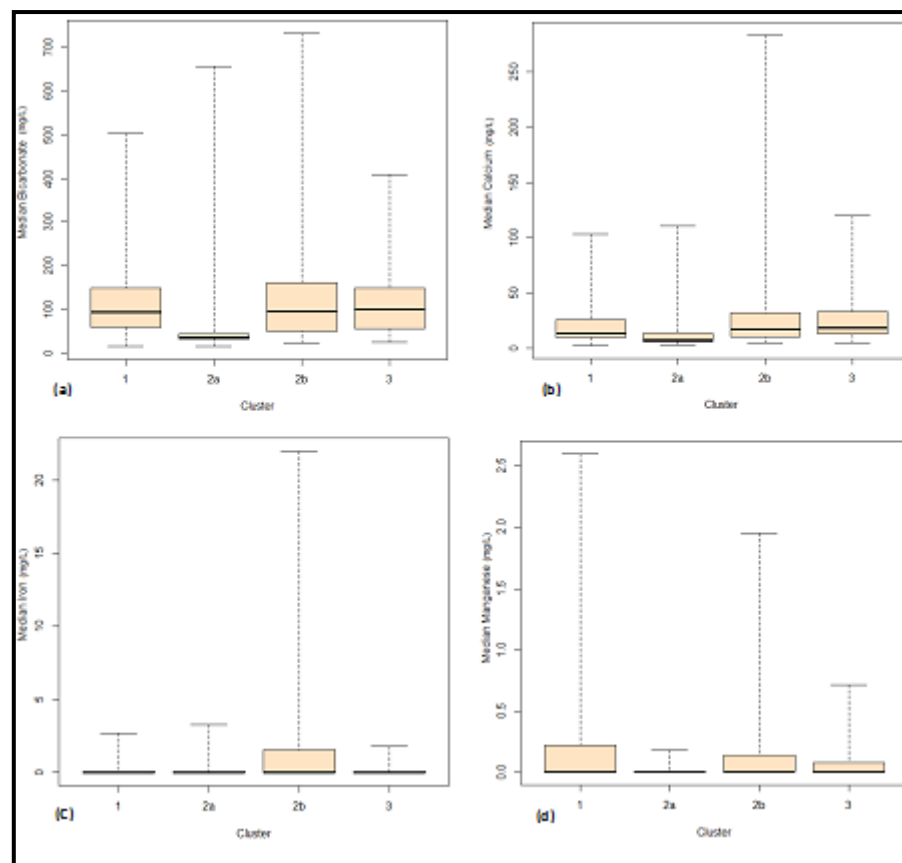


Fig. S5 The Box-and-Whisker Plot of median HCO_3^- (a), Ca (b), Fe (c) and Mn (d) across Bioclusters defined at the 4-cluster threshold

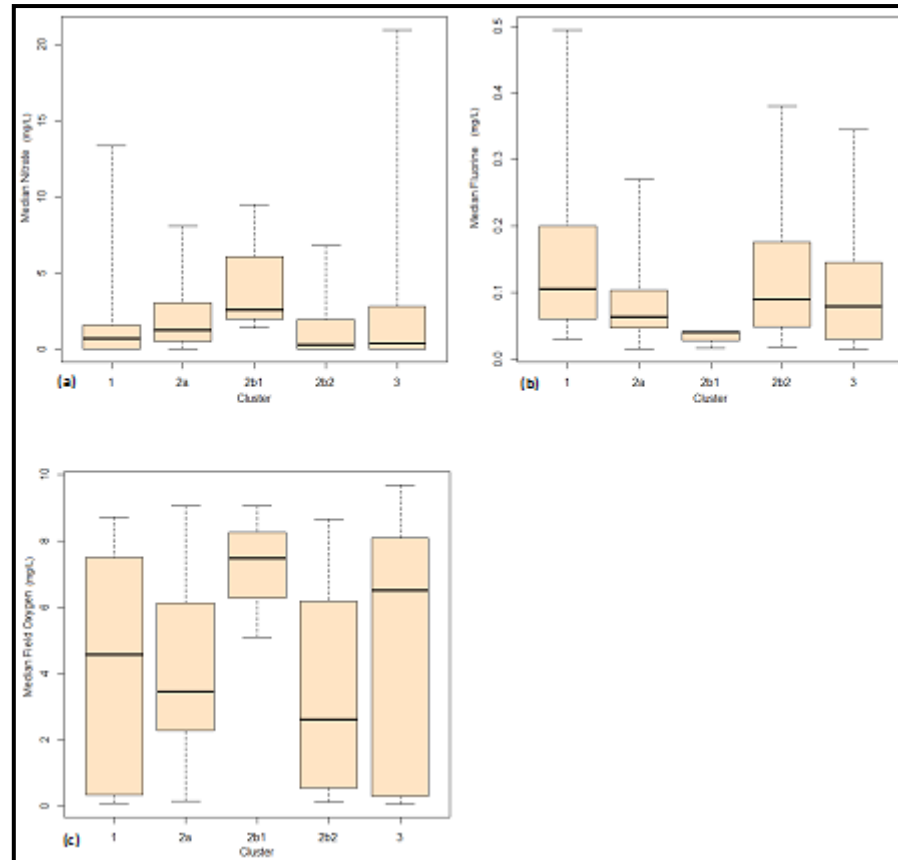


Fig. S6 The Box-and-Whisker Plot of median $\text{NO}_3\text{-N}$ (a), F (b) and Dissolved Oxygen (c) across Bioclusters defined at the 5-cluster threshold

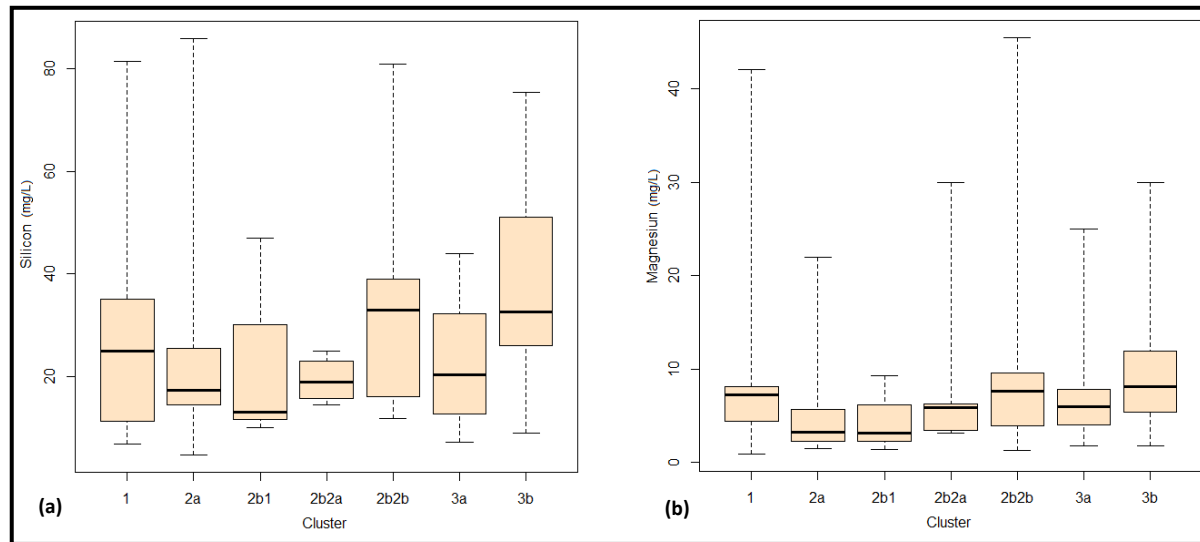


Fig. S7 The Box-and-Whisker Plot of median SiO₂ (a) and Mg (b) across Bioclusters defined at the 7-cluster threshold

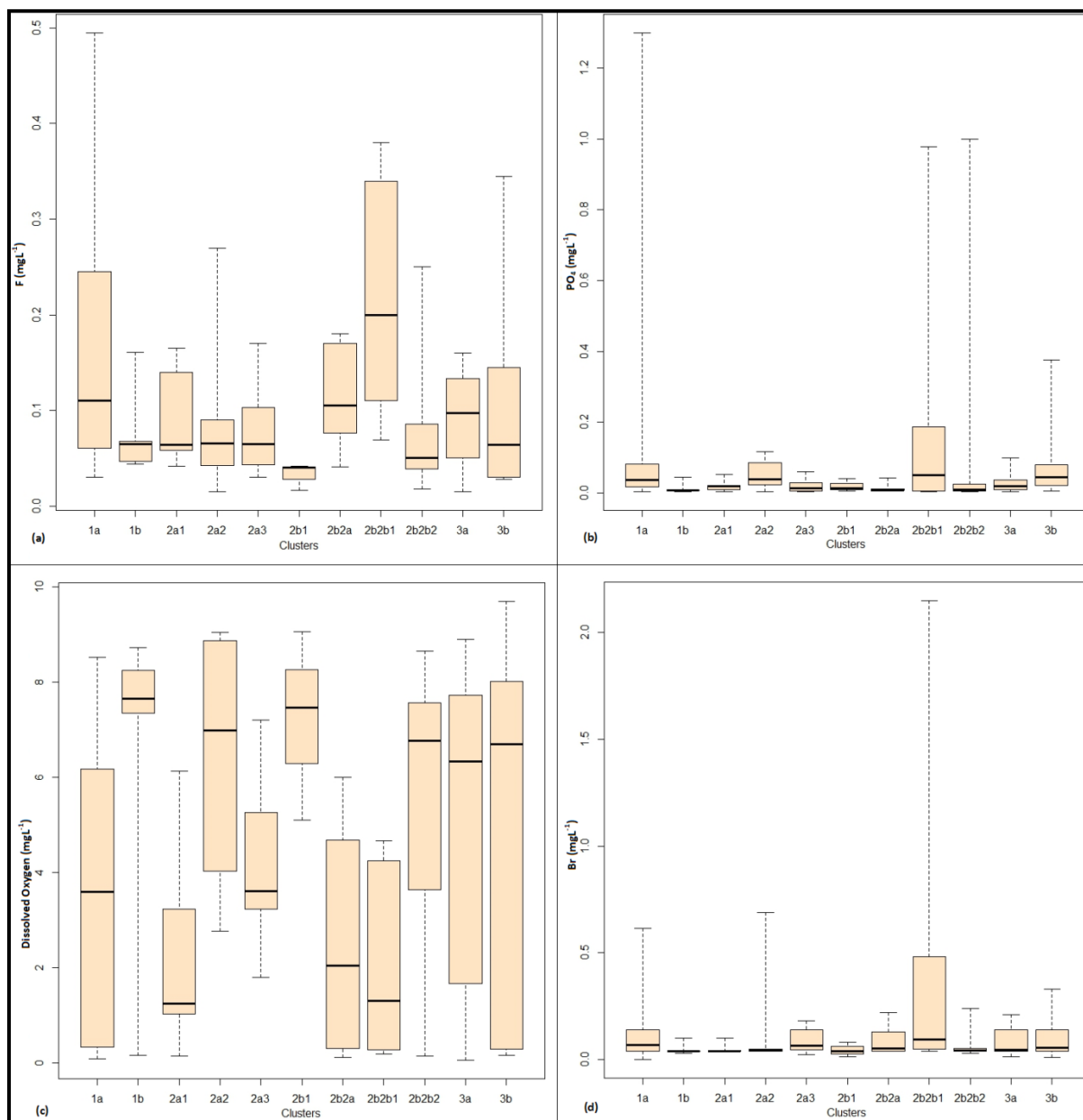


Fig. S8 (i). Box-and-Whisker Plot of median concentrations F (a), $\text{PO}_4\text{-P}$ (b), Dissolved Oxygen (c) and Br (d) across Bioclusters defined at the 11- cluster threshold.

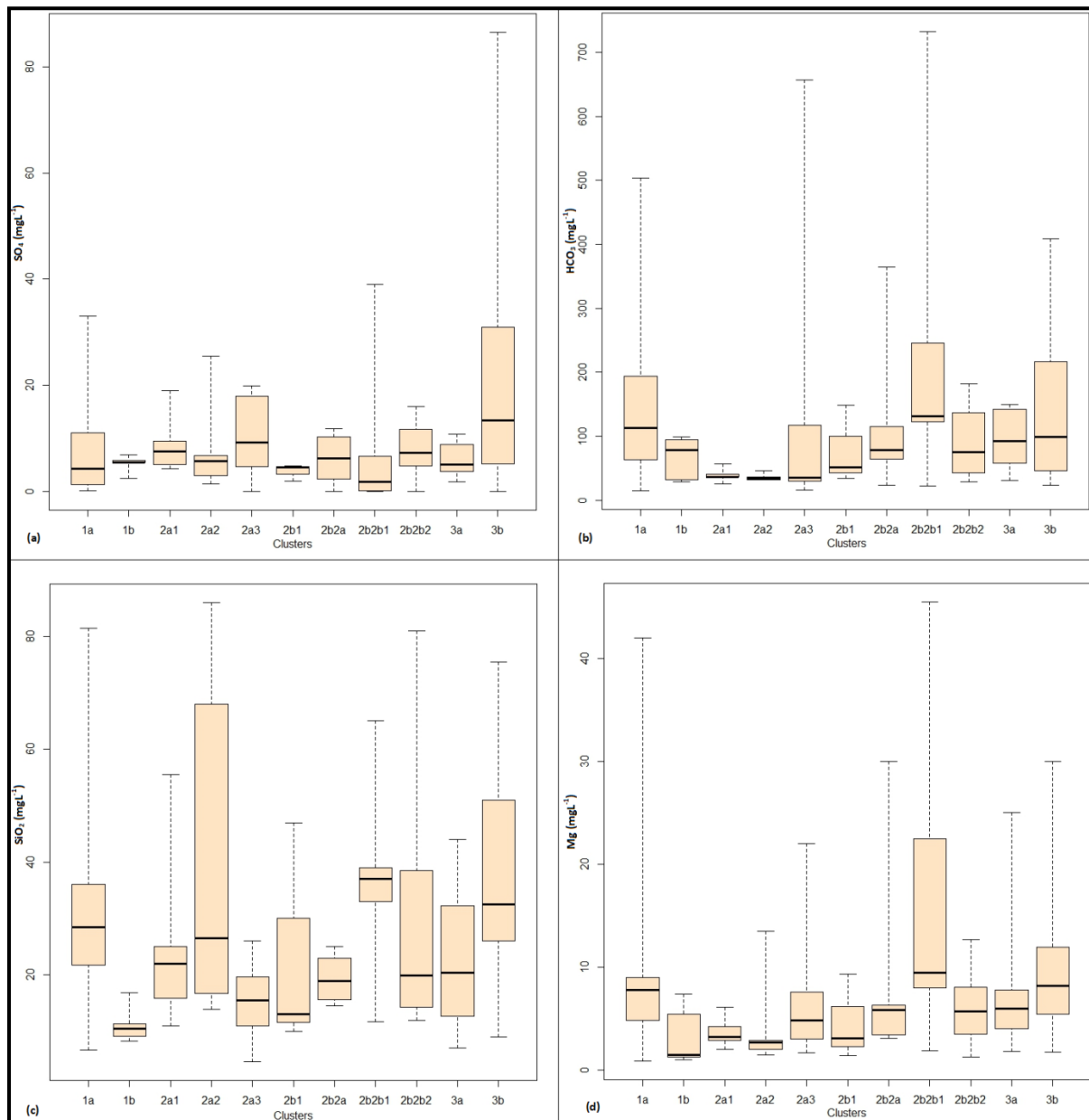


Fig. S8 (ii). Box-and-Whisker Plot of median concentrations of SO_4 (a), HCO_3 (b), SiO_2 (c) and Mg (d) across Bioclusters defined at the 11-cluster threshold.

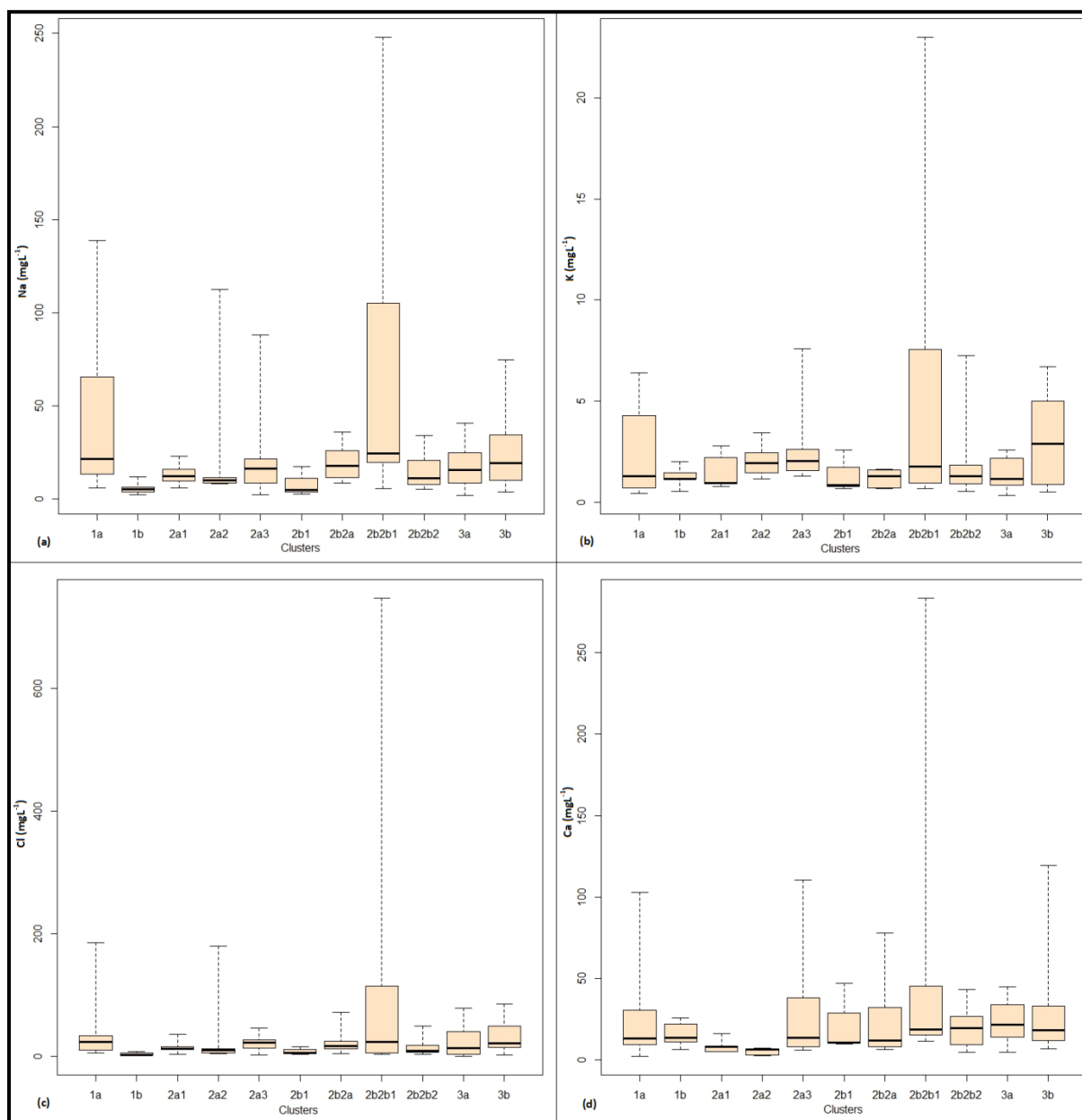


Fig. S8 (iii). Box-and-Whisker Plot of median concentrations of Na (a), K (b), Cl (c) and Ca (d) across Bioclusters defined at the 11-cluster threshold.

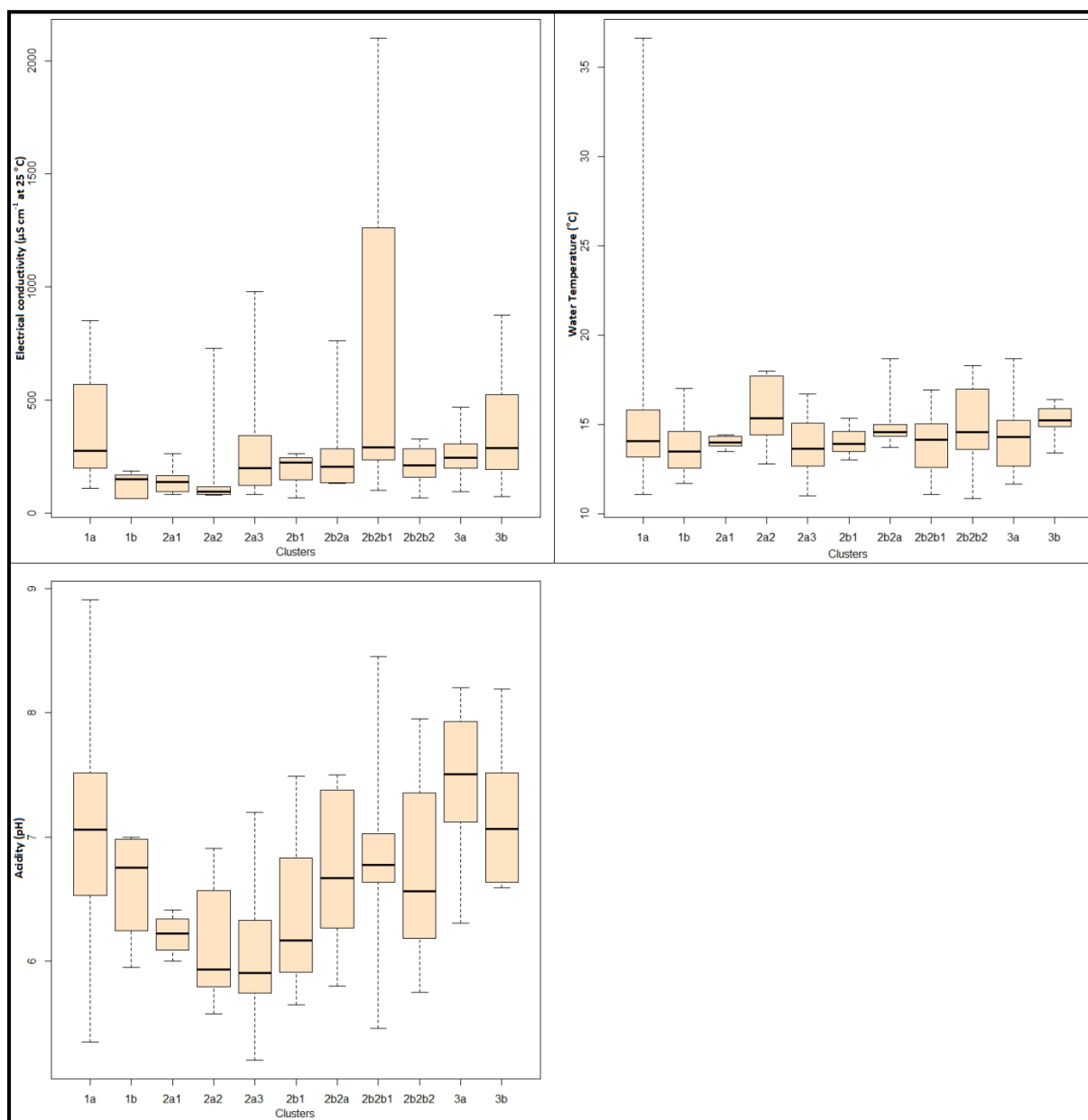


Fig. S8 (iv). Box-and-Whisker Plot of median concentrations of Electrical conductivity (a), Water temperature (b) and Acidity (h) across Bioclusters defined at the 11-cluster threshold.

Relationships between molecular bacterial diversity and chemistry of groundwater in the Wairarapa Valley, New Zealand

Sirisena KA, Daughney CJ, Moreau M, Ryan KG, Chambers GK 2014. Relationships between Molecular Bacterial Diversity and Chemistry of Groundwater in the Wairarapa Valley, New Zealand. New Zealand Journal of Marine and Freshwater Research (*In press*)

Running title: Bacterial diversity in Wairarapa groundwater

Abstract

Groundwater plays an important role in New Zealand water supplies and hence monitoring activities are conducted regularly. Most monitoring programmes aim to evaluate groundwater chemistry and almost completely overlook the microbial component in this ecosystem. In our present study, the bacterial community structure of groundwater in the Wairarapa Valley was examined using the terminal restriction fragment length polymorphism (T-RFLP), and relationships between bacterial community structure and groundwater chemistry, aquifer confinement and groundwater usage were explored. In addition, the results from this study were compared with a previous T-RFLP survey of the same area in an attempt to detect changes in bacterial community structure over time. The data obtained suggested that bacterial community structure was related to groundwater chemistry, especially to redox conditions. Species composition showed minimal variation over time if groundwater chemistry remained unchanged. These findings reflect the potential of using bacterial communities as biological indicators to evaluate the health of groundwater ecosystems. We suggest that it is important to include this type of broad bacterial diversity assessment criteria into regular groundwater monitoring activities.

Keywords

Bacterial Diversity; Groundwater; T-RFLP; DNA Analysis; Microbial Ecology;
Environmental monitoring; New Zealand

Introduction

Groundwater is one of the most valuable natural resources around the globe. A large proportion of the world's population directly depends on groundwater for its water requirements. It is the world's major drinking water source, providing about 60% of drinking water in Europe with an even greater percentage in individual countries and more than 80% in North Africa and the Middle East (Struckmeier et al. 2005; Steube et al. 2009).

Groundwater plays a crucial role in urban and rural water supplies in New Zealand too, where nearly one quarter of the population uses groundwater as its major drinking water source.

Groundwater also supplies a significant fraction of the requirements for the agricultural and industrial sectors (Daughney & Reeves 2005).

Due to this high demand, groundwater monitoring activities are extensively conducted throughout the world to assess quality and availability. However, the majority of these monitoring programmes are restricted to the evaluation of physical and chemical parameters as measures of groundwater quality. In recent years, there has been an increasing trend to consider groundwater not only as a valuable resource for human use, but also as a dynamic ecosystem. Therefore, in addition to chemical monitoring, assessments of ecological status, in some cases including the microbial component, have also been included into national groundwater monitoring policies in some parts of Europe and Australia (Steube et al. 2009; Griebler et al. 2010; Stein et al. 2010; Zhou et al. 2012; Korbel & Hose 2011; Korbel et al. 2013). It is likely that the microbial component plays an important role in subsurface ecosystems, including groundwater, as it provides the driving force for biogeochemical processes taking place in these environments (Falkowski 2008). Therefore, the species composition of groundwater microbiota should be considered in groundwater monitoring programmes. For example, it is crucial to understand the groundwater microbial diversity in

the absence of human influence (i.e. under baseline conditions) to enable identification of its relationships to anthropogenic pressures and other environmental factors (Larned 2012).

In New Zealand, groundwater monitoring is undertaken by many organisations, of which the various regional authorities are the most active. They operate State-of-the-Environment (SOE) groundwater quality monitoring programmes within their own areas of jurisdiction and also collaborate in the operation of the National Groundwater Monitoring Programme (NGMP), which is comprised of 110 monitoring sites around the country (Daughney et al. 2012; Sirisena et al. 2013). The NGMP is a long-term research and monitoring programme that aims to identify spatial patterns and temporal trends in groundwater quality at the national scale and relate them to specific causes (Rosen 2001; Daughney & Reeves 2005, 2006; Morgenstern & Daughney 2012). The regional SOE programmes typically only assess the presence of coliform bacteria (mainly *Escherichia coli*) as a biological factor, because it is an indicator species of faecal contamination that could cause serious human health problems (Ministry for the Environment 2010; Greater Wellington Regional Council 2013).

A preliminary evaluation of microbial biodiversity in New Zealand's groundwater was conducted by van Bekkum et al. (2006). In this pilot study, bacterial community structure was determined using 20 groundwater samples collected from bores around the Hutt Valley and Wairarapa regions. This work provided initial indications of relationships between bacterial community structure and groundwater chemistry. At the national scale, a more recent study evaluated the relative abundance of bacterial species in groundwater at all NGMP sites (Sirisena et al. 2013). This study revealed considerable microbial biodiversity in New Zealand groundwater, finding strong relationships between community structure and groundwater chemistry, in particular with regard to the influence of redox potential and the degree of human impact.

In the present study, we evaluated microbial biodiversity of groundwater in the Wairarapa Valley using one standard, culture independent, DNA-based molecular profiling tool: Terminal Restriction Fragment Length Polymorphism (T-RFLP). It is widely believed that culturing methods do not reveal the full array of bacterial diversity in natural environmental samples, because the majority of species present in such environments cannot be easily grown in artificial culture media (Zhou et al. 1997; Janssen et al. 2002; Neufeld & Mohn 2005; Lozupone & Knight 2007). Hence, T-RFLP is an alternative culture-independent, rapid, cost effective and sensitive technique for characterisation of microbial community structure in environmental samples (Liu et al. 1997; Edlund et al. 2006; van Bekkum et al. 2006; Sirisena et al. 2013). Further, recent studies have demonstrated that T-RFLP can be highly effective, even as efficient as modern high throughput sequencing techniques in revealing bacterial community structure (Camarinha-Silva et al. 2012; Pilloni et al. 2012).

The present study has four objectives. The first is to explore the bacterial community structure in groundwater in parts of the Wairarapa Valley that were not previously investigated by van Bekkum et al. (2006) or by Sirisena et al. (2013). The second objective is to determine the regional-scale relationships between bacterial community structure and groundwater chemistry, aquifer confinement and groundwater bore usage, for comparison with the general conclusions drawn in the national-scale study of Sirisena et al. (2013). The third objective of this study is to compare the present bacterial structure in the Wairarapa Valley groundwater with the results of van Bekkum et al. (2006) in an attempt to measure changes in community structure over time. The fourth and final objective is to compare different approaches for analysis of T-RFLP data. At present, there is no commonly accepted best practice approach for the analysis of T-RFLP data (Blackwood et al. 2003). Thus, we have applied and compared two common methods for standardizing T-RFLP peaks: (1)

standardization to the highest peak in the profile; and (2) standardization to the sum of all peaks in the profile. We also used two common approaches to determine the similarity between T-RFLP profiles: (1) Euclidean distance; and (2) Bray-Curtis similarity. In summary, this study has the overarching goal of providing a solid foundation for more detailed explorations of bacterial diversity in New Zealand groundwater, to move towards inclusion of microbial status assessment criteria into regional and national monitoring programmes and related policies.

Materials and methods

Study area and groundwater sampling

Groundwater samples were collected from 34 groundwater sampling sites across the Wairarapa Valley and one site from the Riversdale area which is located in the eastern coast of the Wellington region (see Fig. 1), in conjunction with the routine quarterly groundwater quality monitoring conducted by the Greater Wellington Regional Council (GWRC). Of these 35 sites, five sites were previously studied by van Bekkum et al. (2006). This provided an opportunity for a partial comparison of bacterial community structure over time.

Groundwater samples (single sample of 2 litres from each site) were collected in September 2009 into individual sterilized plastic bottles according to the National Protocol for State of the Environment Groundwater Sampling (Daughney et al. 2006). All these containers were kept at 4 °C until they were used. Additional samples were collected by GWRC staff at the same time as a part of their routine groundwater monitoring operations. The samples were analysed by Hill Laboratories (Hamilton, New Zealand) for 28 chemical parameters (Na, K, Mg, Ca, Pb, Zn, B, HCO₃, Cl, SO₄, NO₃-N, NO₂-N, NH₄-N, PO₄-P, Fe, Mn, Br, F, SiO₂, total dissolved solids (TDS), total organic carbon (TOC), alkalinity, total hardness, total cations, total anions, total oxidized nitrogen (TON), dissolved oxygen

(DO) and free CO₂) in accordance with standard analytical procedures (Daughney & Reeves 2005; Daughney et al. 2010; Tidswell et al. 2012). Electrical conductivity (EC) and pH were also measured in the field at the time of sampling. The hydrochemical data and site-specific data such as aquifer confinement and bore usage were obtained from GWRC and are summarized in Tables S1 and S2 respectively.

DNA extraction

Two litres of groundwater from each site were filtered through a sterile 0.22 µm nitrocellulose membrane filter (Millipore, Australia) using a vacuum system. The filters were stored at -20 °C in sterilized 50-ml plastic tubes until use. Extreme care was taken to avoid long storage periods and all groundwater filtrations and DNA extractions were carried out immediately as delays may alter the apparent microbial community composition in such samples (Gilpin et al. 2013). Bacterial genomic DNA extractions were performed using ZR Fungal/Bacterial DNA kits (Zymo Research, United States) as described in Sirisena et al. (2013). Extracted DNA was resuspended in 100 µl of molecular biology grade water (MO BIO Laboratories, Inc., United States), quantified using Quant-iTTM High-Sensitivity DNA Assay kits (Invitrogen, United States) and stored at -20 °C until further use.

PCR amplification and T-RFLP analysis

The PCR amplification and restriction enzyme digestion were performed as described in Sirisena et al. (2013). A c. 1400 bp region of bacterial 16S rRNA gene was amplified using two fluorescently labelled bacterial specific oligonucleotide primers: F63 (5'-[6-FAM]CAG GCC TAA CAC ATG CAA GTC-3') and R1389 (5'-[6-HEX]ACG GGC GGT GTG TAC AAG-3'). For T-RFLP analysis, 500 ng of purified PCR product was digested

with 10 U of *AluI* restriction endonuclease (Roche, United States) in a total volume of 25 μ l. Digested products were run on an ABI 3730XL DNA Analyzer (Applied Biosystems Inc., United States) along with a GeneScanTM-400HD ROXTM internal size standard (Applied Biosystems Inc., United States) to separate and precisely determine the sizes of fluorescently labelled terminal restriction fragments (T-RFs) up to 400 bp in length. The resulting T-RFLP electropherograms were transformed to numerical barcodes using GeneMapper[®] v 3.1 software (Applied Biosystems Inc., United States). Binary presence (1) or absence (0), fragment sizes (bp) and heights corresponding to each peak were tabulated using a bin size of 1 bp.

Data analysis

The tabulated T-RFLP data output from GeneMapper[®] v 3.1 was prepared for quantitative analysis using the methods of Sirisena et al. (2013). Briefly, decimal values associated with T-RF lengths (bp) were rounded to the nearest integer value using ± 0.5 bp as the binning threshold (i.e. to the nearest 1 bp). If two or more decimal fragment sizes were assigned to a single bin size after rounding, the heights of the peaks were summed as if they were a single peak. Further, FAM peaks below 21 bp, HEX peaks below 18 bp and both FAM and HEX peaks over 400 bp were eliminated from the analysis, because these correspond to the lengths of the primers or are outside the calibration range of the internal size standard described above. A threshold of 200 relative fluorescence units (RFU) was used to separate true peaks from the background noise based on a negative control T-RFLP profile as described in Sirisena et al. (2013).

In this study, we applied two of the most commonly used approaches to scale the peak heights of T-RFs in each profile: (1) peak heights were standardised relative to the highest peak in the profile (Parkinson 2004, 2009; van Bekkum et al. 2006); (2) peak

heights were standardised relative to the sum of all peaks in each profile (Culman et al. 2008).

The two separate sets of scaled peak heights were subjected to hierarchical cluster analysis (HCA), which is one of the most common approaches to evaluate the similarities between T-RFLP profiles. However, as there is no generally accepted distance measure used to perform HCA, we applied two of the most widely used metrics: (1) Euclidian distance (Dollhopf et al. 2001; Blackwood et al. 2003); and (2) Bray-Curtis similarity (Griebler et al. 2010; Stein et al. 2010). Based on the matrix of distance values obtained between each pair of samples, dendrograms were constructed using Ward's linkage method to display similarities between the samples (van Bekkum et al. 2006). The dendrogram showed those groups of samples having the most similar T-RFLP profiles; these clusters are hereafter referred to as "bioclusters", as first described in Sirisena et al. (2013). They demonstrated that a different number of bioclusters can be formed depending on the separation threshold applied, which should be selected to maximize distinction between the bioclusters while ensuring that each cluster contains enough samples to be representative of the population. They compared a range of thresholds that formed 3, 5, 7 or 11 clusters, as their study consisted of a relatively large number of samples (Sirisena et al. 2013). In the present study, we apply separation thresholds that resulted in formation of two or three bioclusters, as it appears to be the most appropriate according to the scale of the data set. The robustness of the T-RFLP technique for this sort of environmental microbial study was demonstrated by using several combinations of the peak standardization and distance measures.

The relationships between bioclusters and groundwater chemistry and categorical environmental parameters were evaluated as described in Sirisena et al. (2013). The Box-and-Whisker plot representation was used to demonstrate the relationships between

bioclusters and groundwater chemistry. Kruskal-Wallis tests were performed to reveal the statistical significance of these relationships. The cross-tabulation approach was used to reveal the links between bioclusters and categorical parameters such as aquifer confinement or groundwater bore usage. All statistical analyses were performed using the statistical programmes R (version 2.15.0) and SPSS version 19 (SPSS IBM, New York, USA).

We applied four other statistical approaches for comparison to the above-listed data analysis methods employed by Sirisena et al. (2013). First, non-metric multidimensional scaling (nMDS) (Kruskal 1964a,b) was carried out with Bray-Curtis similarities, and we compared the nMDS clustering with the hierarchical cluster analysis (HCA) pattern derived as described above. Second, Permutational Multivariate Analysis of Variance (PERMANOVA) test (Anderson et al. 2008) with 9999 permutations was used to verify whether the nMDS pattern is related to: 1) HCA clustering, 2) aquifer confinement categories, or 3) bore usage groups. Third, the RELATE analysis, a comparative Mantel-type test (Clark & Warwick 2001), was carried out to determine the relationship between bacterial diversity expressed by FAM T-RF structure and groundwater chemistry as a whole rather than individual parameters. Here, the hydrochemical data (x) were transformed to a natural log variable $[\ln(x+1)]$ in order to eliminate uneven values among different parameters. It is suggested that the Euclidian distance is more appropriate for grouping environmental data (hydrochemistry) (Ramette 2007). Therefore, two similarity matrices were computed: 1) the Euclidian distance matrix for the 30 hydrochemical parameters, and 2) Bray-Curtis similarity matrix based on FAM T-RFs. The RELATE analysis determined the correlation between the bacterial diversity and groundwater chemistry. Finally, Canonical Correspondence Analysis (CCA) (ter Braak & Smilauer 2002) was also performed to evaluate the relative contribution of each

hydrochemical parameter for shaping the microbial structure. These multivariate analyses were performed using the PRIMER v.6 statistical programme (Primer-E Ltd., Plymouth, UK) with the additional add-on package PERMANOVA+ (Anderson et al. 2008). The CCA was performed with CANOCO 5 for Windows package (ter Braak & Smilauer 2002).

For the quantitative representation of microbial diversity, Shannon diversity indices (H') were calculated as $H' = - \sum P_i \ln(P_i)$, where P_i is the relative abundance of i th T-RF in a given profile (Griebler et al. 2010; Stein et al. 2010). These calculations were based on T-RF heights, standardized relative to the sum of all peaks in a given profile as this approach more appropriately describes the relative abundance. In this analysis, H' values were determined separately for FAM and HEX T-RFs for all 35 samples individually and mean H' scores were also calculated within each biocluster.

Results

Groundwater bacterial diversity

The T-RFLP analysis detected 46 and 60 total unique bacterial T-RFs for FAM and HEX respectively. The T-RFs ranged in size from 22 bp to 248 bp for FAM and from 26 bp to 339 bp for HEX. The total number of FAM T-RFs ranged from 3 to 15 in individual samples, and HEX T-RFs from 3 to 17 (Fig. 2). The frequency of each FAM and HEX T-RF (i.e. the number of sites at which a particular T-RF occurred) is shown in Fig. 3. The FAM peaks with fragment sizes of 28, 30 and 199 bp occurred with highest frequencies; 25, 25 and 27 profiles respectively. Similarly, HEX peaks corresponding to fragment sizes of 128, 129 and 339 bp were found in 24, 34 and 25 profiles respectively. However, because more than one taxon may be represented by any one peak, a single FAM or HEX T-RF may not precisely represent a single species. Nonetheless, they will still provide a valid comparative insight into

species richness in combined analyses and can, therefore, be termed as operational taxonomic units (OTUs). The average Shannon diversity indices (H') were 1.36 ± 0.47 for FAM OTUs (ranging from 0.37 to 2.29) and 1.39 ± 0.59 for HEX OTUs that varies from 0.24 to 2.49 (Fig. 4).

Validation of T-RFLP analysis

Hierarchical cluster analysis was performed with four different combinations of peak scaling method and distance measure as explained in Materials and methods: (1) T-RFs standardized to highest peak / Euclidean distance; (2) T-RFs standardized to highest peak / Bray-Curtis similarity; (3) T-RFs standardized to all peaks / Euclidean distance; and (4) T-RFs standardized to all peaks / Bray-Curtis similarity. The four resulting dendrograms showed only minimal differences in clustering patterns (Fig. S1). At a distance threshold that results in the formation of three bioclusters (hereafter referred to as the 3-Cluster threshold), the composition of each biocluster (samples belonging to each cluster) was 100% identical for all four analysis methods described above. However, a slight difference in linkage pattern was revealed among the two peak standardization approaches. If the three bioclusters are arbitrarily named as A, B and C, in methods 1 and 2, the cluster representation was A (B, C) at the 2-cluster threshold level (Fig. S1A–B), whereas in methods 3 and 4, the pattern was (A, B) C at the 2-cluster threshold level (Fig. S1C–D). However, the effect of the choice of similarity index was minimal on clustering for this data set, as the two distance measures used in this study resulted in dendrograms with 100% similarity in linkage pattern and cluster composition. Given these minimal differences, a single cluster assignment that obtained from analysis method 4 was chosen (Fig. 5) for the presentation of results for the remainder of this report. The clusters formed at the 3-cluster threshold are henceforth referred to as bioclusters

1A, 1B and 2; these names are arbitrary but are selected to convey the relationship of the clusters to each other as depicted in Fig. 5.

Relationships between bacterial diversity and groundwater chemistry

The bioclusters at the 3-cluster threshold were compared with the 30 hydrochemical parameters (as listed in Materials and methods and shown in Table S1). The Kruskal-Wallis test results showed that the bioclusters were significantly associated ($P < 0.05$) with Na, K, Mg, Ca, B, HCO_3 , Cl, SO_4 , $\text{NO}_3\text{-N}$, $\text{NH}_4\text{-N}$, Fe, F, total dissolved solids (TDS), total hardness, total cations, total anions, dissolved oxygen (DO), total oxidized nitrogen (TON), electrical conductivity (EC) and alkalinity (Table S3). Box-and-whisker plots (Figs. 6 and S2[i-iv]) reflect qualitative aspects of these relationships. For example, bioclusters 1A and 1B were associated with low concentrations of $\text{NH}_4\text{-N}$, Fe, Mn, $\text{NO}_2\text{-N}$, $\text{PO}_4\text{-P}$ and Br and high concentrations of $\text{NO}_3\text{-N}$ and SO_4 compared with biocluster 2. Biocluster 1A can be distinguished from 1B in that the latter is associated with lower concentrations of Na, K, Ca, Mg, HCO_3 , Cl and F. Table 1 summarizes each biocluster's association with different chemical parameters in terms of relative concentration ranges derived from the absolute values shown in Figures 6 and S2[i-iv]. Overall, these results suggested that the groundwater bacterial community structure explained by the bioclusters has distinct relationships with groundwater chemistry.

Relationships between bacterial diversity and environmental factors

Cross-tabular representation demonstrated some interesting qualitative aspects of the relationships between bacterial community structure explained at 3-cluster threshold and aquifer confinement and groundwater bore usage (Fig. 7). The majority of sampling sites belonging to biocluster 1B were located in unconfined aquifers, whereas biocluster 2

contained the least number of sites in unconfined aquifers, and the highest number of sites in confined aquifers. The relationships between bioclusters and groundwater bore usage are not very distinct, as all three clusters contained groundwater bores used for potable, domestic, stock and irrigation purposes. However, it is interesting to note that the sites used for public purposes were not represented in biocluster 1B.

Mean Shannon diversity indices (H') for each biocluster indicated the presence of a considerable difference of bacterial diversity among the three bioclusters (Table 2). For both FAM and HEX OTUs, biocluster 1B represented the highest mean H' while bioclusters 1A and 2 showed medium and the lowest mean diversity indices (Fig. 8). Overall, the additional statistical approaches applied in this study do not provide additional insights into the relationships between groundwater bacterial diversity and hydrochemistry, yet they strongly support the major findings inferred from HCA. The nMDS pattern was shown to be highly comparable to the HCA clustering. The PERMANOVA results also confirmed this observation ($P=0.0001$). No significant relationships were found between nMDS clustering and aquifer confinement ($P=0.1407$) or bore usage ($P=0.3278$). The RELATE analysis confirmed that the groundwater chemistry is highly correlated with bacterial diversity explained by FAM T-RFs ($P=0.0184$). The CCA results suggested that, among the 30 hydrochemical parameters, $\text{NO}_3\text{-N}$, $\text{NO}_2\text{-N}$ and Fe were the main factors influencing the bacterial diversity represented by FAM T-RFs. Interestingly, $\text{NO}_3\text{-N}$, $\text{NO}_2\text{-N}$ and Fe are three major factors contributing to the redox condition of the groundwater. This indicates that bacterial diversity is mainly influenced by the redox potential of groundwater, as previously determined on the basis of HCA and the Kruskal-Wallis test. The results of nMDS, PERMANOVA, RELATE and CCA are therefore not displayed.

Table 1 Summary of the relative magnitudes of chemical parameters in each Biocluster at 3-cluster threshold.

Chemical parameter	Biocluster 1A	Biocluster 1B	Biocluster 2
Na	Medium	Low	High
K	Medium	Low	High
Mg	Medium	Low	High
Ca	Medium	Low	High
Pb	Not clear	Not clear	Not clear
Zn	Not clear	Not clear	Not clear
B	Medium	Low	High
HCO ₃	Medium	Low	High
Cl	Medium	Low	High
SO ₄	High	High	Low
NO ₃ -N	High	Medium	Low
NO ₂ -N	Low	Low	High
NH ₄ -N	Low	Low	High
PO ₄ -P	Low	Low	High
Fe	Low	Low	High
Mn	Low	Low	High
Br	Low	Low	High
F	Medium	Low	High
SiO ₂	High	Low	High
Total Dissolved Solids (TDS)	Medium	Low	High
Total Organic Carbon (TOC)	Not clear	Not clear	Not clear
Alkalinity	Medium	Low	High
Total hardness	Medium	Low	High
Total cations	Medium	Low	High
Total anions	Medium	Low	High
Free CO ₂	Not clear	Not clear	Not clear
DO	Medium	High	Low
Total Oxidized Nitrogen (TON)	High	Medium	Low
EC	Medium	Low	High
pH	Low	Low	High

Table 2 Summary of mean Shannon diversity indices (H') and standard deviations (SD) for each Biocluster, separately calculated for FAM and HEX T-RFs.

Bioclusters	FAM		HEX	
	Mean H'	SD	Mean H'	SD
Biocluster 1A	1.38	0.41	1.63	0.30
Biocluster 1B	1.74	0.60	1.70	0.59
Biocluster 2	1.12	0.36	0.84	0.38

Table 3 Summary of groundwater characteristics in each Biocluster.

Biocluster	Groundwater characteristics
Biocluster 1A	Oxidized water with possibly high human impact, Moderate to high bacterial diversity, Moderate TDS, Low pH, Low to moderate [Na], [K], [Mg], [Ca], [Cl], [HCO ₃] and [F], High alkalinity, Highest [SO ₄]
Biocluster 1B	Oxidized water with possibly low human impacted, Highest bacterial diversity, Lowest TDS, Low pH, Lowest [Na], [K], [Mg], [Ca], [Cl], [HCO ₃] and [F], High alkalinity, majority unconfined aquifers, Moderate to high [SO ₄]
Biocluster 2	Reduced water, Lowest bacterial diversity, Highest TDS, High pH, Highest [Na], [K], [Mg], [Ca], [Cl], [HCO ₃] and [F], High alkalinity, majority confined aquifers, Lowest [SO ₄],

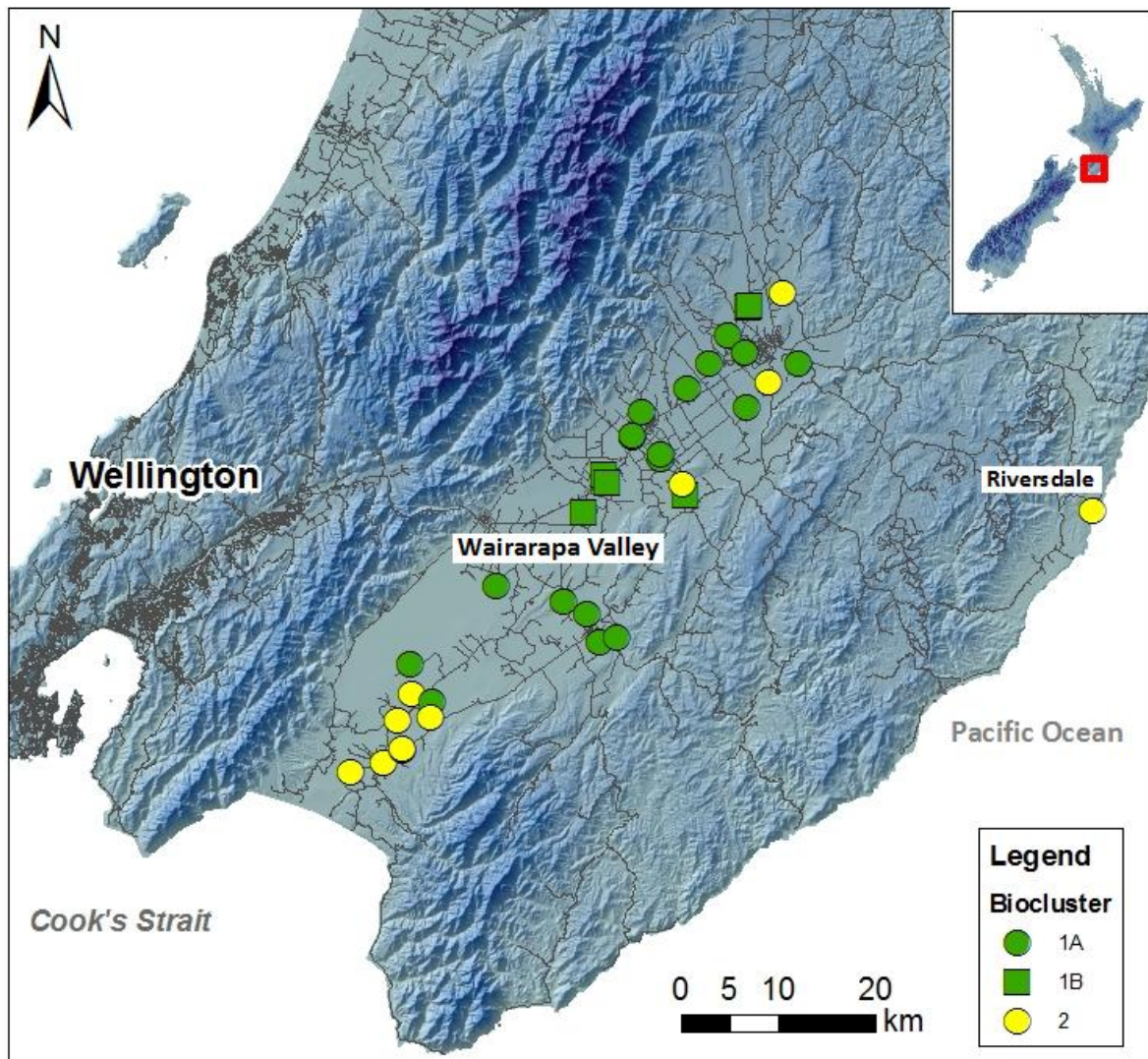


Figure 1 Groundwater sites sampled in the Wairarapa valley and the Riversdale area, New Zealand. These sites are grouped into bioclusters based on their bacterial diversity (see results section). Each site is represented with a relevant symbol in a specific colour and shape to match with the biocluster to which it belongs.

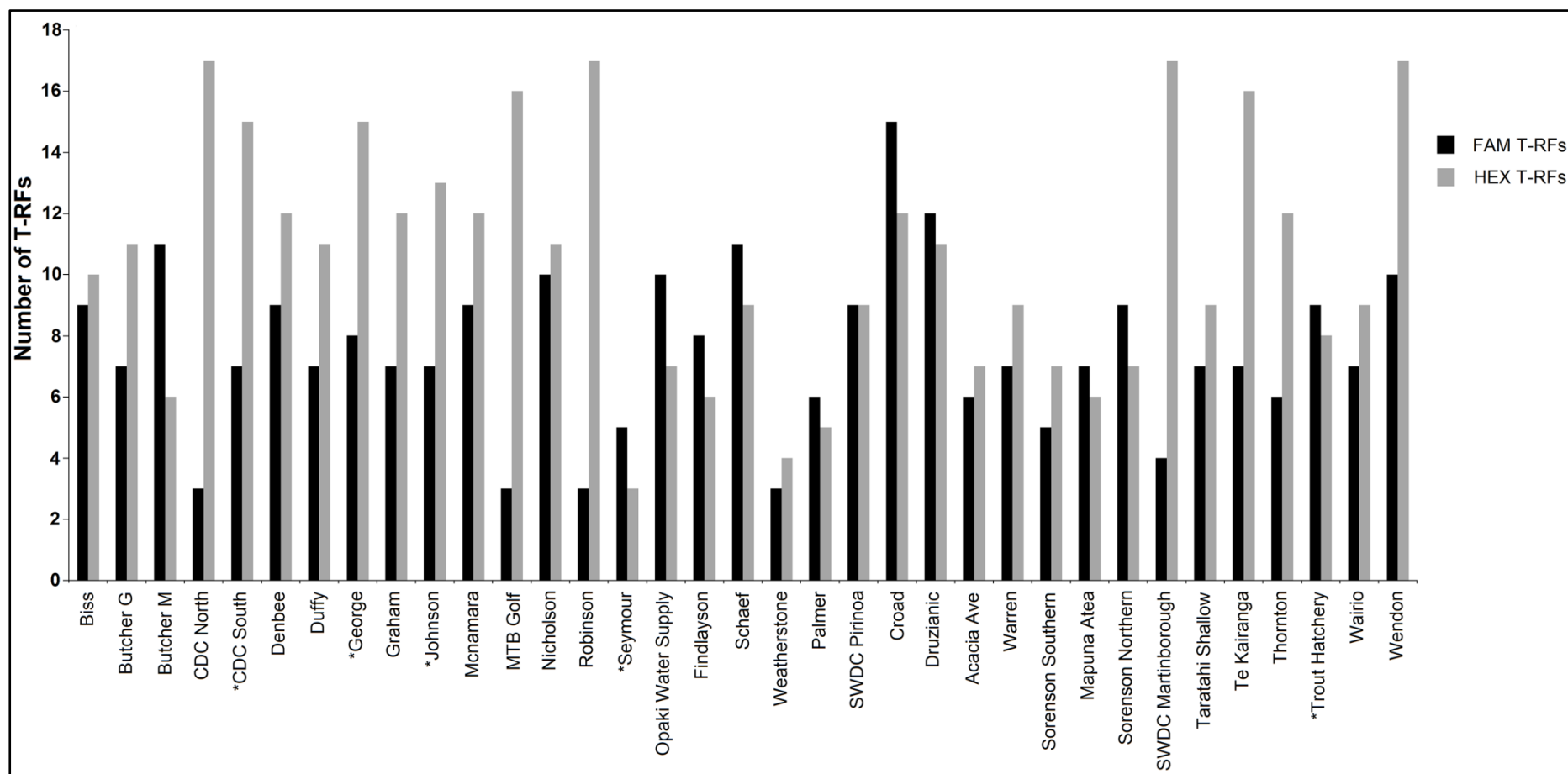


Figure 2 Summary of the total number of FAM (Black) and HEX (Grey) T-RFs over 200 RFU in each sample.

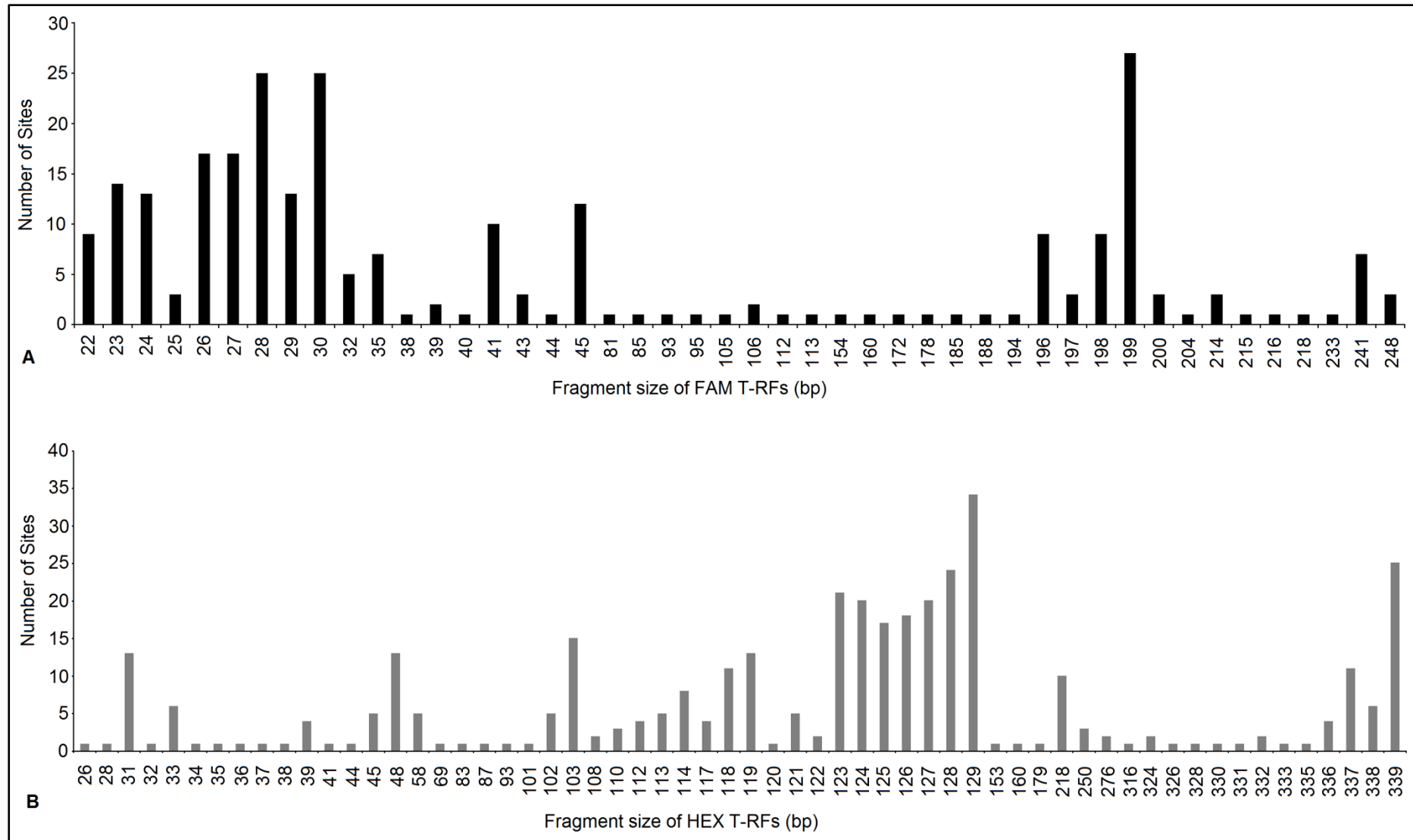


Figure 3 Summary of the frequency of each **(A)** FAM and **(B)** HEX T-RF (i.e. the number of sites at which each T-RF was detected).

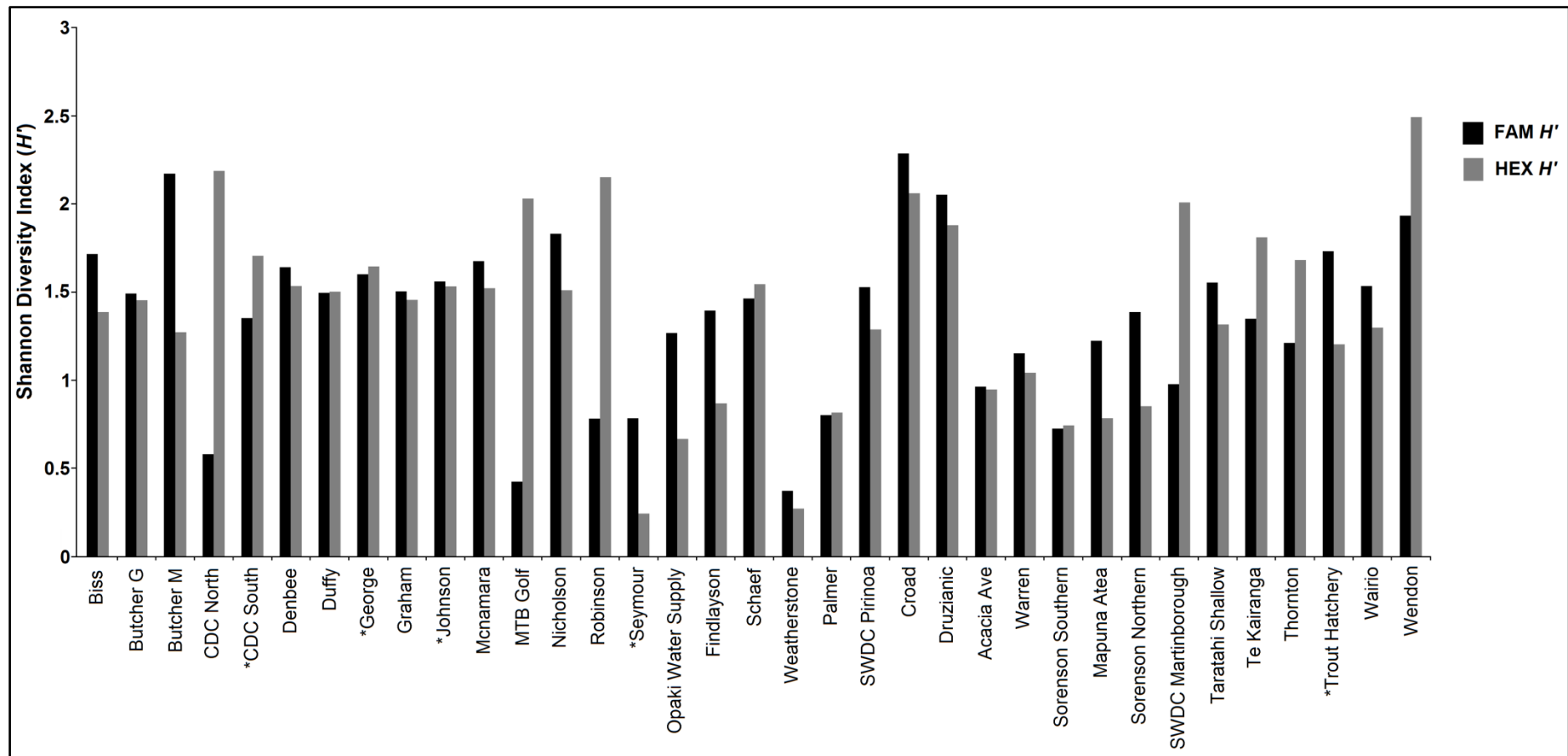


Figure 4 Summary of Shannon diversity index (H') values for each sample using FAM (Black) and HEX (Grey) OTUs.

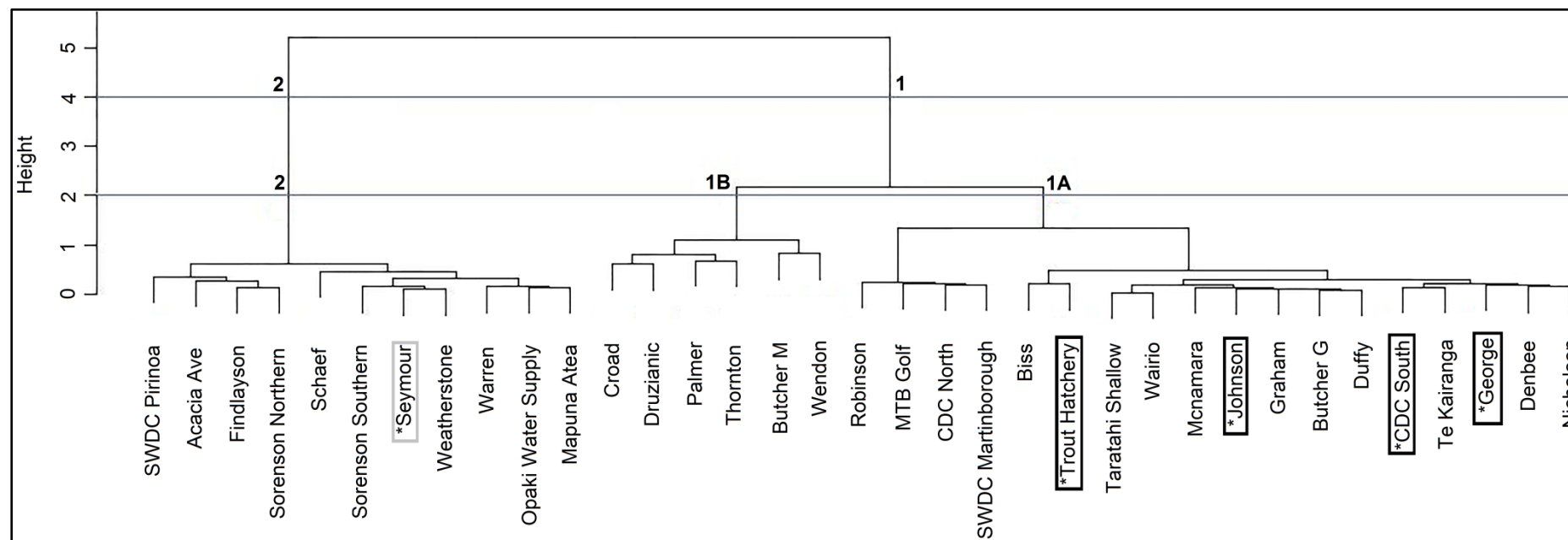


Figure 5 Dendrogram produced by hierarchical cluster analysis performed using Ward's linkage rule with FAM and HEX T-RFs standardized to the sum of all peaks in each profile and the Bray-Curtis similarity index. The five sites that were successfully tested in the previous study by van Bekkum et al. (2006) are labelled with the * symbol. Four sites: Trout Hatchery; Johnson; CDC South; and George indicated with boxes with black margin as they were clustered together in that study, in contrast to Seymour which was clustered separately and is indicated by a grey-margined box.

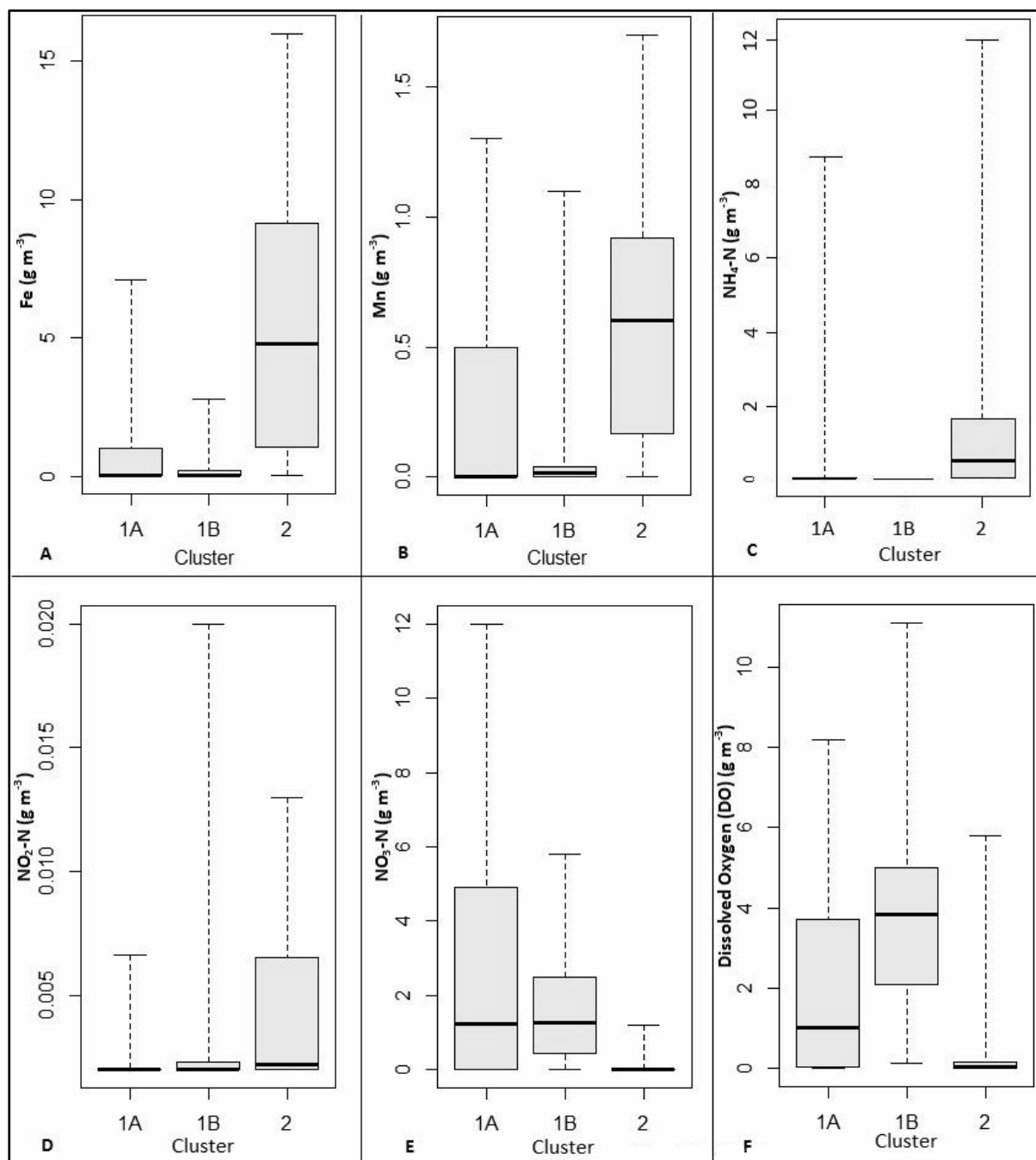


Figure 6 Box-and-Whisker Plot comparisons of concentrations of **(A)** Fe, **(B)** Mn, **(C)** $\text{NH}_4\text{-N}$, **(D)** $\text{NO}_2\text{-N}$, **(E)** $\text{NO}_3\text{-N}$ and **(F)** Dissolved Oxygen across bioclusters defined at the 3-cluster threshold.

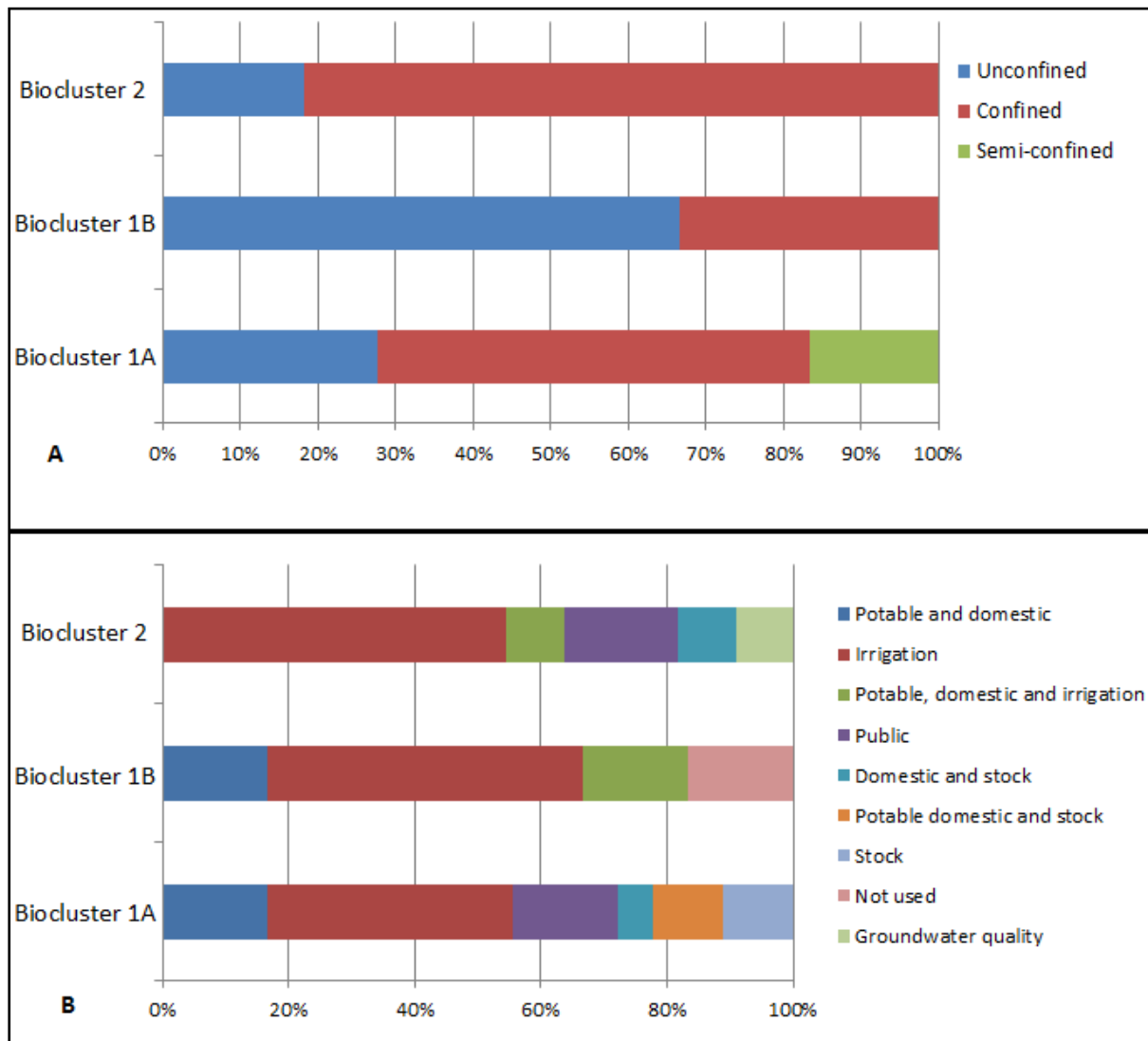


Figure 7 Percentage of samples in each biocluster defined at 3-cluster threshold as a function of **(A)** aquifer confinement and **(B)** groundwater bore usage.

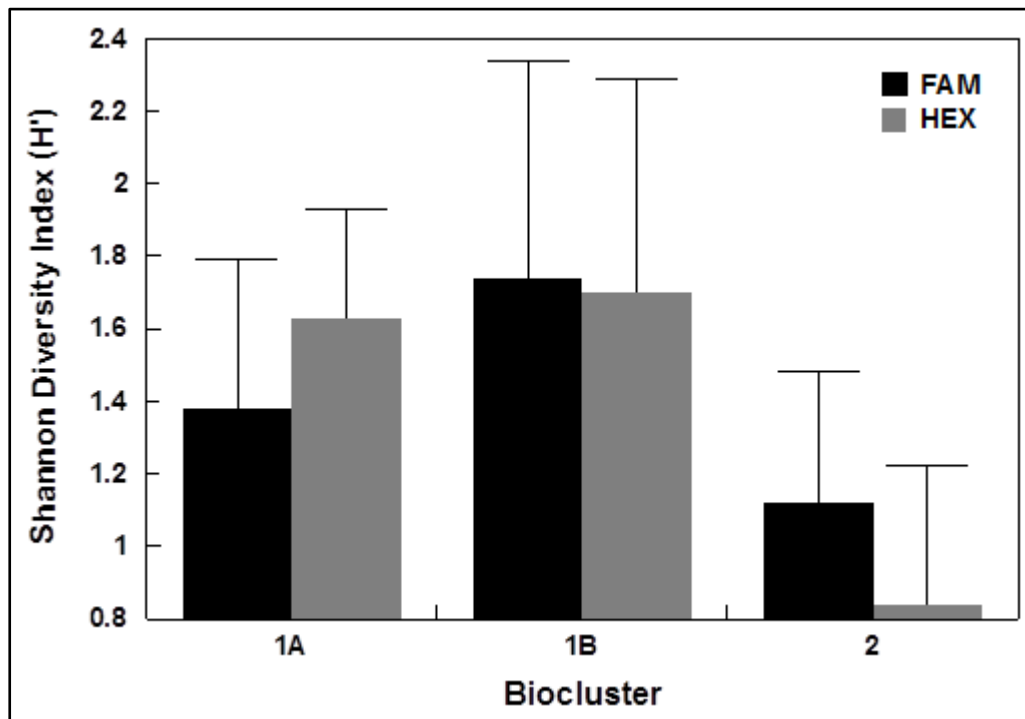


Figure 8 Summary of mean Shannon diversity index (H') values for each biocluster using FAM and HEX T-RFs. Bars represent the mean Shannon Index and the error bars represent one standard deviation.

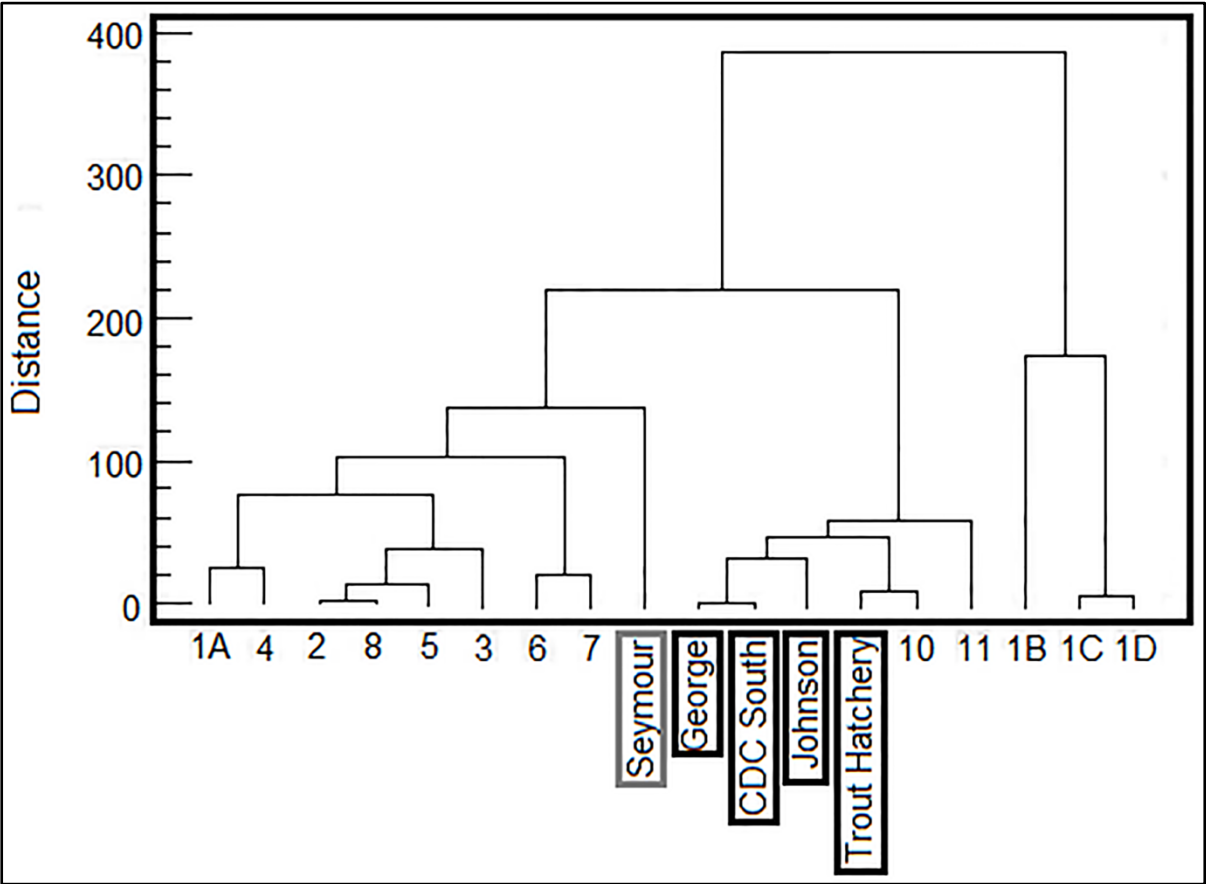


Figure 9 Hierarchical cluster analysis pattern for the five samples: Seymour; Trout Hatchery; Johnson; CDC South; and George revealed by van Bakkum et al. (2006). Two different coloured boxes (black and gray) were used to indicate the sites belonging to the two different clusters represented in Figure 5.

Discussion

The results of this study indicate that considerable bacterial diversity was present in Wairarapa Valley groundwater. The biocluster representation provided a useful framework for evaluating the relationships between bacterial diversity and the chemistry of the groundwater. These were clearly identifiable, in particular, for the redox-sensitive substances such as Fe, Mn, $\text{NO}_3\text{-N}$, $\text{NH}_4\text{-N}$ and SO_4 . Various species of bacteria can use different reduced forms of nitrogen (NH_4^+ , NO_2^-), manganese (Mn^{2+}), iron (Fe^{2+}) and other redox sensitive elements as reducing agents to reduce organic carbon, CO_2 (carbon fixation), oxygen or any other more oxidized forms of nitrogen (NO_3^-), manganese (MnO_4^-), iron (Fe^{3+}) and sulphur (SO_4^{2-}), through which they obtain energy (Chapelle 2000). The presence of high concentrations of $\text{NH}_4\text{-N}$, $\text{NO}_2\text{-N}$, Fe and Mn (reduced forms) and low concentrations of $\text{NO}_3\text{-N}$, SO_4 , total oxidized nitrogen (TON) and dissolved oxygen (oxidized forms) in samples grouped into biocluster 2 indicated that the groundwaters in this cluster were more reduced than groundwater found at sites assigned to bioclusters 1A and 1B. The three bioclusters can also be differentiated by the amounts of total dissolved solids (TDS); where biocluster 2 showed the highest TDS in contrast with bioclusters 1A and 1B which have medium and low TDS respectively. In addition, biocluster 1B reflected relatively low $\text{NO}_3\text{-N}$ and TON concentrations compared with biocluster 1A, possibly indicating that sites assigned to biocluster 1B are less impacted by human activities in the aquifer recharge zone.

Table 3 provides a summary of groundwater chemistry and aquifer features represented in each biocluster. Interestingly, the biocluster representation based on bacterial diversity was comparable to hydrochemical categories previously defined using groundwater chemistry, human impact and aquifer properties (Daughney & Reeves 2005 and see Table S4). Our results suggest that the three bioclusters might provide bacterial community fingerprints for the relevant hydrochemical categories. In other words, oxidised vs. reduced

and impacted vs. non-impacted groundwaters have characteristic bacterial populations, at least for the sites in the Wairarapa Valley that were sampled in this study. This is in agreement with the conclusions of the national survey of groundwater bacterial diversity by Sirisena et al. (2013). The bacterial diversity in each biocluster represented by mean (H') reveals that bioclusters 1A and 1B, with oxidized groundwater, have relatively high diversity compared with biocluster 2, which has a more reduced hydrochemical profile. This observation implies that the sites in bioclusters 1A and 1B may contain diverse bacterial groups such as sulphur oxidizers, nitrifying bacterial groups, iron oxidizers and hydrogen oxidizers which help to oxidize the reduced forms of the redox chemical components as described above. A majority of sites located in unconfined aquifers were assigned to bioclusters 1A and 1B, which contained oxidized groundwater. This is consistent with the relationship between groundwater chemistry and aquifer confinement previously noted by Daughney and Reeves (2005). The bioclusters were not strongly related to groundwater bore usage; although none of the sites belonging biocluster 1B were used for public purposes, this should not be taken to indicate a causal relationship between bacterial diversity and bore usage.

We were not able to collect samples from all the monitoring sites that were tested by van Bekkum et al. (2006). However, we were able to re-test five of their previously sampled sites: Seymour; Trout Hatchery; Johnson; CDC South; and George. This provided an opportunity for partial comparison of the bacterial community structure in Wairarapa groundwater in 2006 and 2009. In the previous study by van Bekkum et al. (2006), Seymour was clustered separately from the other four samples (Fig. 9). This pattern for these five samples remains similar in the present study, for all four T-RFLP data analysis combinations employed (Figs. 5 and S1). The study by van Bekkum et al. (2006) differs from ours in several technical aspects, e.g. T-RFLP was performed using tetrachloro-6-carboxy-

fluorescein (TET) as the fluorescence label for reverse primer R1389 at the 5' end, in contrast to 6-carboxyhexafluorescein (HEX) used in our study. In addition, they evaluated similarity between samples using the Common Area Index (CAI), whereas Euclidean distance and Bray-Curtis similarity were used in the present study. Again, we interpret the similarities of the two sets of result as providing further evidence to demonstrate that the T-RFLP technique is a robust and reliable molecular profiling tool, i.e. that it produces results that are largely independent of the data analysis methods and experimental conditions employed, at least for a study of the scale undertaken in this work. We note that all the chemical parameters remained very similar at each site in 2009 compared with 2006 (Table S5). This implies that the bacterial diversity of the groundwater may have remained constant over time in part because the groundwater chemistry also remained constant over time, although we acknowledge that more data are required to robustly determine the direction of causality of such relationships.

The T-RFLP methodology used in this work provides a reliable and rapid molecular profiling tool that can be used in future studies to investigate the bacterial community structure in groundwater. The comparison between hierarchical cluster analyses performed using four different combinations of data analysis approaches suggests that HCA appeared to be relatively insensitive to T-RF scaling method and distance measure used, at least for the data considered in this study. Therefore, HCA can be effectively used as a robust technique to compare similarities between T-RFLP profiles. However, it remains to be seen whether other T-RF datasets would also show the same properties with respect to the methods used for peak scaling and distance (similarity) measurement, and we recommend that a similar comparison of data analysis methods should be undertaken for future studies of this type. Similarity of dendrograms shows our data are robust regardless of data analysis method. Therefore, they are deemed to be resulting authentic signals of real biological significance. Indeed, this may be a

general feature of the T-RFLP approach which may go some way towards explaining why no single standardized method / distance metric has so far been adapted for widespread use.

We note that several different methods are available for statistical analysis of datasets such as that collected in this study. Following Sirisena et al. (2013), we employed HCA, box-whisker plots, the Kruskal-Wallis test and crosstabulation. For comparison, we also applied several independent techniques (nMDS, PERMANOVA, RELATE and CCA). The fact that the results were highly comparable gives credibility to the conclusions drawn. Given that this similarity of results from different techniques may not extend to other studies, we recommend that independent techniques should be applied for validation of inferences made in future investigations of this type.

We also note that different molecular profiling tools could be applied in such studies. It is hard to argue that one molecular profiling tool is better than the others as each technique has its own advantages and drawbacks. For example, the automated ribosomal intergenic spacer analysis (ARISA) approach is another popular molecular tool that can be effectively used in microbial community analyses (Lear et al. 2013; Washington et al. 2013). But one of the issues in the ARISA methodology is that it is possible to produce PCR products with same length for different species. The T-RFLP technique is susceptible to the same drawback as different species may generate T-RFs with same length. However, in T-RFLP, two fluorescently labelled primers can be used to minimize this problem by generating one or two signals for any one species. As this is a part of an integrated project investigating the bacterial diversity in New Zealand groundwater, we have used T-RFLP as our choice of method to be consistent with the past study by Sirisena et al. (2013). However, we suggest that it would be a good future prospect to compare T-RFLP and ARISA in terms of cost-effectiveness and information obtained.

As the primary objective of this study was strictly to understand the microbial state of groundwater itself, we did not attempt to analyse the microbial diversity of the aquifer materials. However, it is evident that bacterial diversities in groundwater and the aquifer, from which they are derived, may differ from each other and that the two communities may have mutual interactions (Alfreider et al. 1997; Lehman & O'Connell 2002; Flynn et al. 2008; Griebler & Lueders 2009). Therefore, a future extension of this study could be to evaluate bacterial communities of the aquifers themselves and attempt to gain a better understanding of their interactions with bacterial communities in the groundwater, taking these as being two distinct components of the groundwater ecosystem. We note however that it is much more cost effective to collect samples of groundwater than to collect aquifer materials directly, and characterisation of microbial community structure in the groundwater itself is likely most promising for routine State-of-the-Environment monitoring.

Although T-RFLP is a suitable technique for this sort of rapid explorative study, it does not provide taxonomic information about those bacterial species that are present in these groundwater ecosystems (Wood et al. 2013). In addition, the resolution of the technique may not be powerful enough to recognize the least abundant species in the environment (Pilloni et al. 2012). Thus, another future extension of this study could be to apply modern metagenomics approaches based on high-throughput DNA sequencing in an attempt to obtain taxonomic information and capture the microbial biodiversity that is not revealed by T-RFLP.

In summary, the findings of this study indicate that the bacterial diversity of groundwater is mainly related to groundwater chemistry. Further, the diversity is stable over timescales of a few years, at least when the groundwater chemistry also remains stable over the same period. These findings reflect the potential of using bacterial communities as biological indicators to evaluate the health of groundwater ecosystems, beyond what may be inferred from chemical or geological information alone. Therefore, we suggest that it would

be worthwhile to include broad bacterial diversity assessment criteria into regular groundwater monitoring activities, as opposed to the current practice whereby bacterial monitoring of groundwater is restricted to indicator species for faecal contamination.

List of supplementary files

Fig. S1: Dendrograms of the hierarchical cluster analyses performed with different combinations of of peak scaling methods and distance measures.

Fig. S2 (i): Box-and-Whisker Plot comparisons of concentrations of SO_4 , Total Dissolved Solids, Total Oxidized Nitrogen, Na, K and Mg across bioclusters defined at the 3-cluster threshold.

Fig. S2 (ii): Box-and-Whisker Plot comparisons of concentrations of Ca, B, HCO_3 , Cl, Br and F across bioclusters defined at the 3-cluster threshold.

Fig. S2 (iii): Box-and-Whisker Plot comparisons of concentrations of $\text{PO}_4\text{-P}$, SiO_2 , Alkalinity, Total hardness, Total cations and Total anions across bioclusters defined at the 3-cluster threshold.

Fig. S2 (iv): Box-and-Whisker Plot comparisons of Electrical conductivity, Acidity, Free CO_2 , Total organic carbon, concentrations of Pb and Zn across bioclusters defined at the 3-cluster threshold.

Table S1: Concentration values of 30 hydrochemical parameters at each groundwater monitoring site in the September 2009 sampling round.

Table S2: Summary of geographical location (in Northing and Easting), aquifer confinement and usage of groundwater of the GWRC sampling sites.

Table S3: Summary of P values (95.0% confidence level, $n=35$, d. f. = 34) of Kruskal-Wallis tests for each chemical parameter at the 3- and 2-Cluster thresholds.

Table S4: Typical characteristics of chemistry and human impact for hydrochemical categories at the 3-category threshold as defined by Daughney & Reeves (2005).

Table S5: Summary of groundwater chemistry at GWRC sampling sites included in both van Bekkum et al. (2006) study and present study.

Acknowledgments

The authors would like to thank Sheree Tidswell and all the groundwater research staff members of the Greater Wellington Regional Council for their valuable support in sample collection and for providing hydrochemical data and site-specific environmental information.

The authors would also like to thank Craig Cary (University of Waikato, Hamilton, New Zealand) and Els Mass (National Institute of Water and Atmospheric Research, Wellington, New Zealand) for their useful advice regarding T-RFLP data analyses, Dalice Sim (Victoria University of Wellington, Wellington, New Zealand) for advice on statistical analyses and the two anonymous reviewers for their constructive suggestions to improve the quality of the manuscript. This project was financially supported by public research funding provided to GNS Science by the Government of New Zealand.

The authors have no conflicts of interest.

References

- Alfreider A, Krossbacher M, Psenner R 1997. Groundwater samples do not reflect bacterial densities and activity in subsurface systems. *Water Research* 31: 832–840.
- Anderson MJ, Gorley RN, Clarke KR 2008. PERMANOVA+ for PRIMER: guide to software and statistical methods, Plymouth, UK, PRIMER-E Ltd.
- Blackwood CB, Marsh T, Kim SH, Paul EA 2003. Terminal restriction fragment length polymorphism data analysis for quantitative comparison of microbial communities. *Applied and Environmental Microbiology* 69: 926–932.
- Camarinha-Silva A, Wos-Oxley ML, Jáuregui R, Becker K, Pieper DH 2012. Validating T-RFLP as a sensitive and high-throughput approach to assess bacterial diversity patterns in human anterior nares. *FEMS Microbiology Ecology* 79: 98–108.
- Chapelle FH 2000. The significance of microbial processes in hydrogeology and geochemistry. *Hydrogeology Journal* 8: 41–46.
- Clarke KR, Warwick RM 2001. Change in marine communities: an approach to statistical analysis and interpretation, Plymouth, UK, PRIMER-E Ltd. Culman SW, Gauch HG, Blackwood CB, Thies JE 2008. Analysis of T-RFLP data using analysis of variance and ordination methods: a comparative study. *Journal of Microbiological Methods* 75: 55–63.
- Daughney CJ, Reeves RR 2005. Definition of hydrochemical facies in the New Zealand groundwater monitoring programme. *Journal of Hydrology (New Zealand)* 44: 105–130.
- Daughney CJ, Reeves RR 2006. Analysis of temporal trends in New Zealand's groundwater quality based on data from the National Groundwater Monitoring Programme. *Journal of Hydrology (New Zealand)* 45: 41–62.

- Daughney CJ, Jones A, Baker T, Hanson C, Davidson P, Zemansky GM, Reeves RR, Thompson M 2006. A national protocol for state of the environment groundwater sampling in New Zealand: ME 781. Wellington, New Zealand, Ministry for the Environment.
- Daughney CJ, Raiber M, Moreau-Fournier M, Morgenstern U, van der Raaij R 2012. Use of hierarchical cluster analysis to assess the representativeness of a baseline groundwater quality monitoring network: Comparison of New Zealand's national and regional groundwater monitoring programs. *Hydrogeology Journal* 20: 185–200.
- Dollhopf SL, Hashsham SA, Tiedje TM 2001. Interpreting 16S rDNA T-RFLP data: Application of self-organizing maps and principal component analysis to describe community dynamics and convergence. *Microbial Ecology* 42: 495–505.
- Edlund A, Soule T, Sjöling S, Jansson JK 2006. Microbial community structure in polluted Baltic Sea sediments. *Environmental Microbiology* 8: 223–232.
- Falkowski PG, Fenchel T, Delong EF 2008. Microbial engines that derived earth's biogeochemical cycles. *Science* 320: 1034–1038.
- Flynn TM, Sanford RA, Bethke CM 2008. Attached and suspended microbial communities in a pristine confined aquifer. *Water Resources Research* 44: W07425.
- Gilpin B, Devane M, Nourozi F, Robson B, Scholes P, Lin S 2013. Recommendations for the processing and storage of water samples before polymerase chain reaction (PCR) analysis. *New Zealand Journal of Marine and Freshwater Research* 47: 582–586.
- Greater Wellington Regional Council 2013. Groundwater quality. Wellington, Greater Wellington Regional Council.
- www.gw.govt.nz/groundwater-2/ [accessed 7 October 2013]
- Griebler C, Lueders T 2009. Microbial biodiversity in groundwater ecosystems. *Freshwater Biology* 54: 649–677.

- Griebler C, Stein H, Kellermann C, Berkhoff S, Brielmann H, Schmidt S, Selesi D, Steube C, Fuchs A, Hahn HJ 2010. Ecological assessment of groundwater ecosystems – vision or illusion? *Ecological Engineering* 36: 1174–1190.
- Janssen PH, Yates PS, Grinton BE, Taylor PM, Sait M 2002. Improved culturability of soil bacteria and isolation in pure culture of novel members of the divisions Acidobacteria, Actinobacteria, Proteobacteria, and Verrucomicrobia. *Applied and Environmental Microbiology* 68: 2391–2396.
- Korbel KL, Hose GC 2011. A tiered framework for assessing groundwater ecosystem health. *Hydrobiologia* 661:329–349.
- Korbel KL, Hancock PJ, Serov P, Lim RP, Hose GC 2013. Groundwater ecosystems vary with land use across a mixed agricultural landscape. *Journal of Environmental Quality* 42: 380–390.
- Kruskal JB 1964a. Multidimensional-Scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29: 1–27.
- Kruskal JB 1964b. Nonmetric Multidimensional Scaling – a numerical method. *Psychometrika* 29: 115–129.
- Larned ST 2012. Phreatic groundwater ecosystems: research frontiers for freshwater ecology. *Freshwater Biology* 57: 885–906.
- Lear G, Washington V, Neale M, Case B, Buckley H, Lewis G 2013. The biogeography of stream bacteria. *Global Ecology and Biogeography* 22: 544–554.
- Lehman RM, O’Connell SP 2002. Comparison of Extracellular Enzyme Activities and Community Composition of Attached and Free-Living Bacteria in Porous Medium Columns Comparison of Extracellular Enzyme Activities and Community Composition of Attached and Free-Living Bacteria in Porous Medium Columns. *Applied and Environmental Microbiology* 68: 1569–1575.

- Liu W, Marsh TL, Cheng H, LJ Forney 1997. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Applied and Environmental Microbiology* 63: 4516–4522.
- Lozupone CA, Knight R 2007. Global patterns in bacterial diversity. *Proceedings of the National Academy of Sciences USA* 104: 11436–11440.
- Ministry for the Environment 2010. Faecal pollution (bacteria) in groundwater. Wellington, Ministry for the Environment.
www.mfe.govt.nz/environmental-reporting/fresh-water/groundwater-quality-indicator/faecal-pollution.html [accessed 7 October 2013].
- Morgenstern U, Daughney CJ 2012. Groundwater age for identification of baseline groundwater quality and the impacts of land-use intensification - The National Groundwater Monitoring Programme of New Zealand. *Journal of Hydrology* 456/457: 79–93.
- Neufeld JD, Mohn WW 2005. Unexpectedly high bacterial diversity in arctic tundra relative to boreal forest soils, revealed by serial analysis of ribosomal sequence tags. *Applied and Environmental Microbiology* 71: 5710–5718.
- Parkinson R 2004. Forensic DNA profiling of bacterial communities in soil. Unpublished MSc thesis, Victoria University of Wellington, Wellington, New Zealand.
- Parkinson R 2009. Bacterial communities associated with human decomposition. Unpublished PhD thesis, Victoria University of Wellington, Wellington, New Zealand.
- Pilloni G, Granitsiotis MS, Engel M, Lueders T 2012. Testing the limits of 454 pyrotag sequencing: reproducibility, quantitative assessment and comparison to T-RFLP fingerprinting of aquifer microbes. *PLoS ONE* 7: e40467.
- Ramette A 2007. Multivariate analyses in microbial ecology. *FEMS Microbiology Ecology* 62: 142–160.

- Rosen MR 2001. Hydrochemistry of New Zealand's aquifers. In: Rosen MR, White PA eds. Groundwaters of New Zealand. Wellington, The New Zealand Hydrological Society. Pp. 77–110.
- Sirisena KA, Daughney CJ, Moreau-Fournier M, Ryan KG, Chambers GK 2013. National survey of molecular bacterial diversity of New Zealand groundwater: relationships between biodiversity, groundwater chemistry and aquifer characteristics. *FEMS Microbiology Ecology* 86: 490–504.
- Stein H, Kellermann C, Schmidt SI, Brielmann H, Steube C, Berkhoff SE, Fuchs A, Hahn HJ, Thulin B, Griebler C 2010. The potential use of fauna and bacteria as ecological indicators for the assessment of groundwater quality. *Journal of Environmental Monitoring* 12: 242–254.
- Steube C, Richter S, Griebler C 2009. First attempts towards an integrative concept for the ecological assessment of groundwater ecosystems. *Hydrogeology Journal* 17: 23–35.
- Struckmeier W, Rubin Y, Jones JAA 2005. Groundwater - reservoir for a thirsty planet?: Earth Sciences for Society; a Prospectus for a Key Theme of the International Year of Planet Earth. Trondheim, IUGS International Union of Geological Sciences Secretariat, Geological Survey of Norway.
- ter Braak CJF, Smilauer P 2012. Canoco reference manual and user's guide: software for ordination, version 5.0. Ithaca, USA, Microcomputer Power. 496 p.
- Tidswell S, Conwell C, Milne JR 2012. Groundwater quality in the Wellington region: state and trends. Publication No. GW/EMI-T-12/140. Wellington, New Zealand, Greater Wellington Regional Council.
- van Bekkum M, Sainsbury JP, Daughney CJ, Chambers GK 2006. Molecular analysis of bacterial communities in groundwaters from selected wells in the Hutt Valley and the

Wairarapa, New Zealand. *New Zealand Journal of Marine and Freshwater Research* 40: 91–106.

Washington VJ, Lear G, Neale MW, Lewis GD 2013. Environmental effects on biofilm bacterial communities: a comparison of natural and anthropogenic factors in New Zealand streams. *Freshwater Biology* 58:2277–2286. Wood S, Smith K, Banks J et al. 2013. Molecular genetic tools for environmental monitoring of New Zealand's aquatic habitats, past, present and the future. *New Zealand Journal of Marine and Freshwater Research* 47: 90–119.

Zhou Y, Kellermann C, Griebler C 2012. Spatio-temporal patterns of microbial communities in a hydrologically dynamic pristine aquifer. *FEMS Microbiology Ecology* 81: 230–242.

Zhou J, Davey ME, Figueras JB, Rivkina E, Gilichinsky D, Tiedje JM 1997. Phylogenetic diversity of a bacterial community determined from Siberian tundra soil DNA. *Microbiology* 143: 3913-3919.

List of figure captions

Figure 1 Groundwater sites sampled in the Wairarapa valley and the Riversdale area, New Zealand. These sites are grouped into bioclusters based on their bacterial diversity (see results section). Each site is represented with a relevant symbol in a specific colour and shape to match with the biocluster to which it belongs.

Figure 2 Summary of the total number of FAM (Black) and HEX (Grey) T-RFs over 200 RFU in each sample.

Figure 3 Summary of the frequency of each (A) FAM and (B) HEX T-RF (i.e. the number of sites at which each T-RF was detected).

Figure 4 Summary of Shannon diversity index (H') values for each sample using FAM (Black) and HEX (Grey) OTUs.

Figure 5 Dendrogram produced by hierarchical cluster analysis performed using Ward's linkage rule with FAM and HEX T-RFs standardized to the sum of all peaks in each profile and the Bray-Curtis similarity index. The five sites that were successfully tested in the previous study by van Bekkum et al. (2006) are labelled with the * symbol. Four sites: Trout Hatchery; Johnson; CDC South; and George indicated with boxes with black margin as they were clustered together in that study, in contrast to Seymour which was clustered separately and is indicated by a grey-margined box.

Figure 6 Box-and-Whisker Plot comparisons of concentrations of (A) Fe, (B) Mn, (C) $\text{NH}_4\text{-N}$, (D) $\text{NO}_2\text{-N}$, (E) $\text{NO}_3\text{-N}$ and (F) Dissolved Oxygen across bioclusters defined at the 3-cluster threshold.

Figure 7 Percentage of samples in each biocluster defined at 3-cluster threshold as a function of (A) aquifer confinement and (B) groundwater bore usage.

Figure 8 Summary of mean Shannon diversity index (H') values for each biocluster using FAM and HEX T-RFs. Bars represent the mean Shannon Index and the error bars represent one standard deviation.

Figure 9 Hierarchical cluster analysis pattern for the five samples: Seymour; Trout Hatchery; Johnson; CDC South; and George revealed by van Bekkum et al. (2006). Two different coloured boxes (red and blue) were used to indicate the sites belonging to the two different clusters represented in Figure 5.

Supplementary information

Table S1 Concentration values of 30 hydrochemical parameters at each groundwater monitoring site in the September 2009 sampling round. Units are in g m⁻³ for all variables except electrical conductivity (EC) which is in $\mu\text{S cm}^{-1}$ at 25 °C and pH in pH units.

Site Name	Site No:	Na	K	Mg	Ca	Pb	Zn	B	HCO3	Cl	SO4	NO2-N	NO3-N	NH4-N	PO4-P	Fe
Butcher G	S26/0117	11	2.6	3.4	10	4E-04	0.005	0.02	28	13	8.5	<0.002	3.9	<0.01	0.02	<0.02
Nicholson	S26/0223	13	1.1	5.9	12	4E-04	0.006	0.01	17	18	16	<0.002	10	<0.01	0.02	<0.02
Graham	S26/0299	7.6	0.8	2.2	6	4E-04	0.008	0.01	18	9.4	9.9	<0.002	2.1	<0.01	0.02	<0.02
Palmer	S26/0457	4.2	0.8	1.4	8	1E-04	0.018	0.02	28	6.6	5.4	<0.002	0.44	<0.01	0.01	<0.02
Denbee	S26/0568	22	1.2	9.3	21	<0.0001	0.003	0.02	150	12	<0.5	0.002	0	0.41	0.91	3.7
Mcnamara	S26/0576	23	1.1	6.3	16	<0.0001	0.002	0.02	110	19	2.5	0.002	0.01	0.49	0.72	3.3
*CDC South	S26/0705	16	1	3.7	9	2E-04	0.006	0.03	40	12	9.2	<0.002	4.9	<0.01	0.02	<0.02
Wendon	S26/0756	17	1.7	5.9	29	<0.0001	0.025	0.03	110	25	12	0.002	0.01	0.04	0.05	2.8
Schaefer	S26/0762	35	2	6.5	22	<0.0001	0.002	0.05	130	44	1.9	0.009	0.01	0.51	0.01	8.7
CDC North	S26/0824	14	1.1	4.2	9	2E-04	0.01	0.04	38	13	10	<0.002	5.2	<0.01	0.02	<0.02
Druzianic	S26/0846	6.9	0.7	1.8	8	3E-04	0.002	0.01	38	6.4	3.7	<0.002	0.72	<0.01	0.01	0.21
Croad	S27/0202	9.7	1.1	3.4	7	<0.0001	0.011	0.01	16	16	15	<0.002	2.5	<0.01	0.02	<0.02
*Johnson	S27/0299	7.9	0.8	2.3	7	3E-04	0.004	0.02	38	8.6	4.1	<0.002	0.31	<0.01	0	0.03
*George	S27/0344	31	1.7	7	17	<0.0001	0.043	0.04	72	58	11	<0.002	0	0.08	0.06	1
SWDC Martinborough	S27/0396	29	2.4	9.3	75	9E-04	0.016	0.04	270	36	32	<0.002	0.33	<0.01	0.02	0.03
Mapuna Atea	S27/0433	74	8.1	24	52	1E-04	0.005	0.09	360	93	<0.5	0.005	0.01	7.8	<0.004	13
Wairio	S27/0435	34	4.1	7.1	15	4E-04	0.013	0.09	150	34	<0.5	0.002	0.02	8.7	5.1	7.1
Robinson	S27/0442	130	1	3	8	6E-04	0.016	0.18	200	97	<0.5	<0.002	<0.002	0.93	4	1
MTB Golf	S27/0571	29	1.7	6.6	14	2E-04	0.014	0.02	46	33	11	<0.002	8.8	<0.01	0.01	<0.02
SWDC Pirinoa	S27/0588	15	1.3	4	10	3E-04	0.006	0.03	53	28	9	0.004	0.01	0.08	0.14	4.8
Warren	S27/0594	59	2.5	11	36	<0.0001	0.001	0.08	190	79	<0.5	<0.002	0	0.71	0.58	1.7
Weatherstone	S27/0602	41	5.3	8.2	23	1E-04	0.003	0.06	140	62	<0.5	<0.002	<0.002	2.6	2.1	3.1

CHAPTER 3.2

Findlayson	S27/0607	170	8.2	22	52	4E-04	0.01	0.15	250	330	<0.5	<0.002	0.01	12	<0.004	16
Sorenson Southern	S27/0614	48	3.6	8.5	24	3E-04	0.07	0.05	150	65	4.3	0.008	0.01	0.79	0.77	5.5
Sorenson Northern	S27/0615	39	2.6	8.9	19	<0.0001	0.008	0.04	96	60	18	0.013	<0.002	0.5	0.8	9.6
Te Kairanga	S27/0681	25	1.9	6.3	54	1E-04	0.005	0.03	170	32	32	<0.002	0.33	<0.01	0.01	<0.02
Biss	T26/0087	10	1.2	2.5	11	2E-04	0.061	0.02	49	9.1	5.7	0.002	1.7	<0.01	0.02	0.06
Butcher M	T26/0099	13	1.2	5.2	8	2E-04	0.003	0.01	37	11	9.1	<0.002	5.8	<0.01	0.02	<0.02
Thornton	T26/0206	18	0.8	3.9	9	4E-04	0.004	0.02	71	11	1.6	0.02	1.8	<0.01	0.06	<0.02
Opaki Water Supply	T26/0259	7.3	0.8	1.6	9	2E-04	0.003	0.01	32	8.6	6.5	<0.002	1.2	<0.01	0.01	<0.02
Taratahi Shallow	T26/0332	18	1	4.6	11	<0.0001	0.009	0.02	49	32	13	0.007	0.77	<0.01	0.04	<0.02
*Seymour	T26/0413	16	0.9	5.4	13	3E-04	0.005	0.02	97	9.6	<0.5	<0.002	<0.002	0.05	0.03	0.38
*Trout Hatchery	T26/0430	7.6	1.4	2.7	10	<0.0001	0.001	0.02	34	8.8	5.7	<0.002	2	<0.01	0.01	<0.02
Duffy	T26/0489	19	1.5	4.8	25	2E-04	0.032	0.02	58	25	13	<0.002	12	<0.01	0.01	<0.02
Acacia Ave	T27/0063	37	1.5	6.3	70	<0.0001	<0.001	0.04	250	51	15	0.002	1.2	<0.01	0.08	<0.02

*Sites tested in the previous study by van Bekkum et al. (2006).

Table S1 Continued.

Site Name	Site No:	Mn	Br	F	SiO2	TDS	TOC	Alkalinity	Total Hardness	Total Cations	Total Anions	DO	TON	Free CO2	pH	EC
Butcher G	S26/0117	0.002	<0.05	0.06	14	99	1.9	23	38	1.3	1.3	3.71	3.9	30	6.1	149
Nicholson	S26/0223	0.001	<0.05	<0.05	17	140	0.5	14	55	1.7	1.9	4.98	10	28	5.9	215
Graham	S26/0299	0.001	<0.05	<0.05	13	81	1.5	15	25	0.8	0.92	7.69	2.1	24	5.8	106
Palmer	S26/0457	0.001	<0.05	<0.05	9	57	1.7	23	25	0.7	0.79	3.98	0.44	11	6.3	84.5
Denbee	S26/0568	0.88	0.16	0.11	37	180	0.9	130	91	3	3	0.01	0.007	14	7.4	290
Mcnamara	S26/0576	0.63	0.08	0.1	36	150	1.2	86	65	2.5	2.4	0.05	0.01	15	7.2	241
*CDC South	S26/0705	6E-04	<0.05	0.14	26	110	0.8	33	37	1.4	1.6	0.92	4.9	25	6.2	173
Wendon	S26/0756	1.1	0.29	0.08	17	170	2	94	97	2.9	2.9	0.12	0.012	42	6.8	295
Schaef	S26/0762	1	0.17	0.18	25	200	4.1	110	81	3.6	3.5	0.01	0.016	50	6.8	375

CHAPTER 3.2

CDC North	S26/0824	0.001	<0.05	0.13	24	120	1.3	31	41	1.5	1.6	1.8	5.2	29	6	177
Druzianic	S26/0846	0.025	<0.05	<0.05	13	60	1.6	31	26	0.9	0.93	2.1	0.72	5.8	6.7	96.6
Croad	S27/0202	0.002	<0.05	0.06	14	88	1.3	13	32	1.1	1.2	3.67	2.5	24	5.8	141
*Johnson	S27/0299	9E-04	<0.05	0.09	14	60	2.1	31	27	0.9	0.97	0.19	0.32	9.2	6.4	104
*George	S27/0344	0.5	0.16	0.12	24	190	1.8	59	72	2.9	3	0.03	0.004	36	6.3	349
SWDC Martinborough	S27/0396	0.012	<0.05	0.08	11	330	2.2	220	230	5.8	6.1	0.01	0.33	34	6.9	573
Mapuna Atea	S27/0433	1.7	0.34	0.18	39	450	8.7	290	230	9.1	8.5	0.02	0.012	58	6.8	898
Wairio	S27/0435	0.53	0.1	0.3	38	210	6.6	120	66	3.8	4	0.01	0.018	27	6.7	398
Robinson	S27/0442	0.15	0.24	0.48	27	370	1.8	160	32	6.3	6.4	0.02	<0.002	8.3	7.4	670
MTB Golf	S27/0571	0.002	0.2	0.24	27	190	2	37	62	2.5	2.5	1.13	8.8	11	6.5	285
SWDC Pirinoa	S27/0588	0.15	<0.05	0.09	14	120	1.2	43	41	1.7	1.8	0.12	0.011	11	6.9	205
Warren	S27/0594	0.25	0.32	0.27	26	310	2.2	160	130	5.4	5.5	0	0.004	11	7.4	604
Weatherstone	S27/0602	0.6	0.14	0.21	48	270	2.3	110	91	4.1	4.2	0.01	<0.002	14	7	446
Findlayson	S27/0607	1.4	1.6	0.28	38	760	11	210	220	13	13	0.22	0.01	67	6.8	1543
Sorenson Southern	S27/0614	0.84	0.13	0.17	48	280	1.5	120	95	4.4	4.4	0.02	0.014	21	6.9	480
Sorenson Northern	S27/0615	0.68	0.15	0.22	34	230	1.7	79	83	3.8	3.7	0.03	0.011	23	6.7	413
Te Kairanga	S27/0681	<0.0005	<0.05	0.11	9	260	1.5	140	160	4.3	4.5	1.28	0.33	27	7	445
Biss	T26/0087	0.005	<0.05	0.05	12	87	1.5	40	37	1.2	1.3	3.98	1.7	24	6.3	138
Butcher M	T26/0099	0.001	0.14	0.07	19	120	1	31	42	1.4	1.5	5	5.8	14	6.2	167
Thornton	T26/0206	0.04	0.09	0.07	15	110	0.7	58	38	1.6	1.6	11.1	1.8	4.2	7.3	169
Opaki Water Supply	T26/0259	<0.0005	<0.05	<0.05	8.8	58	1.4	26	28	0.9	0.99	4.91	1.2	11	6.1	105
Taratahi Shallow	T26/0332	1.3	0.09	0.2	40	150	2.2	41	47	1.8	2	0.09	0.78	46	6.2	224
*Seymour	T26/0413	0.18	0.06	0.08	21	120	1.2	80	54	1.8	1.9	0.02	<0.002	9.8	7.2	188
*Trout Hatchery	T26/0430	<0.0005	<0.05	0.07	12	67	2.3	28	37	1.1	1.1	8.17	2	28	7	117
Duffy	T26/0489	1E-03	0.17	0.07	32	220	1.6	48	83	2.5	2.8	2.19	12	32	6.4	298
Acacia Ave	T27/0063	<0.0005	0.22	<0.05	6.6	330	2	210	200	5.7	6	5.8	1.2	9.1	7.7	578

*Sites tested in the previous study by van Bekkum et al. (2006).

Table S2 Summary of geographical location (in Northing and Easting), aquifer confinement and usage of groundwater of the GWRC sampling sites.

Site Name	Easting	Northing	Aquifer confinement	Bore Usage
Butcher G	2721500	6018500	Unconfined	Potable and domestic
Nicholson	2726219	6021005	Unconfined	Potable and domestic
Graham	2728370	6023590	Unconfined	Potable and domestic
Palmer	2717675	6012051	Unconfined	Potable, domestic and irrigation
Denbee	2723504	6013642	Confined	Irrigation
Mcnamara	2723479	6014255	Confined	Irrigation
*CDC South	2720489	6015999	Confined	Public
Wendon	2725937	6010018	Confined	Irrigation
Schaeff	2725720	6011070	Confined	Domestic and stock
CDC North	2720564	6016101	Confined	Public
Druzianic	2717921	6011212	Confined	Not used
Croad	2715480	6008240	Unconfined	Irrigation
*Johnson	2706525	6000655	Confined	Irrigation
*George	2713369	5999061	Confined	Irrigation
SWDCMartinborough	2715880	5997683	Confined	Public
Mapuna Atea	2697716	5989557	Confined	Irrigation
Wairio	2697631	5992523	Confined	Stock
Robinson	2699915	5988602	Confined	Potable domestic and stock
MTB Golf	2717180	5994736	Confined	Irrigation
SWDC Pirinoa	2694869	5982431	Confined	Public
Warren	2691376	5981438	Confined	Irrigation
Weatherstone	2699650	5987020	Confined	Irrigation
Findlayson	2696313	5986755	Confined	Irrigation
Sorenson Southern	2696803	5983642	Confined	Irrigation
Sorenson Northern	2696830	5983876	Confined	Irrigation
Te Kairanga	2718974	5995264	Unconfined	Irrigation
Biss	2730310	6026470	Semi-Confined	Potable domestic and stock
Butcher M	2732532	6029339	Unconfined	Potable and domestic
Thornton	2732595	6029549	Unconfined	Irrigation
Opaki Water Supply	2736010	6030840	Unconfined	Public
Taratahi Shallow	2732246	6019123	Semi-Confined	Domestic and stock
*Seymour	2734500	6021700	Confined	Potable, domestic and irrigation
*Trout Hatchery	2732145	6024748	Unconfined	Stock
Duffy	2737585	6023576	Semi-Confined	Irrigation
Acacia Ave	2768035	6008362	Unconfined	Groundwater quality

The Northing and Easting are in NZ Map Grid 1949.

*Sites tested in the previous study by van Bekkum et al. (2006).

Table S3 Summary of *P* values (95.0% confidence level, *n*=35, d. f. = 34) of Kruskal-Wallis tests for each chemical parameter at the 3- and 2-Cluster thresholds.

Parameter	<i>P</i> values	
	3-Cluster	2-Cluster
Na	.006	.006
K	.019	.016
Mg	.019	.013
Ca	.015	.013
Pb	.535	.454
Zn	.307	.135
B	.009	.010
HCO₃	.020	.012
Cl	.008	.007
SO ₄	.195	.074
NO₃-N	.007	.002
NO ₂ -N	.121	.056
NH₄-N	.004	.001
PO ₄ -P	.656	.594
Fe	.008	.002
Mn	.128	.048
Br	.076	.024
F	.036	.117
SiO ₂	.178	.194
Total Dissolved Solids (TDS)	.015	.016
Total Organic Carbon (TOC)	.191	.130
Alkalinity	.022	.012
Total Hardness	.026	.014
Total Cations	.010	.009
Total Anions	.011	.012
Free CO ₂	.222	.643
Dissolved Oxygen (DO)	.034	.041
Total Oxidized Nitrogen (TON)	.018	.005
Electrical Conductivity (EC)	.008	.009
pH	.128	.045

Values in bold show statistical significance ($P < 0.05$) in the relationships between chemical parameters and Bioclusters.

CHAPTER 3.2

Table S4 Summary of groundwater chemistry at GWRC sampling sites included in both van Bakkum et al. (2006) study and present study.

Units are in g m^{-3} for all variables except pH which is in pH units and Electrical conductivity (EC) in $\mu\text{S cm}^{-1}$ at 25 °C. N/A indicates that the parameter value is not available.

Chemical parameter	George		Johnson		Seymour		Trout Hatchery		CDC South	
	2006	2009	2006	2009	2006	2009	2006	2009	2006	2009
Na	33	31	7.5	7.9	16.5	16	7.7	7.6	17.1	16
K	1.7	1.7	0.8	0.78	0.9	0.89	1.2	1.4	1.1	1
Mg	7.2	7	2.4	2.3	5.2	5.4	2.2	2.7	3.9	3.7
Ca	17.2	17	6.4	6.9	13	13	9.7	10	9.8	8.6
Pb	N/A	<0.0001	N/A	0.0003	N/A	0.00031	N/A	<0.0001	N/A	0.00016
Zn	N/A	0.043	N/A	0.0035	N/A	0.0047	N/A	0.0011	N/A	0.0056
B	N/A	0.035	N/A	0.02	N/A	0.018	N/A	0.02	N/A	0.034
HCO ₃	69.8	72	34.8	38	94	97	31.8	34	39.1	40
Cl	54.9	58	8	8.6	9.5	9.6	7	8.8	12.5	12
SO ₄	11.2	11	4	4.1	0.7	<0.5	4.6	5.7	8.7	9.2
NO ₃ -N	< 0.01	0.0025	0.15	0.31	<0.01	<0.002	2.23	2	4.97	4.9
NO ₂ -N	N/A	<0.002	N/A	<0.002	N/A	<0.002	N/A	<0.002	N/A	<0.002
NH ₄ -N	0.08	0.082	<0.01	<0.01	0.04	0.051	<0.01	<0.01	<0.01	<0.01
PO ₄	N/A	0.059	N/A	0.0045	N/A	0.028	N/A	0.014	N/A	0.022
Fe	0.87	1	0.03	0.031	0.27	0.38	0.03	<0.02	0.03	<0.02
Mn	0.37	0.5	<0.05	0.00091	0.12	0.18	<0.05	<0.0005	<0.05	0.00059
Br	N/A	0.16	N/A	<0.05	N/A	0.06	N/A	<0.05	N/A	<0.05
F	N/A	0.12	N/A	0.094	N/A	0.08	N/A	0.068	N/A	0.14
SiO ₂	24	24	14.4	14	21.2	21	12.1	12	26.1	26
Total Dissolved Solids (TDS)	N/A	190	N/A	60	N/A	120	N/A	67	N/A	110
Total Organic Carbon (TOC)	N/A	1.8	N/A	2.1	N/A	1.2	N/A	2.3	N/A	0.75
Alkalinity	N/A	59	N/A	31	N/A	80	N/A	28	N/A	33
Total Hardness	N/A	72	N/A	27	N/A	54	N/A	37	N/A	37
Total Cations	N/A	2.9	N/A	0.9	N/A	1.8	N/A	1.1	N/A	1.4

CHAPTER 3.2

Total Anions	N/A	3	N/A	0.97	N/A	1.9	N/A	1.1	N/A	1.6
Free CO2	N/A	36	N/A	9.2	N/A	9.8	N/A	28	N/A	25
Dissolved Oxygen (DO)	N/A	0.03	N/A	0.19	N/A	0.02	N/A	8.17	N/A	0.92
Total Oxidised Nitrogen (TON)	N/A	0.0036	N/A	0.32	N/A	<0.002	N/A	2	N/A	4.9
Electrical Conductivity (EC)	296	349	86	104	186	188	99	117	169	173
pH	6.2	6.33	6.5	6.39	7.4	7.21	6.1	7.04	6.3	6.17

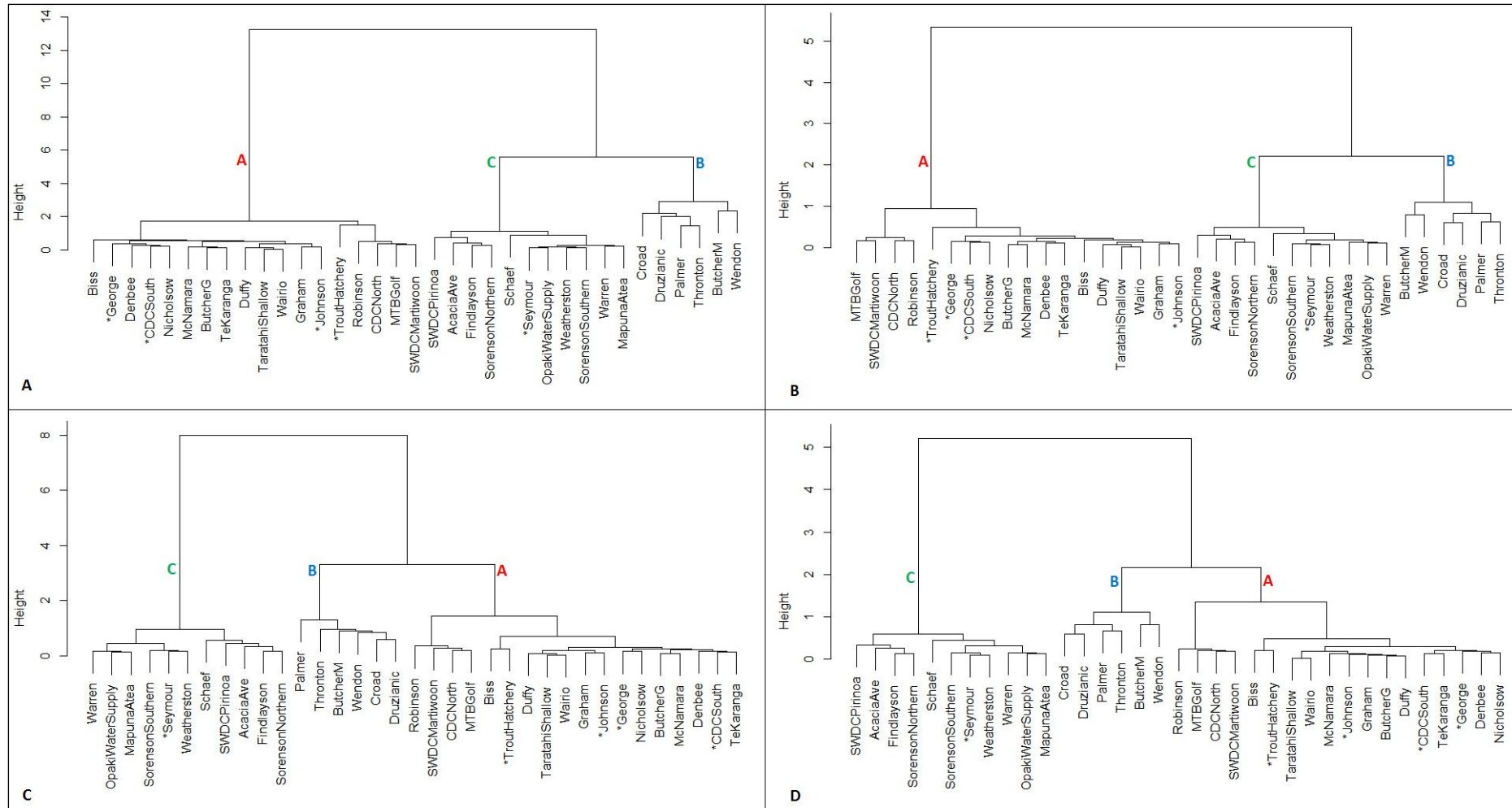


Fig. S1 Dendrograms of the hierarchical cluster analyses performed with different combinations of peak scaling methods and distance measures: **A**, T-RFs standardized to the highest peak in each profile / Euclidean distance; **B**, T-RFs standardized to the highest peak in each profile / Bray-Curtis similarity; **C**, T-RFs standardized to the sum of all peaks in each profile / Euclidean distance; and **D**, T-RFs standardized to the sum of all peaks in each profile / Bray-Curtis similarity. Sites tested in the previous study by van Bekkum et al. (2006) are labelled with * symbol.

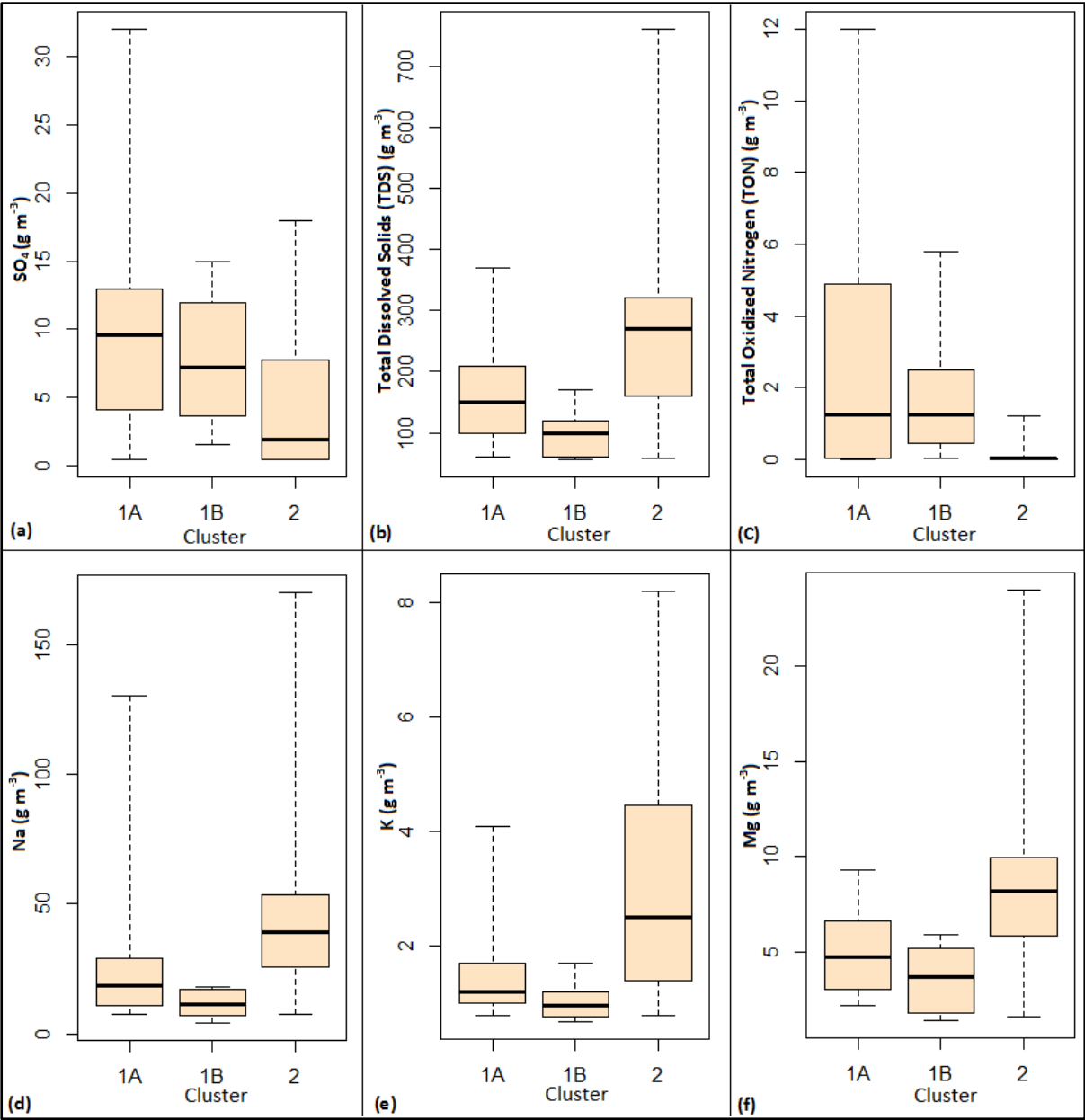


Fig. S2 (i) Box-and-Whisker Plot comparisons of concentrations of SO_4 (a), Total Dissolved Solids (b), Total Oxidized Nitrogen (c), Na (d), K (e) and Mg (f) across Bioclusters defined at the 3-cluster threshold.

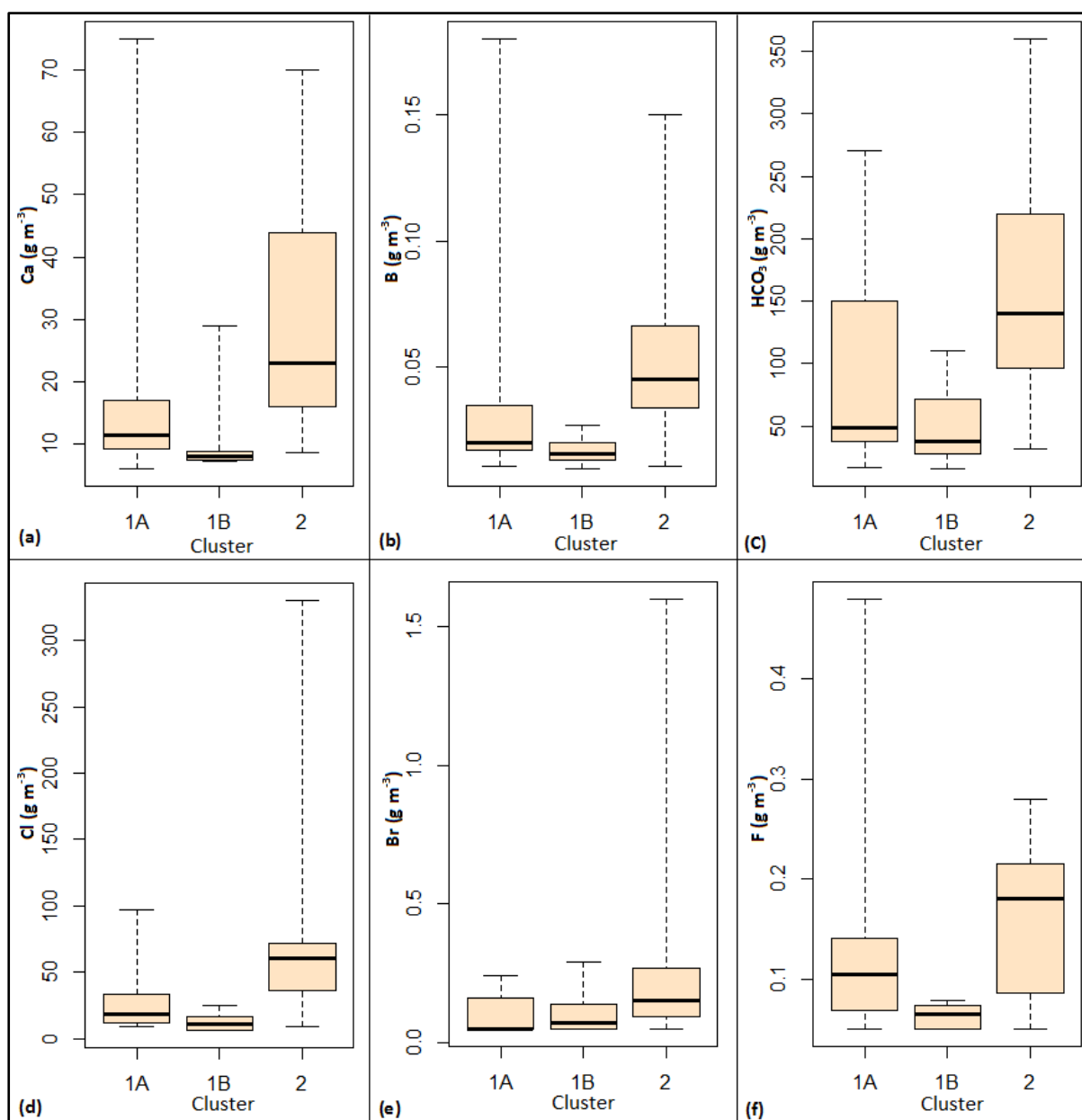


Fig. S2 (ii) Box-and-Whisker Plot comparisons of concentrations of Ca (a), B (b), HCO_3 (c), Cl (d), Br (e) and F (f) across Bioclusters defined at the 3-cluster threshold.

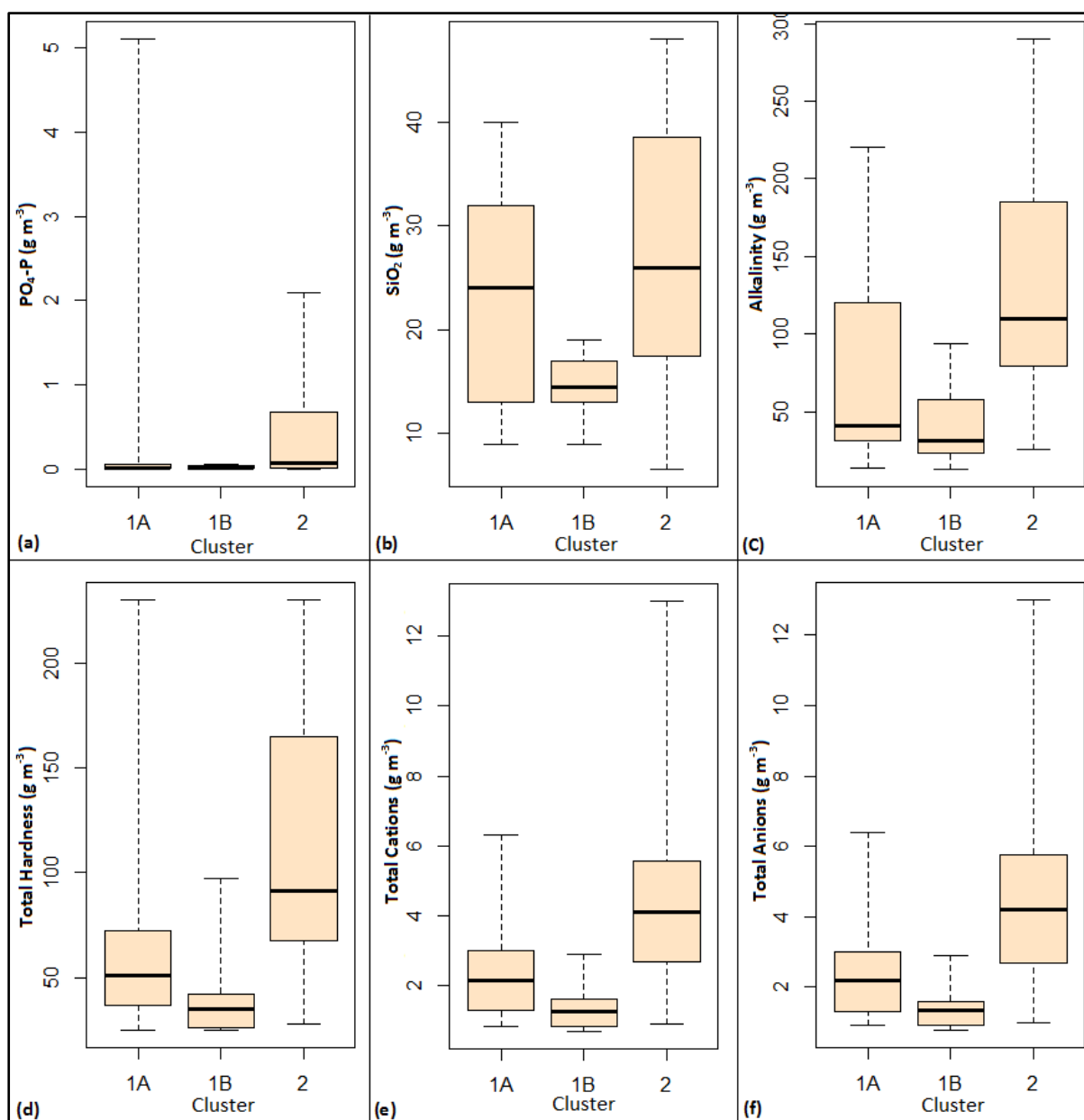


Fig. S2 (iii) Box-and-Whisker Plot comparisons of concentrations of $\text{PO}_4\text{-P}$ (a), SiO_2 (b), Alkalinity (c), Total hardness (d), Total cations (e) and Total anions (f) across Bioclusters defined at the 3-cluster threshold.

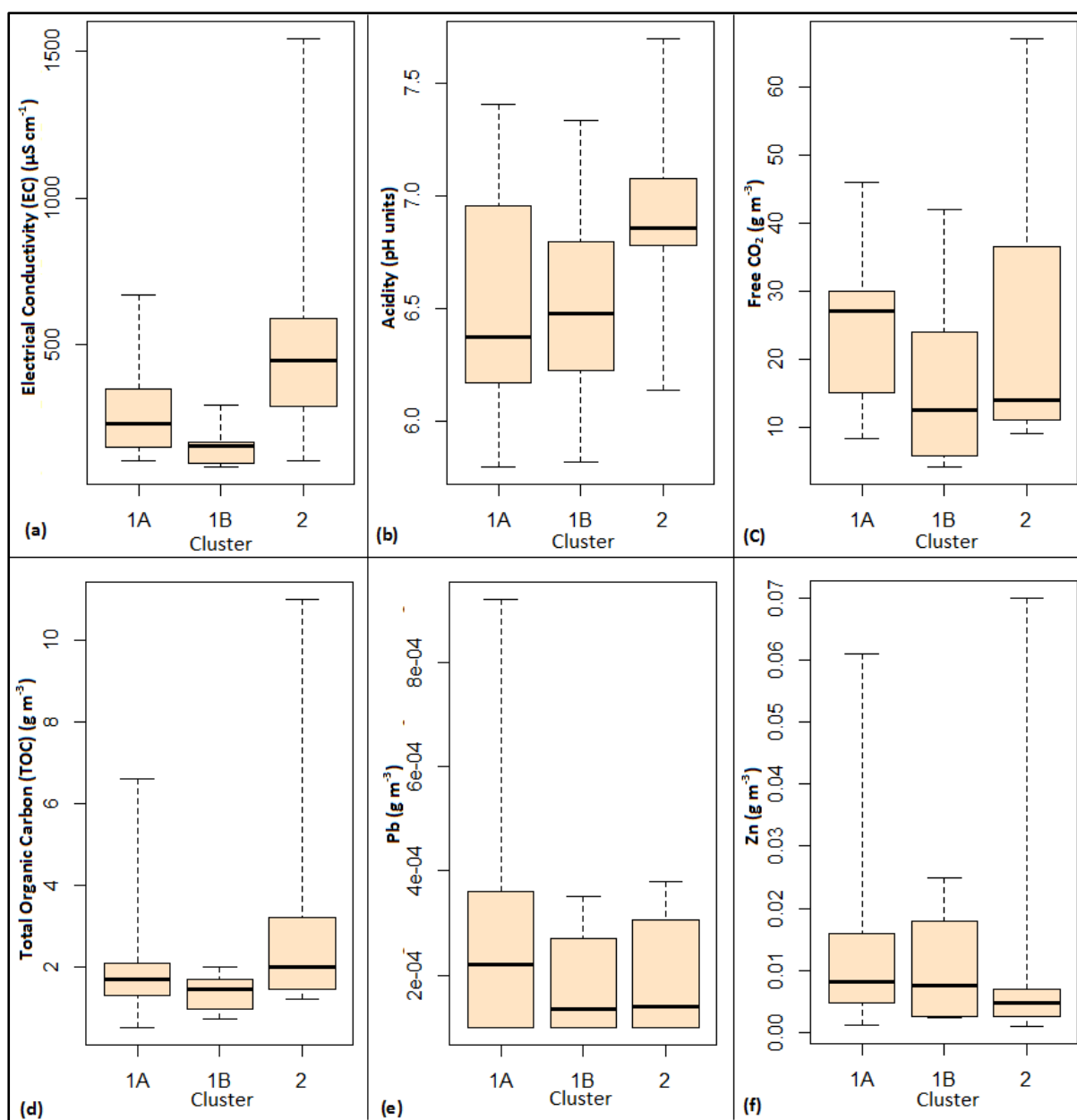


Fig. S2 (iv) Box-and-Whisker Plot comparisons of Electrical conductivity (a), Acidity (b), Free CO_2 (c), Total organic carbon (d), concentrations of Pb (e) and Zn (f) across Bioclusters defined at the 3-cluster threshold.

Pyrosequencing analysis of bacterial diversity in New Zealand groundwater

Sirisena KA, Daughney CJ, Moreau M, Sim DA, Ryan KG, Chambers GK 2014.

Pyrosequencing analysis of bacterial diversity in New Zealand groundwater. *(to be submitted to Molecular Ecology by June 2014)*

Keywords: bacterial diversity; 454 sequencing; groundwater chemistry; microbial ecology; environmental monitoring; New Zealand

Running title:

Groundwater bacterial diversity in New Zealand

Abstract

Bacterial diversity of New Zealand groundwater was characterized using Roche 454 pyrosequencing of the V5-V7 region of the bacterial 16S rRNA gene. Here, we analyzed bacterial assemblages from 35 sites across the country that represent four groundwater hydrochemical categories: oxidized with high human impact; oxidized with low human impact; moderately reduced; and highly reduced. A total of 281,896 partial sequences of 16S rRNA gene were obtained for the 35 samples. We detected 6579 OTUs of which 65 % (4281 OTUs) were singletons. The singletons represented ~ 1.5 % of the overall OTU abundance, while the 10, 100 and 1000 most abundant OTUs accounted for ~ 70 %, 92 % and 97 % respectively. Both non-metric multidimensional and canonical correspondence analysis revealed that bacterial communities were clustered according to the redox potential of the water. We found that the more abundant OTUs mainly contributed to the similarity of bacterial populations within hydrochemical groups whereas the combination of abundant and rare OTUs contributed to the dissimilarities among the different hydrochemistries. Our results indicated that the groundwater hydrochemistry is correlated with the bacterial community structure and NO₃-N, pH, Br and SO₄ were the key parameters that influence bacterial diversity. Across all the hydrochemical categories, *Proteobacteria* was the most abundant phylum and *Pseudomonas* was the most dominant genus. Overall, our study reflects the potential of using bacterial communities as biological signatures to evaluate groundwater ecosystems.

Introduction

Microorganisms including bacteria inhabit almost every environment on earth, from the subsurface to depths of more than a kilometre (Parkes *et al.* 1994). Microorganisms provide the greatest diversity among all groups of living organisms, in terms of metabolic functional capability and habitat (Griebler & Lueders 2009). Therefore, microbes govern many key biogeochemical processes and ultimately they may control the entire global process of energy and material transformation (Chapelle 2000). Terrestrial subsurface environments provide the largest habitat for microorganisms and a large fraction (up to 40%) of prokaryotic biomass may be present in these environments (Whitman *et al.* 1998).

Groundwater ecosystems also provide vastly diverse habitats to harbour this microbial diversity (Griebler & Lueders 2009). Therefore, a proper understanding of the groundwater microbiota will enable us to assess the current state and future trends of the ecosystem, providing crucial information for the sustainable management of groundwater resources. Although initial groundwater microbiological observations date back a few centuries, systematic studies of groundwater microbial diversity have accelerated during the last few decades with the recognition of the important role of microbiota in maintaining ecosystem health (Griebler & Lueders 2009). In recent years, some countries in Europe and some parts of Australia have included microbial parameters into their national groundwater monitoring policies with increasing emphasis on evaluating groundwater bacterial diversity and relating the microbial state to the abiotic properties of the ecosystem (Steube *et al.*, 2009; Griebler *et al.*, 2010; Stein *et al.*, 2010; Korbel & Hose 2011). The recent advances in the rapid development of molecular techniques for environmental samples provide an ideal platform to understand the bacterial diversity in groundwater.

Early subsurface microbial studies were mainly based on culturing techniques in which pure microbial strains obtained from environmental samples were grown in laboratory conditions for evolutionary and metabolic investigations. However, it is now understood that many microbial species present in subsurface environments cannot be cultured in laboratory conditions, probably due to our lack of knowledge about the required conditions or inability to reproduce them in the laboratory. Therefore, culturing methods have become less popular and molecular tools have emerged as the most popular approach for subsurface microbiological studies.

Molecular methods encompass an array of culture-independent techniques that can be selected according to the objective and scale of the study. Simple molecular profiling tools such as terminal restriction fragment length polymorphism (T-RFLP), automated ribosomal intergenic spacer analysis (ARISA) and denaturing gradient gel electrophoresis (DGGE) have been widely used in many subsurface microbial investigations because they are relatively cost effective and provide a rapid overview of the microbial community structure (Wood *et al.* 2013). Each of these molecular profiling tools has its own advantages and disadvantages and can be chosen according the purpose of the investigation (Zinger *et al.* 2012). However, the common drawback in many profiling tools is that they are not capable of directly and confidently providing the taxonomic identities of the microbial species present in the sample. Further, many of these techniques are not powerful enough to identify the less abundant species in the subsurface environments (Bent *et al.* 2007).

Metagenomic methodologies have emerged as a means of overcoming the limitations of molecular techniques such as T-RFLP, ARISA and DGGE. In the earlier metagenomic approaches, the amplified DNA fragments of the gene of interest were cloned into a cloning vector and taxonomic identifications are conducted using the classical Sanger sequencing technique (Handelsman 2004; Handelsman *et al.* 2007). However, this classical approach is

still laborious and time consuming and successful alternatives were found with the introduction of high-throughput next-generation sequencing (NGS) technologies. To date, several NGS platforms have been commercially introduced: Roche 454™; Illumina™; and Ion Torrent™ (Wood *et al.* 2013). Despite the fact that, in general, all these NGS technologies require substantial expert knowledge on bioinformatics and relatively complex library preparation protocols, they provide: an enormous deep sequencing power to identify less abundant taxa in the environment; taxonomic information of the microbial populations; and more precise information on abundance and evenness of microbial communities. Therefore, these NGS technologies enable us to gain more understanding of the subsurface biosphere, including groundwater ecosystems, especially in terms of the interactions between the biotic and abiotic components.

This study investigates the bacterial diversity of New Zealand groundwater using Roche 454 pyrosequencing technology. Groundwater is an important resource for water supply in New Zealand. In recent years, studies have evaluated bacterial community structures in New Zealand groundwater at local and national scales (van Bekkum *et al.* 2006; Sirisena *et al.* 2013, 2014), but none of them were able to obtain precise taxonomic information of the microbial communities. Thus, the first objective of this study was to investigate the bacterial diversity of New Zealand groundwater quantitatively. This enables a more precise estimation of the microbial composition than with the T-RFLP molecular profiling tool (Sirisena *et al.* 2013). The second objective of the study was to reveal the taxonomic identities of the groundwater bacterial species for the first time in New Zealand. This information will significantly contribute to our understanding of the groundwater ecosystem properties and provide a solid platform to investigate the interactions between the biotic and abiotic components. The third objective of the study was to reveal the relationships among bacterial assemblage composition and groundwater hydrochemistry, in

particular the redox potential and environmental factors such as aquifer lithology, confinement, depth of the groundwater bore, mean residence time of the water and land use activities around the aquifer recharge zone. Here, we hypothesized that the bacterial community composition is mainly related to the redox condition of the water. Therefore, we expected that bacterial communities collected from different geographical regions but with similar hydrochemistries would be more similar than the communities collected from the same geographical regions but with different groundwater chemistries. Overall, our present study will provide a novel and deep insight into New Zealand groundwater ecosystems.

Materials and methods

Study site and sampling

In this study, we have analysed a set of 35 representative samples (Fig. 1) that were chosen from the national survey of bacterial diversity conducted by Sirisena *et al.* (2013). These samples were collected from groundwater monitoring sites (bores) that are geographically and hydrologically isolated and located across New Zealand. A single groundwater sample was collected from each site in June 2010 according to the National Protocol for State of the Environment Groundwater Sampling (Daughney *et al.* 2006). The samples were kept at 4 °C until further processing. Site-specific information and hydrochemical data related to these sites are summarized in Table S1 and S2 (Supporting information). Further information pertaining to these sites is available in the GNS Science Geothermal and Groundwater (GGW) Database (<http://ggw.gns.cri.nz/ggwdata/mainPage.jsp>).

The sites were selected to represent different hydrochemical categories based on the hydrochemical facies introduced by Daughney & Reeves (2005). In this categorization, groundwater samples are initially assigned to one of two categories, based on redox state, as oxidized (hydrochemical category 1) or reduced (hydrochemical category 2). The oxidized groundwaters were further subdivided on the basis of degree of human impact as impacted water (category 1A) or unimpacted water (category 1B). The reduced groundwaters were subdivided on the basis of redox state as moderately reduced (category 2A) or highly reduced water (category 2B). Our sampling design was based on these four hydrochemical categories, with sites assigned to categories on the basis of median values of 19 hydrochemical parameters over the period from March 2008 to March 2012. Each category was represented by more or less an equal number of samples.

DNA extraction and 454 pyrosequencing

DNA extraction. In this study, we used the groundwater microbial genomic DNA that was previously extracted from the same samples by Sirisena *et al.* (2013). Briefly, 2 L of groundwater from each site was filtered through a sterile 0.22 µm nitrocellulose membrane filter (Millipore, Australia). The filtrations were carried out immediately as delays could alter the natural microbial community composition in these samples (Gilpin *et al.* 2013). The filters were subsequently stored at -20°C in sterilized 50 mL plastic tubes until use. Bacterial genomic DNA extractions were performed as outlined by Sirisena *et al.* (2013) using ZR Fungal/Bacterial DNA kits (Zymo Research, United States) according to the manufacturer's instructions. The recovered DNA was quantified using Quant-iT™ High-Sensitivity DNA Assay kits (Invitrogen, United States), dissolved in 100 µl of molecular biology grade water (MO BIO Laboratories, Inc., United States) and stored at -20°C until further processing.

PCR amplicon library preparation and 454 pyrosequencing.

Amplicon library preparation involved two rounds of PCR amplification on the V5-V7 region of the bacterial 16S rRNA gene (Bottos *et al.* 2014). The first round of PCR was performed in triplicates for each sample using the primers Tx9 (5'-GGATTAGAWACCCBGGTAGTC-3') and 1391R (5'-GACGGGCRGTGWGTRCA-3') (Ashby *et al.* 2007). Each 30 µl reaction mixture contained 1X PrimeSTAR buffer, 0.2 mM dNTPs, 0.025 U PrimeSTAR HS DNA Polymerase (Takara Holdings, Kyoto, Japan), 0.4 µM of primers Tx9 and 1391R (Integrated DNA Technologies, Iowa, USA), 5 µl (~ 5 ng) of bacterial genomic DNA and molecular biology grade water (MO BIO Laboratories, Inc., USA) to adjust the final volume. In this step, we determined the optimum number of thermal cycles for each sample. Thermal cycling conditions consisted of an initial denaturation step of 3 min at 94 °C followed by 18, 24 and 30 cycles for the three aliquots of each sample respectively at 94 °C for 20 s, 52 °C for 20 s and 72°C for 45 s. The reactions were completed with a final extension at 72°C for 4 min. All amplifications were completed on a Mastercycler® Pro S PCR system (Eppendorf, Hamburg, Germany). PCR products were run on a 1% agarose gel stained with SYBR® Safe in 1X Tris-Acetate-EDTA (TAE) buffer (Life Technologies, CA, USA) and visualized by a Safe Imager™ 2.0 Blue Light Transilluminator (Life Technologies, CA, USA). The optimum PCR cycling condition for each sample was determined by visual comparison of the amplified products in gel electrophoresis. PCR was repeated for each sample in duplicates with the optimum cycling conditions as described above. Then the three reactions (duplicate aliquots from this PCR and one aliquot obtained with the optimum conditions in the previous PCR) were pooled together and amplicons were run on 2% agarose gels as described above. The targeted PCR bands were excised from the gel and purified using MO BIO Gel Extraction Kits (MO BIO Laboratories, CA, USA). Gel purified products were purified again

using an Agencourt AMPure XP system (Beckman Coulter, CA, USA) according to the manufacturer's instructions and quantified using Quant-iT™ High-Sensitivity DNA Assay kits (Life Technologies, CA, USA).

To prepare the amplicons for pyrosequencing, the second round of PCR was completed using the primers with 454 adaptors and unique Multiplex Identifier (MID) sequences. The forward primer included a MID sequence unique to each sample. PCR was performed in triplicates and reactions were prepared as described above with these exceptions: only 10 ng of purified PCR product was used as the template; the primers Bac(X)A-Tx9F (5'-CCATCTCATCCCTGCGTGTCTCCGACTCAG-MID(X)-GGATTAGAWACCCBGGTAGTC-3') and BacB-1391R (5'-CCTATCCCCTGTGTGCCTTGGCAGTCTCAG-GACGGGCRGTGWGTRCA-3') were used; thermal cycling was reduced to 10 cycles. Triplicate PCR reactions were pooled, purified and quantified as described above. The remaining steps were performed at the University of Waikato DNA Sequencing Facility. The number of amplifiable molecules was quantified using a KAPA Library Quantification Kit for Roche 454 Titanium/Universal (Kapa Biosystems, MA, USA). The amplicons were subjected to unidirectional pyrosequencing by the GS Junior System using the GS Junior Titanium emPCR Kit (Lib-L), the GS Junior Titanium Sequencing Kit and the PicoTitrePlate Kit (Roche 454 Life Sciences, CT, USA).

Sequence analysis and identification of Operational taxonomic Units (OTUs)

Raw pyrotags were analyzed using AmpliconNoise v1.25 and chimeric sequences were eliminated using Perseus (Quince *et al.* 2011). Sequences were assigned to the relevant sample according to the MIDs that were used to label each sample. Amplicons that had at

least a single-base mismatch to the PCR primer sequence were not considered for further analyses. The identification of unique sequences and operational taxonomic units (OTUs) was conducted using Mothur v1.24.1 (Schloss *et al.* 2009) and ESPRIT (Sun *et al.* 2009) software packages. Two sequences were considered to be unique if they differed in at least a single base mismatch or similar in the nucleotide sequence, but there is at least one extra base at the end in one sequence. Pair-wise distances between reads were computed and the sequences that had similarities of 97% or more were grouped into OTUs. Taxonomic identities of the unique sequences were obtained using the Classifier function of the Ribosomal Database project (RDP) Release 10 (Wang *et al.* 2007).

Accession numbers and data availability

We will deposit the sequence data and supporting data generate in this study in a public database upon the acceptance of the manuscript for publication.

Diversity estimation and statistical analysis

Diversity indices. The programme Mothur v1.24.1 (Schloss *et al.* 2009) was used to calculate the following: rarefaction curves; the Chao 1 and the non-parametric Abundance Coverage Estimator (ACE) as OTU richness indices; and the Shannon diversity index (H') and Simpson diversity index as measures of diversity. The Kruskal-Wallis test was used to investigate the relationship among the richness and diversity indices and the hydrochemical categories. Further, the number of OTUs and the Shannon diversity indices observed in this study were compared with the same parameters that were generated in a previous study using the simple molecular tool T-RFLP (Sirisena *et al.* 2013).

Cluster analysis. We applied non-metric multidimensional scaling (nMDS) (Kruskal 1964a,b) to determine the similarities between samples. Here, we used Bray-Curtis similarities that were derived from the relative abundance of (i) all OTUs; (ii) all OTUs except singletons; and (iii) the 100 most abundant OTUs to calculate how closely related the samples were to each other. In this method, similarities are represented in a multidimensional space by plotting the samples as clusters of more similar samples. This representation enables the identification of the key environmental factors that relate to the clustering pattern of bacterial assemblages. We used the Permutational Multivariate Analysis of Variance (PERMANOVA) test (Anderson *et al.* 2008) with 9999 permutations to verify that the bacterial populations from the same cluster were more similar than those from different clusters. Further, the test was used to evaluate the statistical significance of discriminating bacterial communities as represented by nMDS, considering two hydrochemical groups (categories 1 or 2) or four hydrochemical groups (categories 1A, 1B, 2A or 2B), and the following categorical environmental parameters: aquifer lithology, aquifer confinement, groundwater mean residence time (MRT) class (cf. Daughney *et al.* 2010), well depth code, land use activities in the aquifer recharge zone; and the geographical region (Table S2 Supporting information).

SIMPER analysis. Similarity percentage (SIMPER) analysis (Clark & Warwick 2001) was employed to reveal the average similarity within and the average dissimilarity among the four hydrochemical categories based on the OTU diversity. Further, it was used to identify those OTUs that contributed mostly to the similarity/dissimilarity within/between the four hydrochemical categories.

RELATE analysis. This is a comparative Mantel-type test that can be used to determine the correlation between two sets of continuous variables (Clark & Warwick 2001). We

employed the RELATE analysis to understand the correlation between bacterial diversity represented by all the OTUs and the groundwater chemistry as a whole rather than individual parameters. The hydrochemical data (x) for the 19 parameters included in Table S1 were transformed to natural log variables [$\ln(x+1)$] in order to eliminate uneven values among different parameters. A similarity matrix was computed based on these hydrochemical variables. The Euclidian distance was used in this purpose as it a more appropriate measure than the Bray-Curtis similarity for grouping environmental data (Ramette 2007). Another similarity matrix was computed based on OTU diversity using Bray-Curtis similarity. The two similarity matrices were used in RELATE analysis to reveal the relationship between hydrochemistry and bacterial diversity.

Cluster analysis, PERMANOVA test, RELATE analysis and SIMPER analysis were performed using the PRIMER v.6 statistical programme (Primer-E Ltd., Plymouth, UK) with the additional add-on package PERMANOVA+ (Anderson *et al.* 2008).

Canonical Correspondence Analysis (CCA). A Canonical Correspondence Analysis (ter Braak & Smilauer 2012) was performed to reveal the relative contribution of each hydrochemical parameter in determining the bacterial community structure explained by all OTUs. Further, we assumed that microbial communities are related to hydrochemistry and tested this hypothesis using a Monte Carlo test with 499 permutations under a constrained (species versus environmental variables) model. The CCA was performed with the CANOCO 5 for Windows package (ter Braak & Smilauer 2012).

Results

Analysis of bacterial operational taxonomic unit (OTU) diversity

The pyrosequencing of 35 groundwater DNA samples resulted in 281,896 partial sequences of 16S rRNA gene after quality filtration and chimera removal. We detected 6579 OTUs of which 65 % (4281 OTUs) were singletons at 97 % similarity based on the average neighbour algorithm. The singletons represented ~1.5 % of the overall OTU abundance, while the 10, 100 and 1000 most abundant OTUs accounted for ~70 %, 92 % and 97 % respectively.

The bacterial diversity and richness estimates significantly varied within and among the four hydrochemical categories (Table 1). The overall observed bacterial species richness ranged from 29 to 947 OTUs and both these extreme values were found in oxidized groundwater with high human impact (category 1A). However, it is important to note that second highest richness recorded in the category 1A was 277 OTUs, which is substantially less than the highest value (947) in this hydrochemical category. The ranges of species richness in other hydrochemical categories fell within the observed span for category 1A: from 60 to 494 OTUs in oxidized groundwater with low human impact (category 1B), from 87 to 366 OTUs in moderately reduced groundwater (category 2A) and from 41 to 481 OTUs in highly reduced groundwater (category 2B).

The Shannon diversity index (H') ranged from 0.34 to 3.98 across the 35 groundwater samples. The average diversity for each hydrochemical category was: 1A – 2.06, 1B – 1.67, 2A – 1.87 and 2B – 2.12, and this indicated that groundwaters with high human impact possess slightly greater diversity than the groundwaters with low human impact. Similarly, highly reduced groundwaters also had a relatively higher diversity than moderately reduced waters. The abundant OTUs (≥ 10 reads) represented only 5 to 21 % of the total bacterial

community in each sample while rare OTUs (≤ 2 reads) provided the major contribution with an average of 76 ± 5.9 % of the diversity in each sample. Further, 88.5 % of all OTUs (5827 OTUs) were found in only one sample (Unique OTUs) whereas only 35 OTUs were detected in 10 or more samples. However, the unique OTUs contributed a total of 19,621 reads which is around 7 % of the overall abundance, while most common OTUs (shared among 10 or more samples) comprised 207,496 reads reflecting 73.6 % of total abundance.

The Kruskal-Wallis test indicated that there was no statistically significant difference in any of the richness estimates or diversity indices between the four hydrochemical categories: Number of OTUs ($P = 0.938$); Chao 1 ($P = 0.956$); ACE ($P = 0.987$); Shannon index ($P = 0.853$); or Simpson index ($P = 0.847$). This suggests that the hydrochemical categories may not properly differentiate the bacterial diversity.

A comparison between the present study and the previous study by Sirisena *et al.* (2013) revealed that the 454 pyrosequencing technology identified a significantly greater number of OTUs than the T-RFLP methodology for all samples (Table S5 Supporting Information). Further, the diversity explained by the Shannon index in the two studies indicates that the 454 approach generally captured a higher diversity than T-RFLP. However, the rarefaction curves (Fig. S1 Supporting information) reflect that our sampling of bacterial richness is not completed yet and we may find additional low abundance OTUs if more sequences are obtained for each sample.

Non-metric multidimensional scaling (nMDS) plots were generated using relative abundances of (i) all OTUs; (ii) all OTUs except singletons; and (iii) the 100 most abundant OTUs. All three approaches provided a more or less similar clustering pattern (Fig S2 Supporting information). Therefore, the remainder of the results in this paper explained considering the plots generated with all OTUs. The nMDS analysis indicated that the pattern of groundwater bacterial diversity coincided with the hydrochemical categories, especially

when just two hydrochemical categories are considered (categories 1 and 2), according a pattern that reflected the redox potential of the groundwater (Fig. 2). For example, the bacterial communities in oxidized groundwater were more similar to each other than the populations in reduced water. The PERMANOVA test also confirmed that there was a statistically significant difference in the bacterial community composition between oxidized and reduced waters ($P = 0.022$). Although the nMDS clustering did not reflect a clear separation of bacterial populations when considering four hydrochemical categories (categories 1A, 1B, 2A and 2B), PERMANOVA analysis revealed a significant variability in bacterial diversity among these categories ($P = 0.033$). Interestingly, both nMDS plots and PERMANOVA analysis showed that bacterial populations are not discriminated by aquifer lithology ($P = 0.775$), aquifer confinement ($P = 0.098$), MRT class ($P = 0.256$), well depth code ($P = 0.272$), land use activities ($P = 0.074$) or geographical region ($P = 0.432$).

The SIMPER analysis revealed that there were significant dissimilarities in microbial communities between each pair of hydrochemical groups (Fig. S3 Supporting Information). The lowest dissimilarities were between groups 1A & 1B and 1B & 2A. However, the average similarity of bacterial communities within each hydrochemical category was relatively low: 1A – 13.12 %; 1B – 16.02 %; 2A – 16.10%; and 2B – 9.91 %. Further, the more abundant OTUs mainly contributed to the similarity of bacterial populations within hydrochemical groups whereas the combination of abundant and rare OTUs contributed to the dissimilarities among the different hydrochemical groups (Table S6 Supporting Information).

The relationship between microbial diversity and hydrochemistry was investigated by correlating the two similarity matrices using RELATE analysis. This indicated that the groundwater hydrochemistry was correlated with the bacterial community structure ($r = 0.25$, $P = 0.002$). Canonical correspondence analysis (CCA) describes the relative contribution of

each hydrochemical parameter to the variation in the bacterial communities. The results suggested that the 19 hydrochemical variables accounted for 59 % of the total variability in the relative abundance of all OTUs found in the groundwater samples (Monte Carlo permutation test, $P = 0.002$). Further, CCA indicated that samples are generally clustered according to the redox condition of the water (Fig. 3). In addition, $\text{NO}_3\text{-N}$, pH, Br and SO_4 were the key explanatory variables, where the first two parameters separate the samples along the first axis and the other two separate along the second axis.

Analysis of bacterial community taxonomic composition

Proteobacteria was the most abundant phylum across all the hydrochemical categories (Fig. 4A). Further, the phylum *Cyanobacteria* comprised a small percentage of the taxonomic diversity of oxidized groundwater having high human impact (category 1A). At the class level, *Gammaproteobacteria* and *Betaproteobacteria* were predominant among all hydrochemical categories (Fig. 4B). However, the relative proportions of the two classes were different between oxidized (1A & 1B) and reduced (2A & 2B) waters:

Betaproteobacteria was the most dominant in oxidized groundwater whereas

Gammaproteobacteria was the most dominant in reduced groundwater. At the order level,

Burkholderiales was the major group present in oxidized groundwater, whereas

Pseudomonadales was the next dominant order. In reduced water, *Pseudomonadales* was the most abundant group while *Burkholderiales* was present as the next abundant component (Fig. 4C). In addition, the moderately reduced groundwater samples (category 2A) contained a considerable percentage (18%) of the order *Campylobacterales* whereas *Methylophilales* and *Rhodocyclales* were also present in significant fractions (19% and 14% respectively) in highly reduced water (category 2B). The diversity at family level revealed that

Oxalobacteraceae was predominant in oxidized groundwaters with *Pseudomonadaceae* as the next dominant component (Fig. 4D). In addition, category 1B also included of a significant percentage (15%) of *Comamonadaceae*. In moderately reduced waters (category 2A), *Pseudomonadaceae* was the predominant family, but the highly reduced groundwaters (category 2B) consisted of three equally dominant families: *Methylophilaceae*; *Pseudomonadaceae*; and *Rhodocyclaceae*. However, despite the hydrochemical differences, *Pseudomonas* was the most dominant genus in all four hydrochemical categories: 1A – 26 %; 1B – 32 %; 2A – 56 %; and 2B – 17 % (Fig. 5). Overall, the analysis of taxonomic diversity of bacterial communities suggested that each hydrochemical group consisted of a unique combination of dominant bacterial genera enabling us to discriminate hydrochemical groups according to taxonomic composition.

CHAPTER 3.3

Table 1 Summary of bacterial diversity and richness estimates based on 454-pyrosequencing operational taxonomic units (OTUs) defined at 0.03 cut-off level

Hydrochemical Category ^a	GGW ID ^b	Number of OTUs	Chao 1 (95 % CI)	ACE (95 % CI)	Simpson (95 % CI)	Shannon (95 % CI)
1A Oxidized groundwater with high human impact	1993	947	2106 (1866–2408)	2023 (1824–2267)	0.110 (0.104–0.116)	3.98 (3.91–4.05)
	54	277	838 (638–1150)	790 (625–1034)	0.208 (0.202–0.214)	2.26 (2.22–2.29)
	36	268	622 (499–810)	628 (513–797)	0.532 (0.520–0.543)	1.33 (1.29–1.37)
	395	255	643 (504–860)	664 (532–858)	0.050 (0.045–0.056)	3.96 (3.86–4.06)
	52	240	518 (414–685)	512 (418–657)	0.078 (0.073–0.082)	3.27 (3.22–3.32)
	389	178	523 (371–795)	454 (348–624)	0.541 (0.529–0.554)	1.25 (1.20–1.29)
	388	126	357 (243–581)	296 (221–430)	0.572 (0.560–0.584)	1.18 (1.15–1.22)
	18	83	211 (144–354)	217 (148–363)	0.483 (0.476–0.490)	0.99 (0.97–1.01)
1B Oxidized groundwater with low human impact	17	29	63 (40–137)	78 (45–177)	0.866 (0.857–0.875)	0.34 (0.32–0.36)
	380	494	1517 (1228–1921)	1589 (1305–1973)	0.310 (0.305–0.315)	1.89 (1.86–1.91)
	53	471	1233 (1019–1529)	1245 (1046–1514)	0.197 (0.191–0.203)	2.58 (2.53–2.62)
	49	258	789 (593–1099)	767 (596–1025)	0.170 (0.160–0.179)	2.75 (2.69–2.81)
	362	217	504 (393–687)	495 (395–650)	0.412 (0.400–0.424)	1.71 (1.67–1.75)
	383	171	402 (306–566)	402 (311–550)	0.393 (0.387–0.399)	1.35 (1.32–1.39)
	458	116	270 (199–402)	284 (210–415)	0.609 (0.579–0.639)	1.30 (1.20–1.41)
	74	97	237 (166–379)	253 (177–402)	0.613 (0.601–0.625)	0.93 (0.90–0.96)
2A Moderately reduced groundwater	39	86	190 (132–320)	188 (136–294)	0.200 (0.195–0.205)	1.99 (1.95–2.02)
	69	60	119 (84–206)	130 (90–222)	0.800 (0.789–0.810)	0.55 (0.52–0.58)
	12	366	949 (768–1210)	933 (770–1162)	0.103 (0.097–0.110)	3.38 (3.31–3.45)
	83	290	611 (504–771)	632 (526–787)	0.418 (0.408–0.428)	1.70 (1.67–1.74)
	3327	268	726 (563–977)	735 (585–955)	0.366 (0.349–0.383)	2.10 (2.03–2.17)
	8	220	452 (362–599)	444 (366–564)	0.213 (0.202–0.225)	2.55 (2.48–2.62)
	42	206	772 (539–1168)	1147 (925–1438)	0.352 (0.337–0.368)	1.93 (1.86–2.00)

CHAPTER 3.3

	6	191	393 (308–540)	363 (298–467)	0.579 (0.572–0.586)	1.04 (1.02–1.06)
	467	128	321 (229–496)	320 (236–469)	0.254 (0.241–0.267)	2.08 (2.02–2.14)
	27	104	230 (166–362)	228 (169–340)	0.389 (0.386–0.393)	1.23 (1.22–1.25)
	30	87	146 (115–211)	153 (134–179)	0.749 (0.735–0.764)	0.78 (0.73–0.82)
	80	481	953 (818–1141)	901 (795–1045)	0.141 (0.134–0.147)	3.27 (3.21–3.33)
	364	452	1305 (1062–1644)	1245 (1044–1513)	0.217 (0.210–0.224)	2.48 (2.42–2.54)
2B	456	397	911 (757–1132)	909 (769–1101)	0.139 (0.133–0.145)	2.90 (2.86–2.95)
Highly	35	216	498 (385–689)	483 (385–635)	0.154 (0.148–0.161)	2.68 (2.63–2.72)
reduced	82	199	474 (366–653)	472 (371–634)	0.269 (0.261–0.278)	1.97 (1.93–2.00)
groundwater	31	167	337 (265–461)	343 (274–455)	0.316 (0.308–0.324)	1.79 (1.75–1.83)
	14	149	295 (231–411)	285 (230–378)	0.667 (0.647–0.688)	1.16 (1.09–1.23)
	338	41	65 (49–116)	68 (51–117)	0.714 (0.697–0.730)	0.71 (0.67–0.75)

^a defined as explained by Daughney & Reeves (2005) which represents the hydrochemistry and degree of human impact at each sampling site

^b Site identification number in the GNS Science Geothermal and Groundwater (GGW) Database (<http://ggw.gns.cri.nz/ggwdata/mainPage.jsp>)

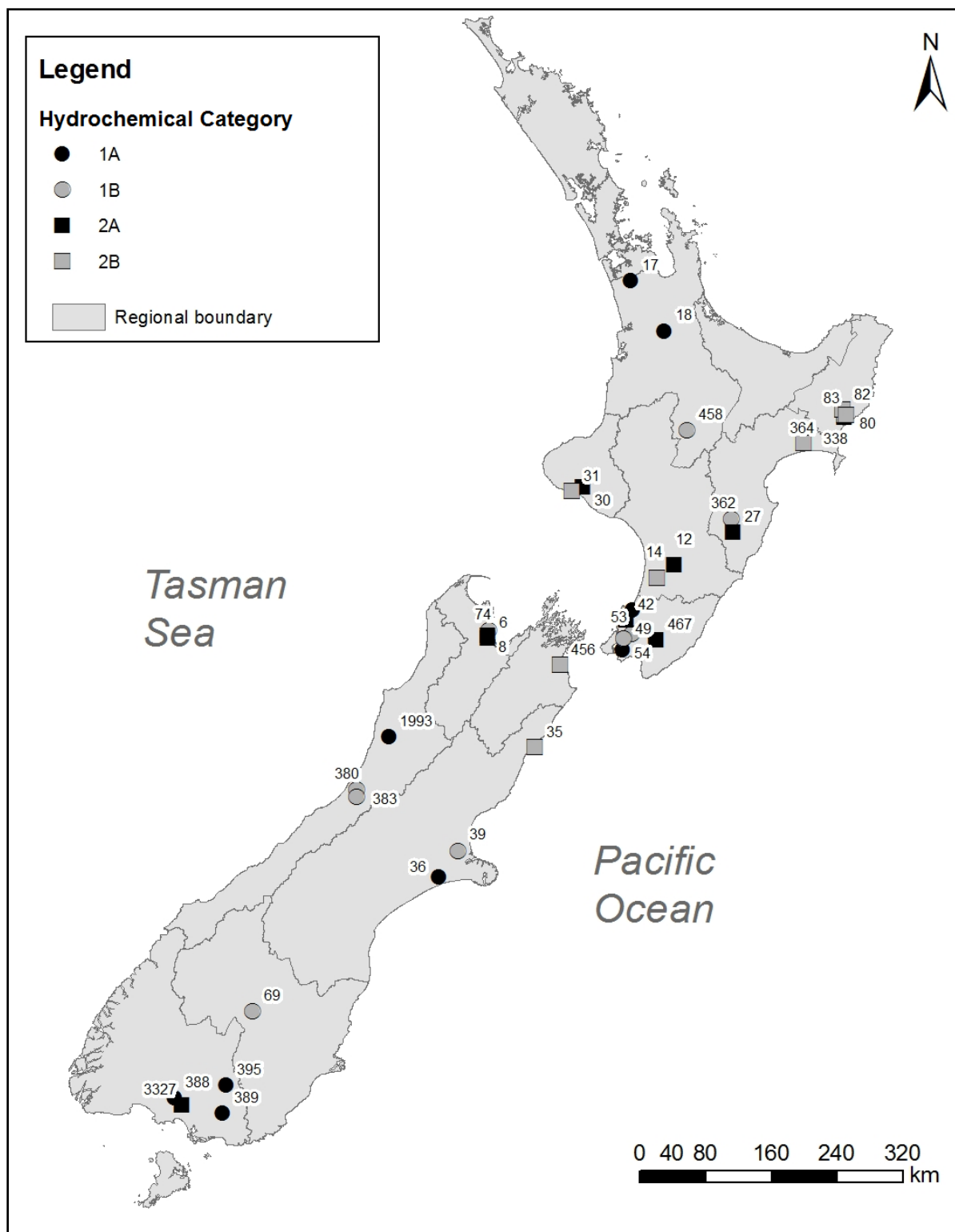


Fig. 1 Groundwater sampling sites across New Zealand. GW ID of each site is displayed next to the site. Hydrochemical categories are determined on the basis of median values of 19 hydrochemical parameters over the period from March 2008 to March 2012, as introduced by Daughney & Reeves (2005).

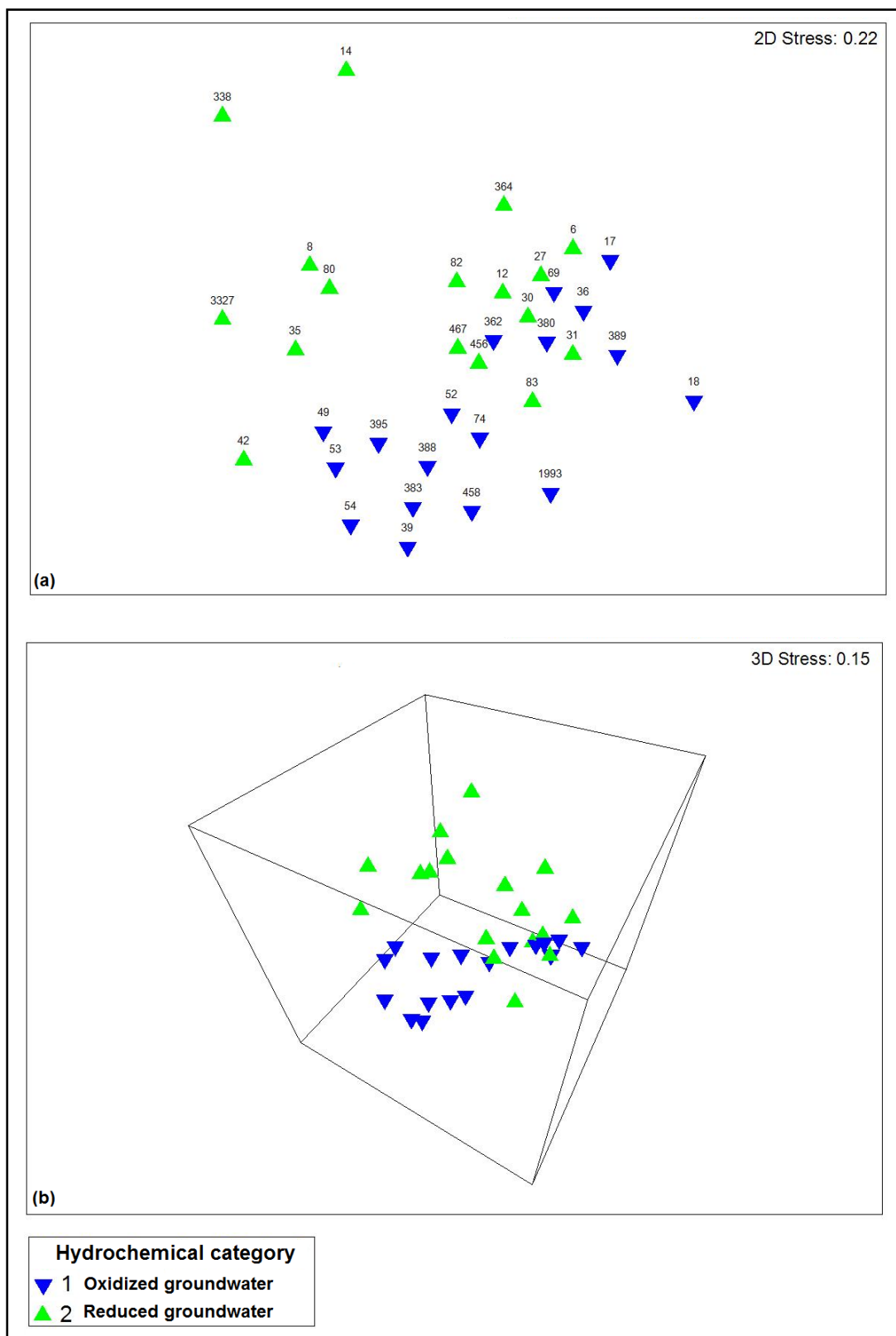


Fig. 2 Non-metric multidimensional scaling based on the relative abundances of all OTUs. Discrimination of samples according to the redox state of the groundwater is displayed: (a) on a 2D plot with a final stress of 0.22; and (b) on a 3D plot with a final stress of 0.15.

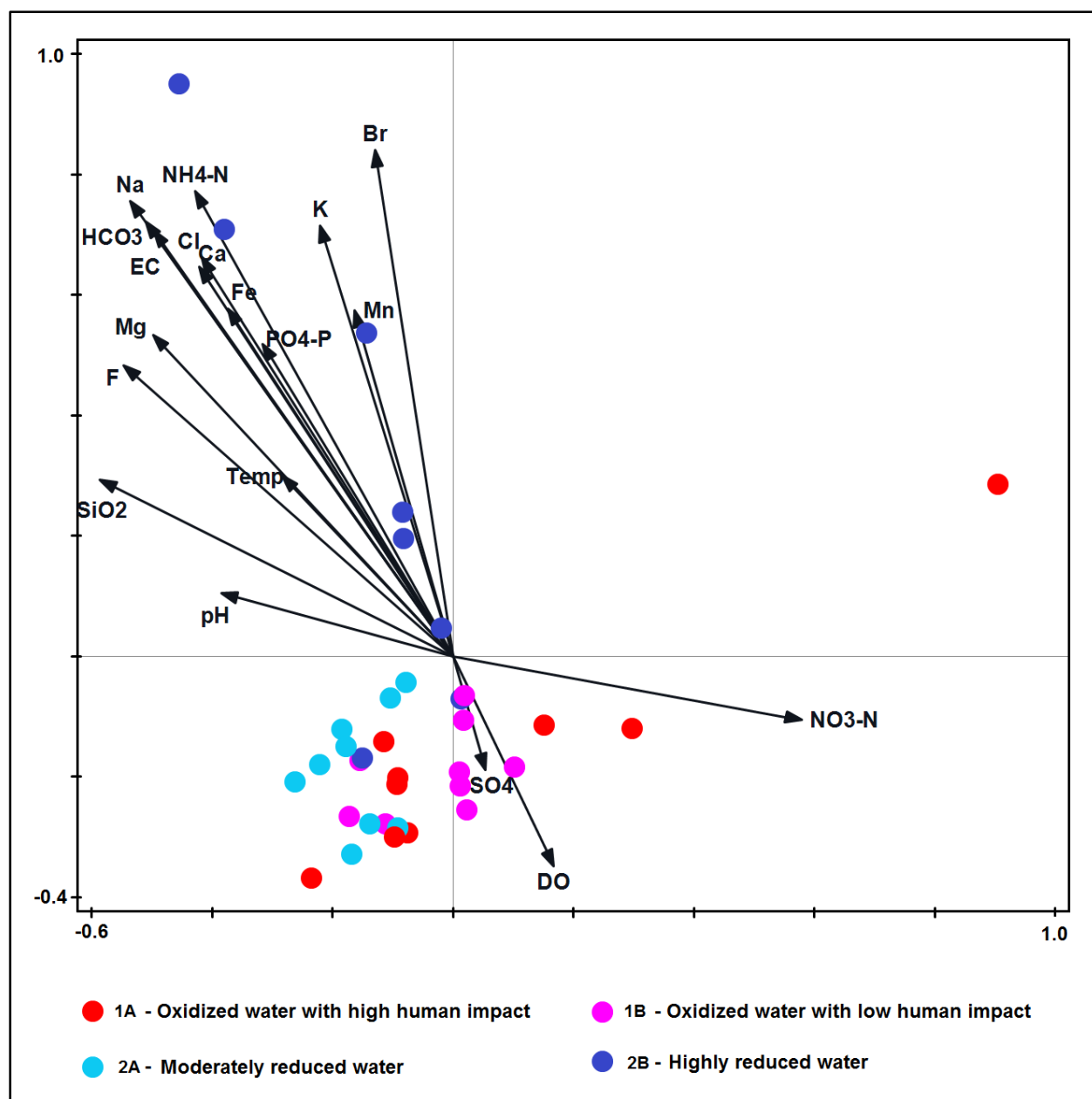


Fig. 3 Canonical correspondence analysis of the relative abundance of all OTUs with the 19 hydrochemical parameters summarized in Table S1 and S2 of the Supporting Information.

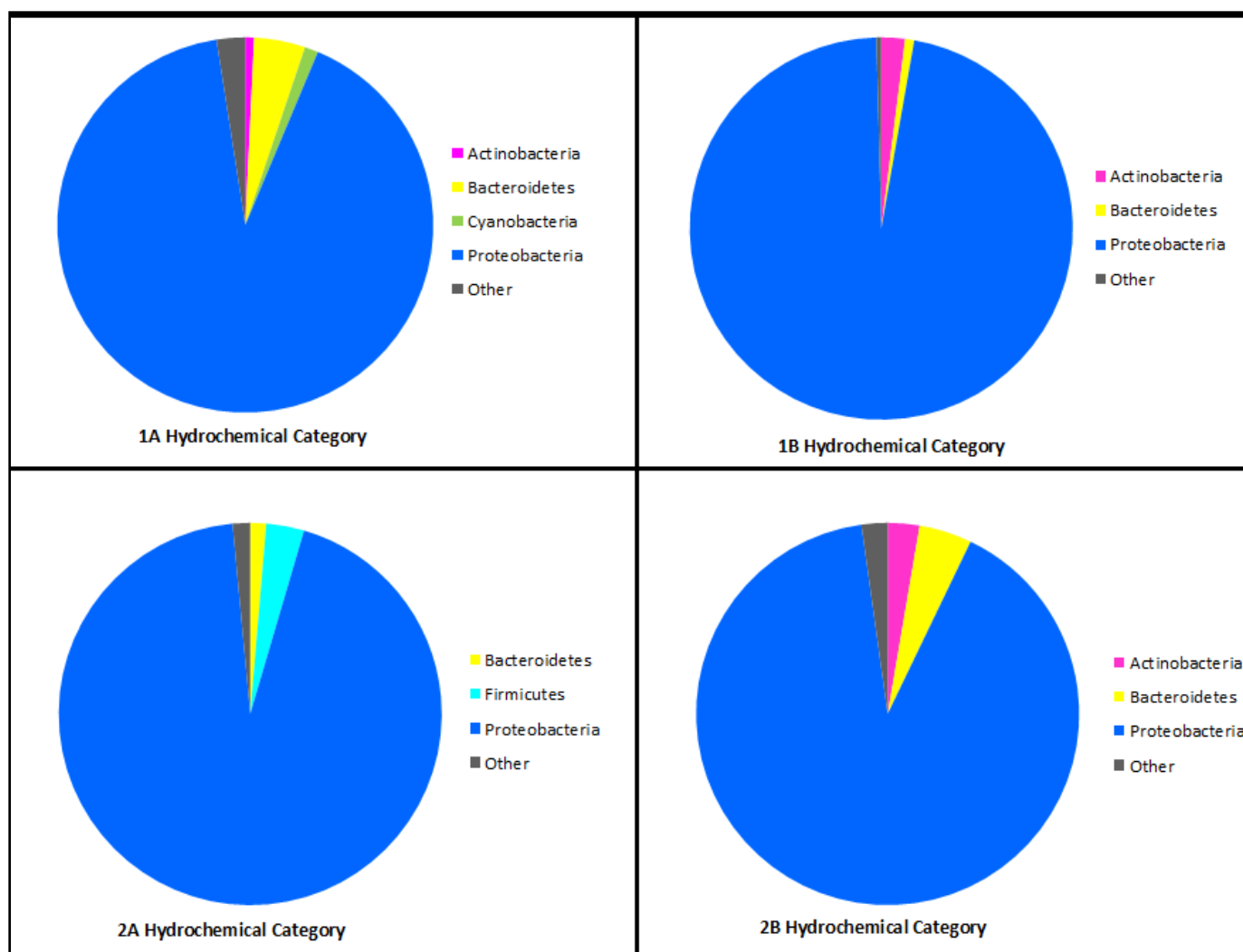


Fig. 4A Groundwater bacterial taxonomic diversity at phylum level. Total number of reads for different OTUs but assigned to the same phylum were summed up to obtain the total number of reads for the particular phylum.

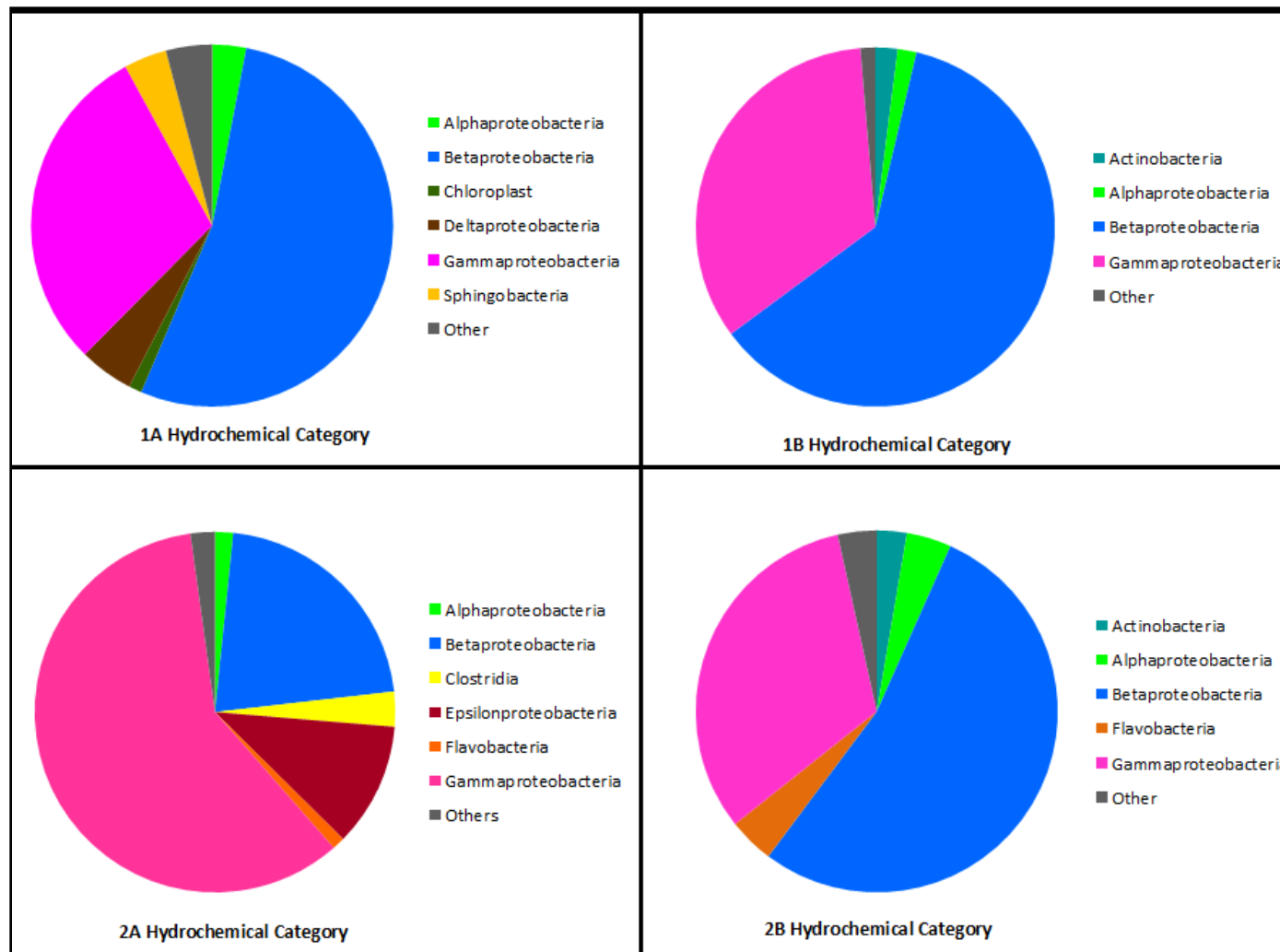


Fig. 4B Groundwater bacterial taxonomic diversity at class level. Total number of reads for different OTUs but assigned to the same class were summed up to obtain the total number of reads for the particular class.

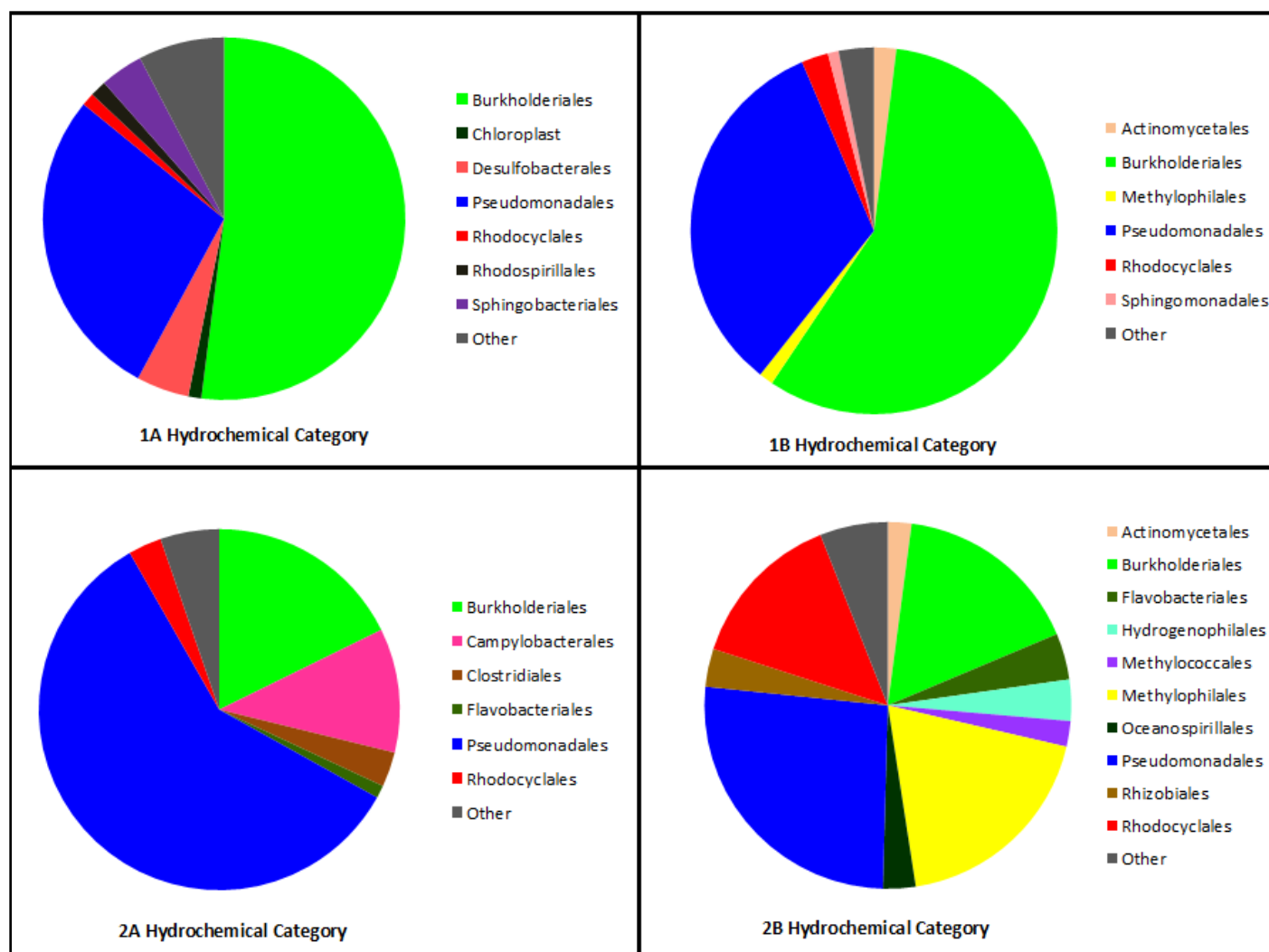


Fig. 4C Groundwater bacterial taxonomic diversity at order level. Total number of reads for different OTUs but assigned to the same order were summed up to obtain the total number of reads for the particular order.

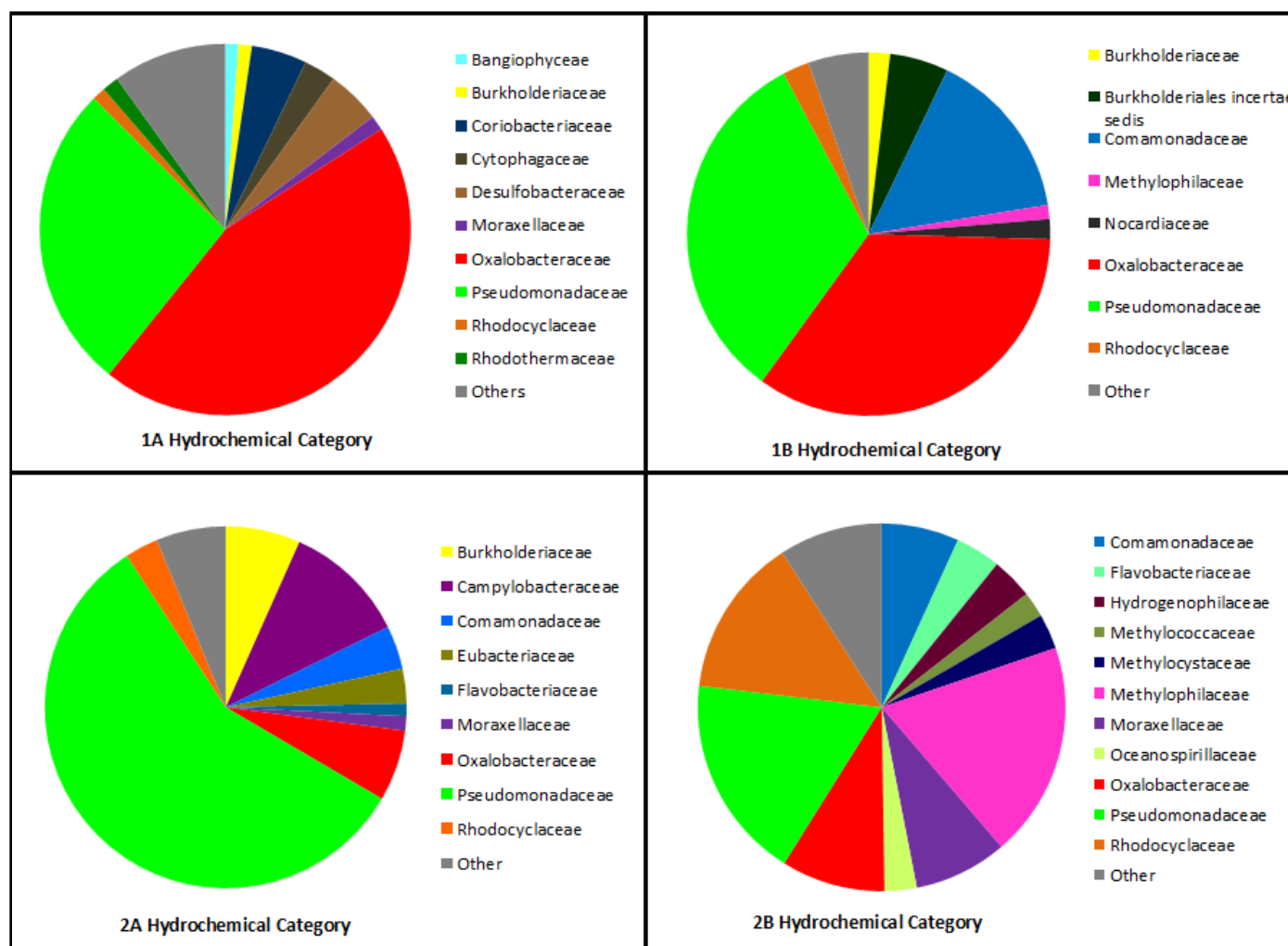


Fig. 4D Groundwater bacterial taxonomic diversity at family level. Total number of reads for different OTUs but assigned to the same family were summed up to obtain the total number of reads for the particular family.

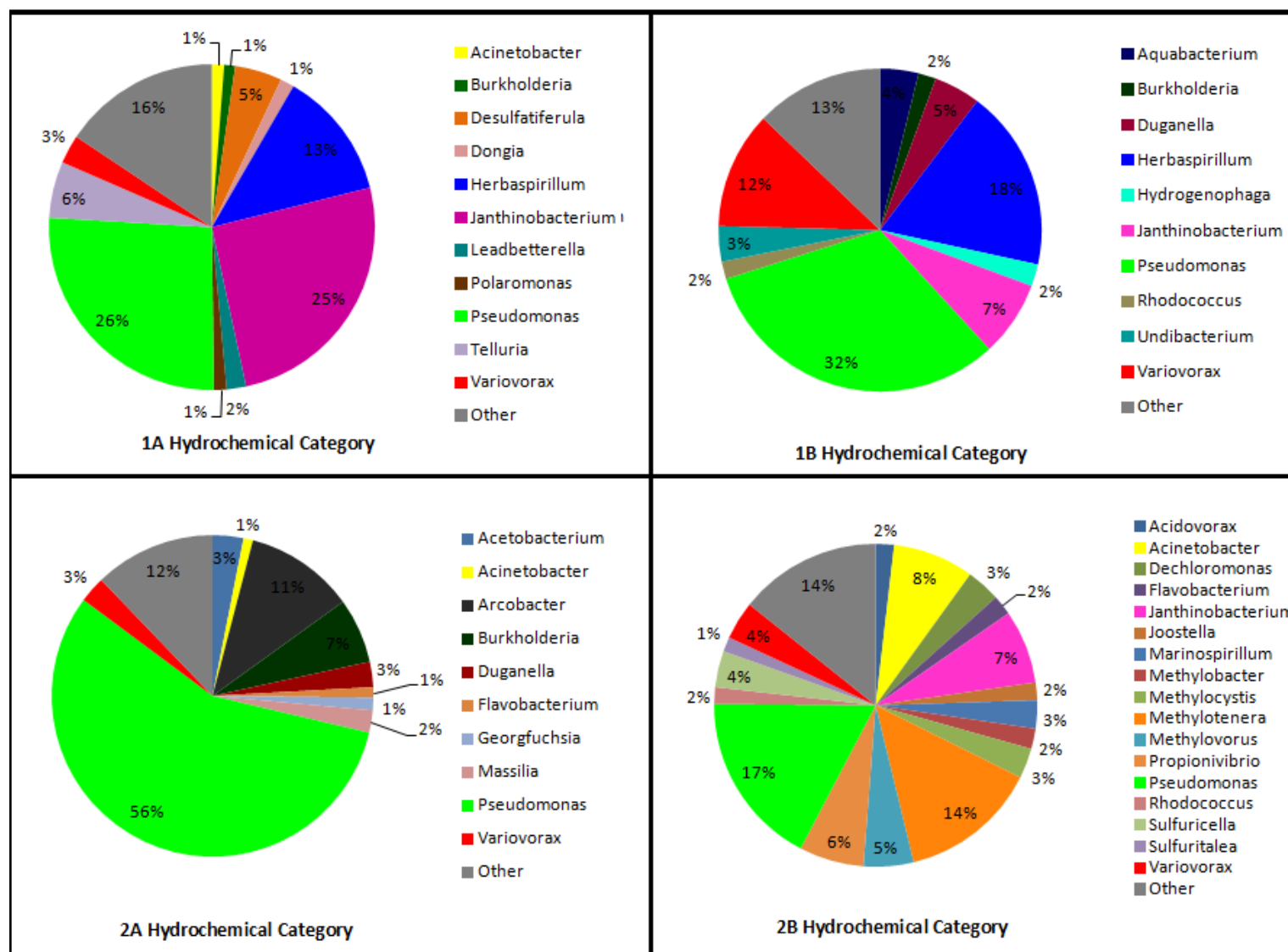


Fig. 5 Groundwater bacterial taxonomic diversity at genus level. Total number of reads for different OTUs but assigned to the same genus were summed up to obtain the total number of reads for the particular genus.

Discussion

Our pyrosequencing approach has allowed us to detect low abundant bacterial taxa, and to quantify the microbial diversity more precisely than our two previous studies (Sirisena *et al.* 2013, 2014, Ch 3.1, 3.2). Overall, 20 times the number of bacterial OTUs were found in this study compared to our previous T-RFLP analysis (Sirisena *et al.* 2013) due to the higher resolution of the pyrosequencing method (Wood *et al.* 2013). In addition, the results suggest that the bacterial community structure is shaped in a way that the most commonly shared OTUs are present with low richness and higher abundances whereas the unique OTUs are represented with higher richness and lower abundance. This observation is more or less consistent across all hydrochemical categories except 1A in which unique OTUs appeared to be more highly abundant than other groups (Tables S3 and S4 Supporting Information). The Shannon diversity indices (H') obtained from the two studies did not exactly reflect a clear pattern that one method always provide higher H' than the other and comparable to each other although Pilloni *et al.* (2012) demonstrated that this is possible if the same target region of the 16S r RNA gene is used in both approaches. It is important to note that the previous study (Sirisena *et al.* 2013) was performed using the full length of bacterial 16S rRNA gene whereas in the present study we have amplified a shorter region of this gene with different primers. These technical differences might have generated an inconsistent H' pattern. Further, the majority of the OTUs identified by pyrosequencing were low abundant species and identifying a vast amount of such taxa may not necessarily increase Shannon diversity indices as both species richness and relative abundance are important aspects in this calculation. Interestingly, the quantitative measures of bacterial diversity have not shown a clear relationship with hydrochemical categories. This implies that, in groundwater ecological perspective, qualitative aspects of bacterial communities may be more important

than its quantitative characteristics, i. e. who are they rather than how many (Lozupone *et al*, 2007).

The bacterial communities with similar hydrochemistries but from different geographical regions were more similar than the communities in the same regions, but with different chemistries. Further, the redox state of groundwater was the most important parameter that shaped the bacterial community structure. These results are generally in accord with the conclusions drawn by Sirisena *et al.* (2013), even though our present study revealed a greater OTU richness and was more effective in detecting rare taxa. This provides a cross-validation for the T-RFLP methodology used in our previous study. However, while the unique combination of more abundant OTUs mainly contributed to the similarities among bacterial populations with similar chemistries, the unique combination of both abundant and relatively rare OTUs shaped the dissimilarities among the samples. Interestingly, NO₃-N and SO₄ that were recognized as important chemical components in CCA analysis are also crucial factors in determination of the redox state of the water. This further supports the observed bacterial diversity-hydrochemistry relationship. Although, none of the environmental parameters indicated a significant pattern with microbial diversity, the land use activities in the aquifer recharge zone tend to reflect a relationship with microbiota ($p=0.074$). This speculation is indirectly supported as NO₃-N, the major parameter that was used to determine the human impact on groundwater, is recognized as one of the key factors for differentiation microbial diversity. However, we acknowledge that further studies are required to confirm such a trend.

The present study, for the first time, has generated the taxonomic identities of bacterial communities present in New Zealand ground water ecosystems. Interestingly, the metabolic activities of some of the major microbial species in each hydrochemical categories are generally supported by the oxidative state of groundwater. For example,

Janthinobacterium found as a major component in our oxidized water samples, was identified as a Mn-oxidizing bacteria (Carmichael *et al*, 2013). *Telluria* that was present in hydrochemical category 1A is a methane-oxidizing bacteria (Brigmon *et al*, 2002). These species are capable of contributing to the oxidized state of groundwater. In addition, *Methylothermobacter* that was present in highly reduced groundwater (Category 2b), is recognized as an obligate methylotroph, capable of degrading methanol and methylamine (Kalyuzhnaya *et al*. 2006; Lidstrom 2006). Its metabolic activities can result in reduced groundwaters. However, it is important to note that only approximately 600 base pair region of the 16S rRNA gene was used for pyrosequencing and it may not provide accurate taxonomic identities of the bacteria especially at the genus level or even higher taxonomic levels for novel species.

Interestingly, *Pseudomonas* was the most dominant genus regardless of hydrochemical conditions. We suggest that this species could also follow the general trend shown by other species if the genetic information it contains is gained from other species in the ecosystem by horizontal gene transfer. This hypothesis supports the idea that the universal properties of an ideal bacterial species may not be reflected by taxonomically named species, but by the ecotypes that are occupying the same ecological niche (Cohan 2002). Therefore, a whole genome analysis for the bacterial isolates obtained from samples that contain *Pseudomonas* as the dominant genus in diverse chemistries should be conducted.

It is interesting to observe the presence of *Cyanobacteria* in oxidized groundwater with high human impact as *Cyanobacteria* are usually photosynthetic microorganisms and it is unusual to reveal them in subsurface groundwater ecosystems. However, we speculate that possibly this could be the phylum *Melainabacteria*, a sibling phylum of *Cyanobacteria* that does not have photosynthetic capability and was recently identified in groundwater and

human gut (Di Rienzi *et al.* 2013; Hofer 2013). Further studies are required to confirm the identity of this phylum as little sequence information is available on *Melainabacteria*.

One of the key aspects in any DNA-based microbial diversity analysis is to extract DNA from all the species present in the particular environment. However, it is not guaranteed that this is possible as some of the species may have thick cell walls that obstruct DNA recovery. Donachie *et al.* (2007) revealed that culture based methods can find new species that are not identified with molecular approaches. Therefore, we suggest that it will be useful to analyse some of these samples using culturing-based methods to detect the missing taxa in our pyrosequencing approach. Further, the relative abundance of a particular taxon determined by rDNA may not necessarily reflect the fraction of actively present microorganisms, as 16S rRNA:rDNA ratios can be influenced by environmental factors (Campbell & Kirchman, 2013). Hence, a combination of 16S rRNA and rDNA analysis would also provide a better insight into the interactions between bacterial taxa and their environment.

Overall, our findings provide a novel insight into the bacterial diversity of groundwater ecosystems and generate a solid platform for further studies on more specific interactions between the biotic and abiotic components. Further, our study reflects the potential of using bacterial communities as biological signatures to evaluate the health of groundwater ecosystems because certain environmental pressures or trends may not be visible through hydrochemical monitoring alone.

Acknowledgments

We would like to thank the groundwater research staff members from the 15 regional councils for their valuable support in sample collection. We would also like to thank D. A. Eccles for his useful advice regarding pyrosequencing data analyses.

Funding: This project was financially supported by public research funding provided to GNS Science by the Government of New Zealand.

Competing interests: The authors declare that there are no conflicts of interest.

References

- Anderson MJ, Gorley RN, Clarke KR 2008. PERMANOVA+ for PRIMER: guide to software and statistical methods, PRIMER-E Ltd, Plymouth, UK.
- Ashby MN, Rine J, Mongodin EF, Nelson KE, Dimster-Denk D (2007) Serial analysis of rRNA genes and the unexpected dominance of rare members of microbial communities. *Applied and Environmental Microbiology*, **73**, 4532–4542.
- Bent SJ, Pierson JD, Forney LJ (2007) Measuring species richness based on microbial community fingerprints: the emperor has no clothes. *Applied and Environmental Microbiology*, **73**, 2399–2401.
- Bottos EM, Woo AC, Zawar-Reza P, Pointing SB, Cary SC (2014) Airborne Bacterial Populations Above Desert Soils of the McMurdo Dry Valleys, Antarctica. *Microbial Ecology*, **67**, 120–128.
- Campbell BJ, Kirchman DL (2013) Bacterial diversity, community structure and potential growth rates along an estuarine salinity gradient. *ISME Journal*, **7**, 210–220.
- Carmichael MJ, Carmichael SK, Santelli CM, Strom A & Bräuer SL (2013) Mn (II)-oxidizing bacteria are abundant and environmentally relevant members of ferromanganese deposits in caves of the upper Tennessee River Basin. *Geomicrobiology Journal* **30**, 779–800.
- Clarke KR, Warwick RM (2001) Change in marine communities: an approach to statistical analysis and interpretation. PRIMER-E Ltd, Plymouth, UK.
- Cohan FM (2002) What are bacterial species? *Annual Review of Microbiology* **56**, 457–87.
- Daughney CJ, Reeves RR (2005) Definition of hydrochemical facies in the New Zealand National Groundwater Monitoring Programme. *Journal of Hydrology (New Zealand)*, **44**, 105–130.

- Daughney CJ, Morgenstern U, van der Raaij R, Reeves RR (2010) Discriminant analysis for estimation of groundwater age from hydrochemistry and well construction: Application to New Zealand aquifers. *Hydrogeology Journal*, **18**, 417–428.
- Di Rienzi SC, Sharon I, Wrighton KC, Koren O, Hug LA, Thomas BC, Goodrich JK, Bell JT, Spector TD, Banfield JF et al. (2013) The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. *eLife*, **2**.
- Donachie SP, Foster JS & Brown MV (2007) Culture clash: challenging the dogma of microbial diversity. *ISME Journal* **7**, 97–102.
- Gilpin B, Devane M, Nourozi F, Robson B, Scholes P, Lin S (2013) Recommendations for the processing and storage of water samples before polymerase chain reaction (PCR) analysis. *New Zealand Journal of Marine and Freshwater Research*, **47**, 582–586.
- Griebler C, Lueders T (2009) Microbial biodiversity in groundwater ecosystems. *Freshwater Biology*, **54**, 649–677.
- Griebler C, Stein H, Kellermann C, Berkhoff S, Brielmann H, Schmidt S, Selesi D, Steube C, Fuchs A, Hahn HJ (2010) Ecological assessment of groundwater ecosystems – vision or illusion? *Ecological Engineering*, **36**, 1174–1190.
- Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews*, **68**, 669–685.
- Handelsman J, Tiedje JM, Alvarez-Cohen L, Ashburner M, Cann IKO, Delong EF et al. (2007) *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. The National Academies Press, Washington, DC, United States.
- Hofer U (2013) Bacterial evolution: Getting to the bottom of Cyanobacteria. *Nature Reviews Microbiology*, **11**, 818–819.

- Kalyuzhnaya MG, Bowerman S, Lara JC, Lidstrom ME & Chistoserdova L (2006) *Methylothermobacter mobilis* gen. nov., sp. nov., an obligately methylamine-utilizing bacterium within the family Methylophilaceae. *International Journal of Systematic and Evolutionary Microbiology* **56**, 2819–2823.
- Korbel KL, Hose GC (2011) A tiered framework for assessing groundwater ecosystem health. *Hydrobiologia* **661**, 329–349.
- Kruskal JB (1964a) Multidimensional-Scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, **29**, 1–27.
- Kruskal JB (1964b) Nonmetric Multidimensional Scaling – a numerical method. *Psychometrika*, **29**, 115–129.
- Lidstrom ME (2006) Aerobic methylophilic prokaryotes. *The Prokaryotes*. Edited by Balows A, Truper HG, Dworkin M, Harder W, Schleifer KH. New York, United States, Springer.
- Lozupone CA, Hamady M, Kelley ST & Knight R (2007) Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology* **73**, 1576–1585.
- Parkes RJ, Cragg BA, Bale SJ, Getliff JM, Goodman K, Rochelle PA, Fry JC, Weightman AJ, Harvey SM (1994) Deep bacterial biosphere in Pacific Ocean sediments. *Nature*, **371**, 410–413.
- Pilloni G, Granitsiotis MS, Engel M & Lueders T (2012) Testing the limits of 454 pyrotag sequencing: reproducibility, quantitative assessment and comparison to T-RFLP fingerprinting of aquifer microbes. *PLoS One* **7**, e40467.
- Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics*, **12**, 38.

- Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA et al. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, **75**, 7537–7541.
- Sirisenka KA, Daughney CJ, Moreau-Fournier M, Ryan KG, Chambers GK 2013 National survey of molecular bacterial diversity of New Zealand groundwater: relationships between biodiversity, groundwater chemistry and aquifer characteristics. *FEMS Microbiology Ecology*, **86**, 490–504.
- Stein H, Kellermann C, Schmidt SI, Briellmann H, Steube C, Berkhoff SE, Fuchs A, Hahn HJ, Thulin B, Griebler C (2010) The potential use of fauna and bacteria as ecological indicators for the assessment of groundwater quality. *Journal of Environmental Monitoring*, **12**, 242–254.
- Steube C, Richter S & Griebler C (2009) First attempts towards an integrative concept for the ecological assessment of groundwater ecosystems. *Hydrogeology Journal* **17**, 23–35.
- Sun Y, Cai Y, Liu L, Yu F, Farrell ML, McKendree W, Farmerie W (2009) ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Research*, **37**, e76.
- ter Braak CJF, Smilauer P (2012) Canoco reference manual and user's guide: software for ordination, version 5.0. Microcomputer Power, Ithaca, USA. 496 p.
- Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, **73**, 5261–5267.
- Whitman WB, Coleman DC, Wiebe WJ (1998) Perspective Prokaryotes : The unseen majority. *Proceedings of the National Academy of Science of the United States of America*, **95**, 6578–6583.

- Wood S, Smith K, Banks J et al. (2013) Molecular genetic tools for environmental monitoring of New Zealand's aquatic habitats, past, present and the future. *New Zealand Journal of Marine and Freshwater Research*, **47**, 90–119.
- Yao S, Ni J, Chen Q & Borthwick AG (2013) Enrichment and characterization of a bacteria consortium capable of heterotrophic nitrification and aerobic denitrification at low temperature. *Bioresource Technology* **127**, 151–157.
- Zinger L, Gobet A, Pommier T (2012) Two decades of describing the unseen majority of aquatic microbial diversity. *Molecular Ecology*, **21**, 1878–96.

Supplementary information

Table S1 Median values of 19 hydrochemical parameters derived from the actual values measured quarterly from March 2008 to March 2012 across the groundwater monitoring sites.

GGW ID	SO4	Na	SiO2	K	PO4-P	NO3-N	Mn	Mg	Fe	F	Cl	Ca	Br	HCO3	NH4-N	EC	Temp	DO	pH
456	3.70	17.3	11.8	2.4	0.033	<0.01	0.062	4.7	0.62	0.088	6.9	4.9	0.04	74	0.020	148	14.80	0.145	8.17
27	1.50	33.0	20.0	1.5	0.112	<0.01	0.220	4.4	0.03	0.260	9.5	15.0	0.04	149	0.290	256	14.20	0.430	7.37
362	3.75	8.4	35.0	0.9	0.039	1.50	0.005	3.2	0.02	0.110	6.7	9.1	0.10	46	0.010	112	13.25	6.170	6.88
364	0.12	85.5	61.5	6.2	1.300	<0.01	0.510	8.9	2.70	0.340	74.0	83.5	0.40	421	3.800	849	15.30	0.210	7.12
82	5.25	30.5	34.0	5.0	0.120	<0.01	0.680	9.0	1.30	0.345	25.0	101.5	0.14	392	0.600	660	15.90	0.190	6.90
83	86.50	34.5	31.5	4.5	0.008	0.27	0.360	12.2	0.22	0.260	71.0	119.5	0.33	286	0.020	790	16.00	0.300	7.00
338	0.00	88.0	14.9	7.6	0.060	<0.01	0.140	22.0	3.30	0.090	47.0	110.5	0.18	657	5.500	980	15.45	1.795	7.20
80	0.12	105.0	37.0	9.3	0.054	0.01	0.340	25.0	8.50	0.340	115.0	158.0	0.48	733	4.100	1260	15.05	1.610	7.10
17	2.60	15.9	52.0	1.8	0.080	2.60	0.005	10.6	0.02	0.063	21.0	12.2	0.05	87	0.010	208	15.30	5.765	ND
18	13.50	14.8	33.0	3.4	0.022	7.00	0.005	5.4	0.02	0.071	11.8	10.8	0.04	46	0.010	182	15.90	8.800	ND
30	5.15	23.0	26.0	2.4	0.072	0.01	0.008	12.0	0.03	0.066	23.0	31.0	0.01	170	0.490	301	15.60	ND	8.14
31	0.03	30.0	53.0	6.7	0.377	0.03	0.715	7.3	1.80	0.340	26.0	17.3	0.10	143	1.300	275	15.00	ND	7.52
36	11.40	9.1	13.1	1.1	0.004	5.10	0.005	4.5	0.02	0.050	9.5	26.0	0.04	75	0.010	212	12.70	8.650	6.45
74	14.20	4.6	14.4	0.9	0.006	2.25	0.005	7.4	0.02	0.030	6.3	23.0	0.04	85	0.010	196	13.40	7.420	6.59
69	6.00	7.4	13.5	0.9	0.007	0.49	0.005	4.4	0.02	0.120	2.8	32.0	0.04	124	0.010	204	13.80	7.250	7.22
53	6.50	10.4	13.9	1.2	0.004	1.20	0.009	2.5	0.12	0.051	12.3	6.9	0.04	32	0.010	116	14.50	2.775	6.01
39	3.00	8.6	18.5	0.7	0.007	0.27	0.005	3.1	0.02	0.120	5.1	14.2	0.04	70	0.010	132	13.70	6.000	7.50
49	4.25	9.9	10.9	0.8	0.004	0.27	0.005	2.0	0.02	0.064	13.2	5.4	0.04	26	0.010	97	13.80	3.230	6.27
467	10.80	32.0	23.0	1.4	0.043	0.03	0.480	7.5	0.97	0.155	54.0	17.0	0.18	74	0.085	331	14.00	0.060	6.34
54	9.10	18.3	17.3	2.0	0.013	1.65	0.005	4.0	0.01	0.070	25.5	6.2	0.06	33	0.010	176	14.00	3.170	5.68
42	19.00	23.0	22.0	2.8	0.020	0.04	0.020	6.1	0.36	0.140	36.0	16.1	0.10	57	0.030	265	14.40	0.150	6.42
388	14.10	15.8	26.0	0.7	0.008	6.80	0.005	11.2	0.02	0.048	21.0	27.0	0.04	94	0.010	303	10.85	8.360	6.55
3327	4.55	23.0	37.0	0.9	0.004	0.03	0.140	9.6	7.80	0.210	25.0	15.3	0.12	125	0.030	285	11.30	1.955	6.72

CHAPTER 3.3

1993	9.40	3.9	4.6	3.0	0.004	4.70	0.180	2.0	0.04	0.030	10.0	12.4	0.18	16	0.010	106	13.32	3.300	5.21
35	0.04	26.0	33.0	0.9	0.978	<0.01	0.570	9.3	1.80	0.315	4.4	36.0	0.10	220	0.280	343	12.60	1.000	7.03
458	5.05	6.0	55.5	2.2	0.054	0.26	0.005	2.9	0.12	0.042	4.1	5.3	0.04	37	0.010	84	14.00	6.135	ND
8	2.70	19.7	65.0	1.0	0.113	0.03	0.315	6.6	3.10	0.380	4.7	16.2	0.04	128	0.040	218	13.98	4.265	6.91
14	0.05	23.0	51.0	7.3	1.000	<0.01	0.380	8.5	1.60	0.250	15.7	24.0	0.05	172	3.000	315	15.03	4.655	7.36
12	9.10	13.2	44.0	2.2	0.100	0.03	0.044	6.5	0.03	0.147	16.8	35.0	0.04	138	0.350	283	14.60	0.340	7.91
380	5.80	2.4	8.3	2.0	0.007	1.10	0.005	1.0	0.08	0.047	3.2	11.1	0.04	32	0.010	66	12.57	7.350	5.95
6	1.80	24.0	26.0	0.5	0.030	0.01	0.016	8.1	0.02	0.200	5.6	33.0	0.04	196	0.020	303	13.59	1.530	7.92
395	19.65	14.2	12.8	2.1	0.015	5.70	0.005	5.2	0.02	0.050	22.0	16.2	0.05	30	0.010	225	11.00	7.200	5.82
52	11.85	26.0	19.3	1.4	0.010	2.80	0.005	6.3	0.02	0.170	25.0	9.7	0.13	65	0.010	232	15.00	1.730	6.27
389	10.20	13.6	15.0	0.8	0.021	8.00	0.006	4.8	0.02	0.062	24.0	13.4	0.14	15	0.010	205	11.10	8.530	5.35
383	4.80	2.9	13.1	2.6	0.006	1.40	0.005	1.4	0.02	0.042	3.3	9.8	0.04	34	0.010	69	13.04	5.100	5.65

Units are in mg L⁻¹ for all variables except pH which is in pH units, Electrical conductivity (EC) in (μS cm⁻¹ at 25 °C) and Temperature in (°C). ND indicates that the parameter value was not determined.

CHAPTER 3.3

Table S2 Summary of site-specific information: aquifer lithology, confinement, well depth (depth code), groundwater mean residence time (MRT class), land use activities in the aquifer recharge zone, geographical region and hydrochemical category to which each site belongs.

GGW ID	Aquifer Lithology	Aquifer Confinement	Well Depth	Depth Code	Mean Residence Time (MRT)	MRT Class	Land Use	Region	Hydrochemical Category – 2 Levels	Hydrochemical Category – 4 Levels
456	Sand	Unconfined	83.50	Deep	32	B	Horticultural	Marlborough	2	2B
27	Gravel	Confined	25.19	Mid	121	D	Urban	Hawke's Bay	2	2A
362	Gravel	Unknown	37.50	Mid	47	C	Agriculture	Hawke's Bay	1	1B
364	Unknown	Unknown	32.00	Mid	74	C	Agriculture	Hawke's Bay	2	2B
82	Unknown	Semi-confined	24.40	Mid	43	C	Horticultural	Gisborne	2	2B
83	Unknown	Unconfined	4.10	Shallow	50	C	Horticultural	Gisborne	2	2A
338	Gravel	Confined	99.00	Deep	170	D	Horticultural	Gisborne	2	2B
80	Gravel	Confined	127.00	Deep	150	D	Horticultural	Gisborne	2	2B
17	Basalt	Unknown	35.00	Mid	125	D	Agriculture	Waikato	1	1A
18	Gravel	Unconfined	5.00	Shallow	1	A	Dairy	Waikato	1	1A
30	Sand	Confined	234.60	Deep	144	D	Agriculture	Taranaki	2	2A
31	Sand	Confined	171.30	Deep	152	D	Agriculture	Taranaki	2	2B
36	Gravel	Unknown	8.00	Shallow	3	A	Agriculture	Canterbury	1	1A
74	Gravel	Unconfined	14.50	Mid	2	A	Horticultural	Tasman	1	1B
69	Gravel	Unconfined	32.30	Mid	40	B	Agriculture	Otago	1	1B
53	Gravel	Unconfined	20.40	Mid	1	A	Industrial	Wellington	1	1B
39	Gravel	Confined	200.20	Deep	155	D	Urban	Canterbury	1	1B
49	Gravel	Semi-confined	25.00	Mid	1	A	Urban	Wellington	1	1B
467	Gravel	Confined	16.00	Mid	55	C	Dairy	Wellington	2	2A
54	Gravel	Unconfined	8.00	Shallow	3	A	Park/Reserve	Wellington	1	1A
42	Gravel	Confined	14.80	Mid	42	C	Park/Reserve	Wellington	2	2A

CHAPTER 3.3

388	Gravel	Unconfined	5.00	Shallow	1	A	Agriculture	Southland	1	1A
3327	Unknown	Unknown	31.00	Mid	NA	NA	NA	Southland	2	2A
1993	Gravel	Unknown	15.00	Mid	35	B	Agriculture	West Coast	1	1A
35	Gravel	Confined	28.00	Mid	163	D	Agriculture	Canterbury	2	2B
458	Sand	Unknown	5.00	Shallow	1	A	Agriculture	Waikato	1	1B
8	Gravel	Confined	236.00	Deep	110	D	Horticultural	Tasman	2	2A
14	Gravel	Confined	33.50	Mid	107	D	Horticultural	Manawatu-Wanganui	2	2B
12	Gravel	Confined	34.00	Mid	150	D	Horticultural	Manawatu-Wanganui	2	2A
380	Gravel	Unknown	3.00	Shallow	2	A	Agriculture	West Coast	1	1B
6	Gravel	Confined	337.00	Deep	110	D	Horticultural	Tasman	2	2A
395	Gravel	Unconfined	8.00	Shallow	1	A	Agriculture	Southland	1	1A
52	Gravel	Semi-confined	10.20	Mid	35	B	Urban	Wellington	1	1A
389	Gravel	Unconfined	9.40	Shallow	10	A	Agriculture	Southland	1	1A
383	Unknown	Unknown	10.00	Shallow	NA	NA	NA	West Coast	1	1B

Well depths are expressed in meters (m) and depth codes are defined as: Shallow – ≤ 10 m; Mid – 11 to 50 m; and Deep – ≥ 51 m. Groundwater mean residence times (MRT) are expressed in years and MRT classes are defined as: A – ≤ 10 years; B – 11 to 40 years; C – 41 to 100 years; and D - ≥ 101 years (Daughney *et al.* 2010). Hydrochemical categories are defined using the median hydrochemical values as described by Daughney & Reeves (2005).

Table S3 Summary of richness and abundance of unique OTUs in each sample.

Hydrochemical categories ^a	GGW ID	Richness of unique OTUs			Abundance of unique OTUs		
		# of unique OTUs	Total # of OTUs	% of unique OTU richness	# of reads in unique OTUs	Total # of reads	% of unique OTU abundance
1A Oxidized groundwater with high human impact	54	220	277	79.42	3900	8706	44.8
	1993	849	947	89.65	2268	6391	35.49
	395	173	255	67.84	442	1313	33.66
	52	157	240	65.42	1485	4593	32.33
	36	184	268	68.66	263	9649	2.73
	389	109	178	61.24	195	7940	2.46
	388	59	126	46.83	77	9594	0.8
	18	45	83	54.22	57	13836	0.41
	17	9	29	31.03	10	10007	0.1
1B Oxidized groundwater with low human impact	49	195	258	75.58	432	3848	11.23
	53	347	471	73.67	720	9056	7.95
	458	64	116	55.17	77	1714	4.49
	39	46	86	53.49	159	5473	2.91
	362	131	217	60.37	224	7938	2.82
	380	317	494	64.17	576	21410	2.69
	383	100	171	58.48	175	8616	2.03
	74	52	97	53.61	87	9564	0.91
	69	19	60	31.67	31	9231	0.34
2A Moderately reduced groundwater	12	257	366	70.22	730	3690	19.78
	3327	202	268	75.37	776	4089	18.98
	8	144	220	65.45	267	3710	7.2
	42	131	206	63.59	163	3590	4.54
	83	217	290	74.83	487	12881	3.78
	467	60	128	46.88	94	3340	2.81
	27	61	104	58.65	437	21334	2.05
	30	47	87	54.02	94	5795	1.62
	6	110	191	57.59	317	25332	1.25
2B Highly reduced groundwater	35	162	216	75	1592	6685	23.81
	80	373	481	77.55	1330	6030	22.06
	364	359	452	79.42	752	5834	12.89
	456	298	397	75.06	745	7351	10.13
	14	97	149	65.1	263	3620	7.27
	82	120	199	60.3	187	7603	2.46
	31	87	167	52.1	130	7524	1.73
	338	26	41	63.41	79	4609	1.71

^a defined as explained by Daughney & Reeves (2005)

Table S4 Summary of richness and abundance of shared OTUs in each sample.

Hydrochemical categories ^a	GGW ID	Richness of shared OTUs			Abundance of shared OTUs		
		# of OTUs shared ^b	Total # of OTUs	% of shared OTU richness	# of reads in shared OTUs ^b	Total #of reads	% of shared OTU abundance
1A Oxidized groundwater with high human impact	17	7	29	24.14	9978	10007	99.71
	388	16	126	12.7	9022	9594	94.04
	389	19	178	10.67	7441	7940	93.72
	36	19	268	7.09	9002	9649	93.29
	18	11	83	13.25	9679	13836	69.96
	52	21	240	8.75	1964	4593	42.76
	395	19	255	7.45	546	1313	41.58
	1993	17	947	1.8	2507	6391	39.23
	54	13	277	4.69	2575	8706	29.58
1B Oxidized groundwater with low human impact	69	11	60	18.33	8989	9231	97.38
	74	11	97	11.34	9229	9564	96.5
	383	12	171	7.02	7976	8616	92.57
	362	16	217	7.37	7127	7938	89.78
	458	16	116	13.79	1512	1714	88.21
	380	29	494	5.87	17415	21410	81.34
	53	20	471	4.25	6883	9056	76
	49	14	258	5.43	1514	3848	39.35
	39	12	86	13.95	1856	5473	33.91
2A Moderately reduced groundwater	30	12	87	13.79	5396	5795	93.11
	6	19	191	9.95	23066	25332	91.05
	42	16	206	7.77	3148	3590	87.69
	83	16	290	5.52	10622	12881	82.46
	467	16	128	12.5	2683	3340	80.33
	8	20	220	9.09	2200	3710	59.3
	27	18	104	17.31	11476	21334	53.79
	12	24	366	6.56	1720	3690	46.61
	3327	14	268	5.22	304	4089	7.43
2B Highly reduced groundwater	31	13	167	7.78	6610	7524	87.85
	14	14	149	9.4	3081	3620	85.11
	338	3	41	7.32	3911	4609	84.86
	82	12	199	6.03	6266	7603	82.41
	456	19	397	4.79	5340	7351	72.64
	364	12	452	2.65	3785	5834	64.88
	35	13	216	6.02	1940	6685	29.02
	80	17	481	3.53	733	6030	12.16

^a defined as explained by Daughney & Reeves (2005)^b OTUs shared among 10 or more samples

Table S5 Shannon diversity indices and number of OTUs based on 454 pyrosequencing data and T-RFLP data presented in Sirisena *et al* (2013)

Hydrochemical categories ^a	GGW ID	Shannon Index (H')			Number of OTUs		
		454-(H')	T-RFLP (H')		454 OTUs	T-RFLP OTUs	
			FAM-(H')	HEX-(H')		FAM OTUs	HEX OTUs
1A Oxidized groundwater with high human impact	17	0.34	1.49	1.47	29	8	12
	18	0.99	2.14	2.17	83	13	15
	36	1.33	0.49	0.19	268	5	3
	52	3.27	2.54	2.29	240	18	15
	54	2.26	2.12	2.29	277	13	14
	388	1.18	1.94	1.10	126	12	6
	389	1.25	1.75	0.92	178	9	7
	395	3.96	2.96	2.42	255	24	14
	1993	3.98	2.11	2.18	947	13	9
1B Oxidized groundwater with low human impact	39	1.99	1.32	1.51	86	7	7
	49	2.75	2.30	2.03	258	16	13
	53	2.58	2.56	1.27	471	18	9
	69	0.55	1.87	1.49	60	12	12
	74	0.93	2.18	0.89	97	13	5
	362	1.71	0.90	1.07	217	7	7
	380	1.89	2.26	2.18	494	14	13
	383	1.35	2.14	1.98	171	12	9
	458	1.30	2.59	2.20	116	18	13
2A Moderately reduced groundwater	6	1.04	1.59	1.37	191	8	9
	8	2.55	1.83	1.90	220	9	8
	12	3.38	1.98	1.82	366	10	15
	27	1.23	0.67	0.40	104	5	4
	30	0.78	1.54	1.21	87	7	9
	42	1.93	2.40	2.26	206	16	14
	83	1.70	1.51	1.79	290	7	16
	467	2.08	1.58	1.82	128	8	11
	3327	2.10	2.40	2.02	268	13	13
2B Highly reduced groundwater	14	1.16	2.39	1.89	149	16	10
	31	1.79	1.88	1.50	167	10	11
	35	2.68	1.43	1.16	216	7	5
	80	3.27	1.63	1.34	481	13	9
	82	1.97	1.81	2.27	199	11	22
	338	0.71	1.66	1.58	41	9	8
	364	2.48	0.27	0.10	452	2	2
	456	2.90	2.00	2.09	397	13	18

^a defined as explained by Daughney & Reeves (2005)

Table S6 Summary of the contribution of each bacterial species for the similarity within each hydrochemical category

Hydrochemical category	Species	Relative abundance	% Contribution for similarity	Cumulative % contribution for similarity
1A Oxidized groundwater with high human impact	<i>Janthinobacterium</i>	20.90	43.76	43.76
	<i>Pseudomonas</i>	21.06	29.12	72.87
	<i>Variovorax</i>	3.38	9.25	82.12
	<i>Herbaspirillum</i>	10.77	5.70	87.82
	<i>Polaromonas</i>	2.22	2.86	90.69
1B Oxidized groundwater with low human impact	<i>Variovorax</i>	18.85	40.02	40.02
	<i>Pseudomonas</i>	23.24	32.42	72.44
	<i>Herbaspirillum</i>	16.85	19.40	91.84
2A Moderately reduced groundwater	<i>Pseudomonas</i>	31.76	76.84	76.84
	<i>Variovorax</i>	3.52	8.04	84.87
	<i>Georgfuchsia</i>	3.18	2.73	87.60
	<i>Burkholderia</i>	6.06	2.54	90.14
2B Highly reduced groundwater	<i>Pseudomonas</i>	13.51	46.60	46.60
	<i>Methylothera</i>	15.74	20.83	67.43
	<i>Variovorax</i>	3.21	10.82	78.25
	<i>Marinospirillum</i>	2.54	5.45	83.70
	<i>Acidovorax</i>	1.50	3.47	87.17
	<i>Methylobacter</i>	1.91	2.15	89.32
	<i>Acinetobacter</i>	6.27	1.34	90.66

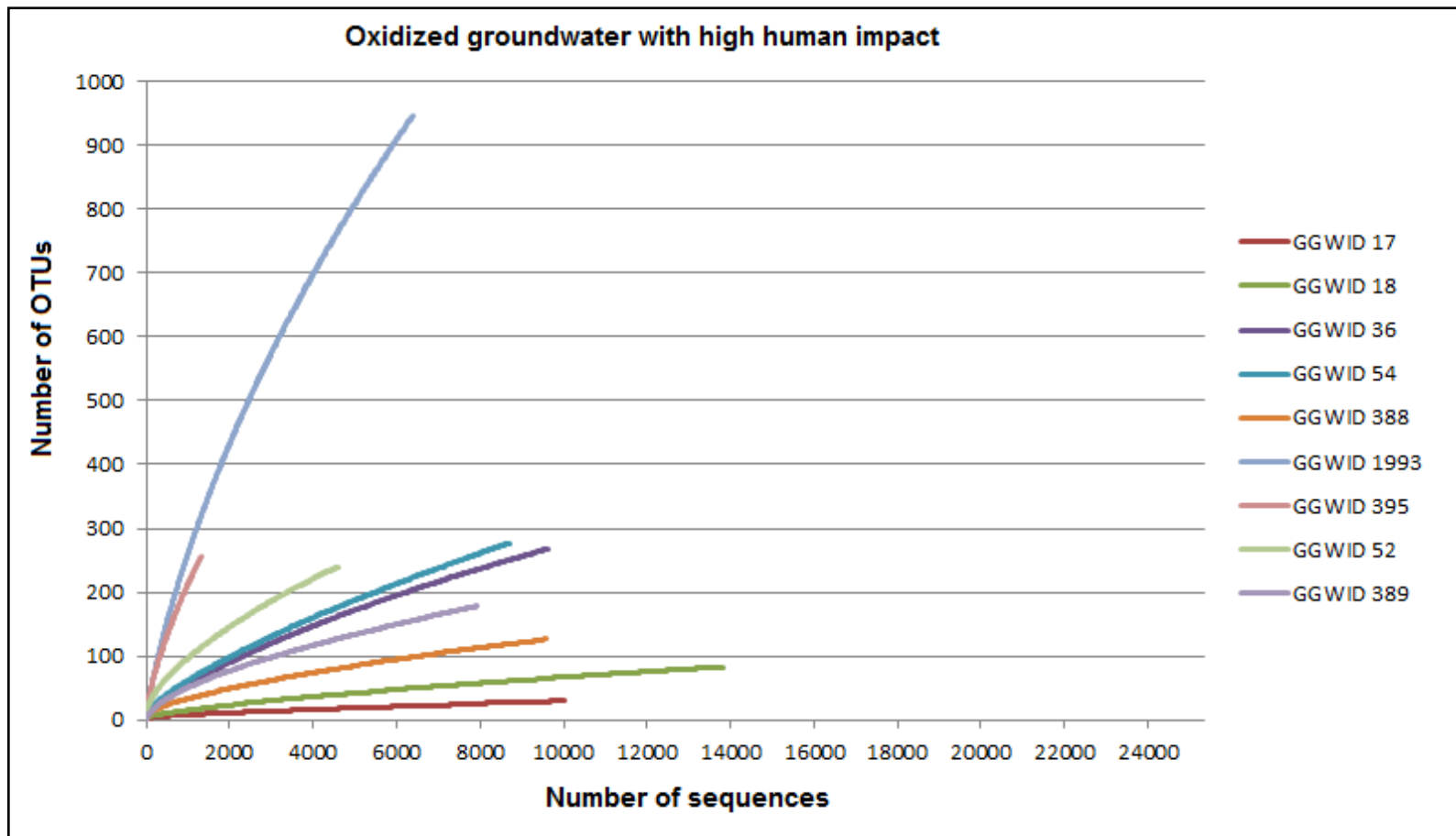


Figure S1 A. Rarefaction curves for oxidized groundwater samples with high human impact.

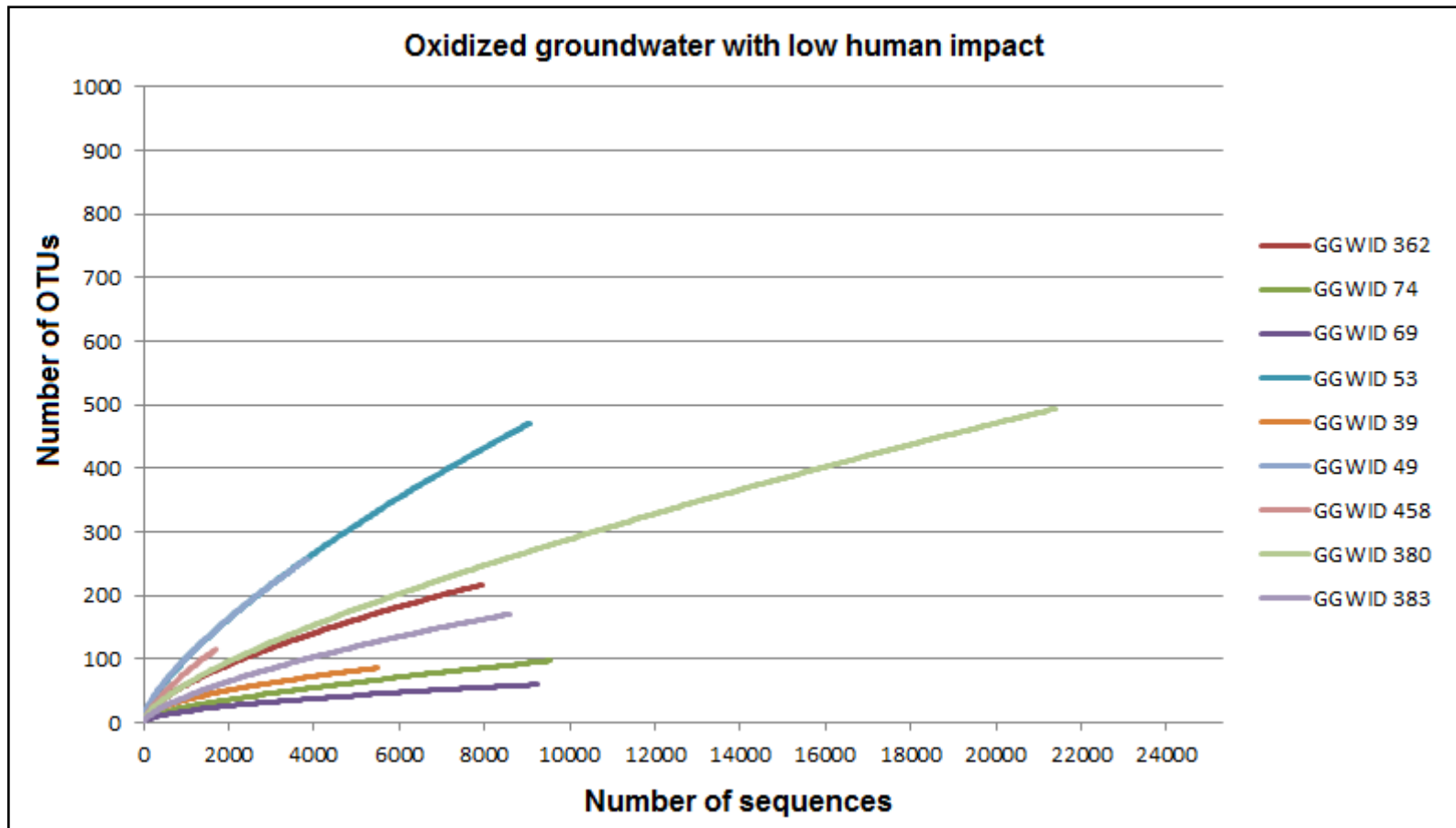


Figure S1 B. Rarefaction curves for oxidized groundwater samples with low human impact.

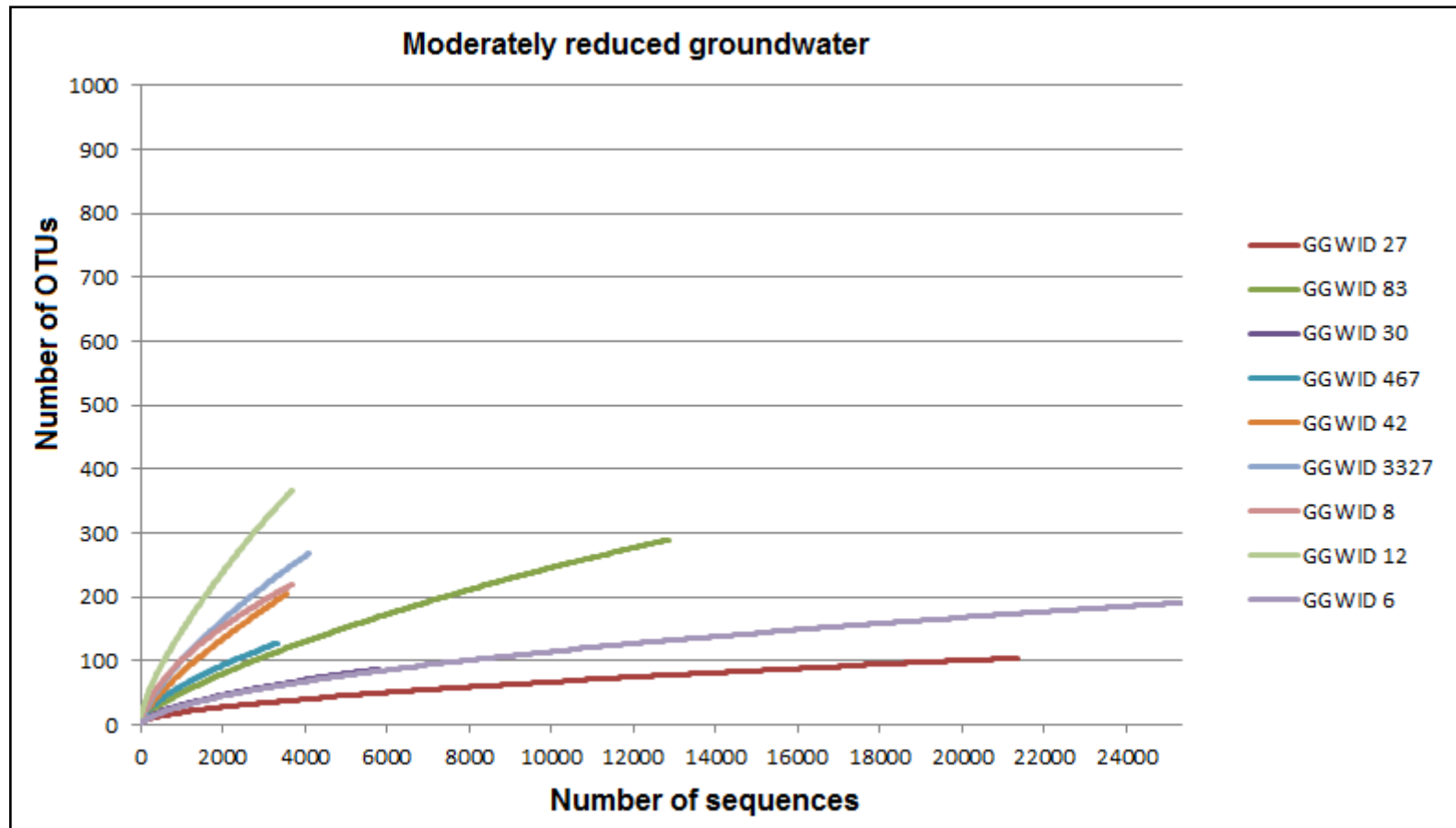


Figure S1 C. Rarefaction curves for moderately reduced groundwater samples.

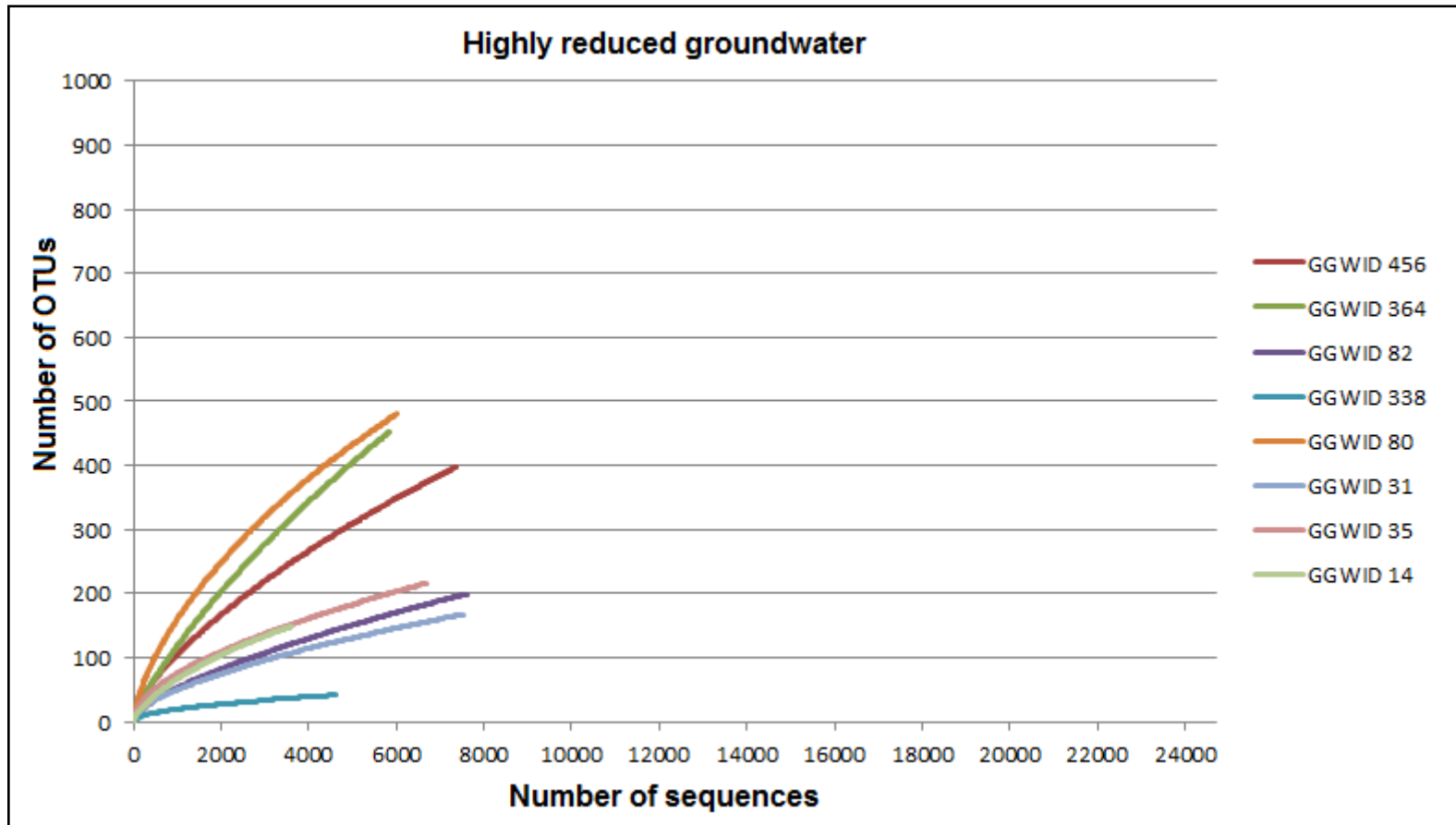


Figure S1 D. Rarefaction curves for highly reduced groundwater samples.

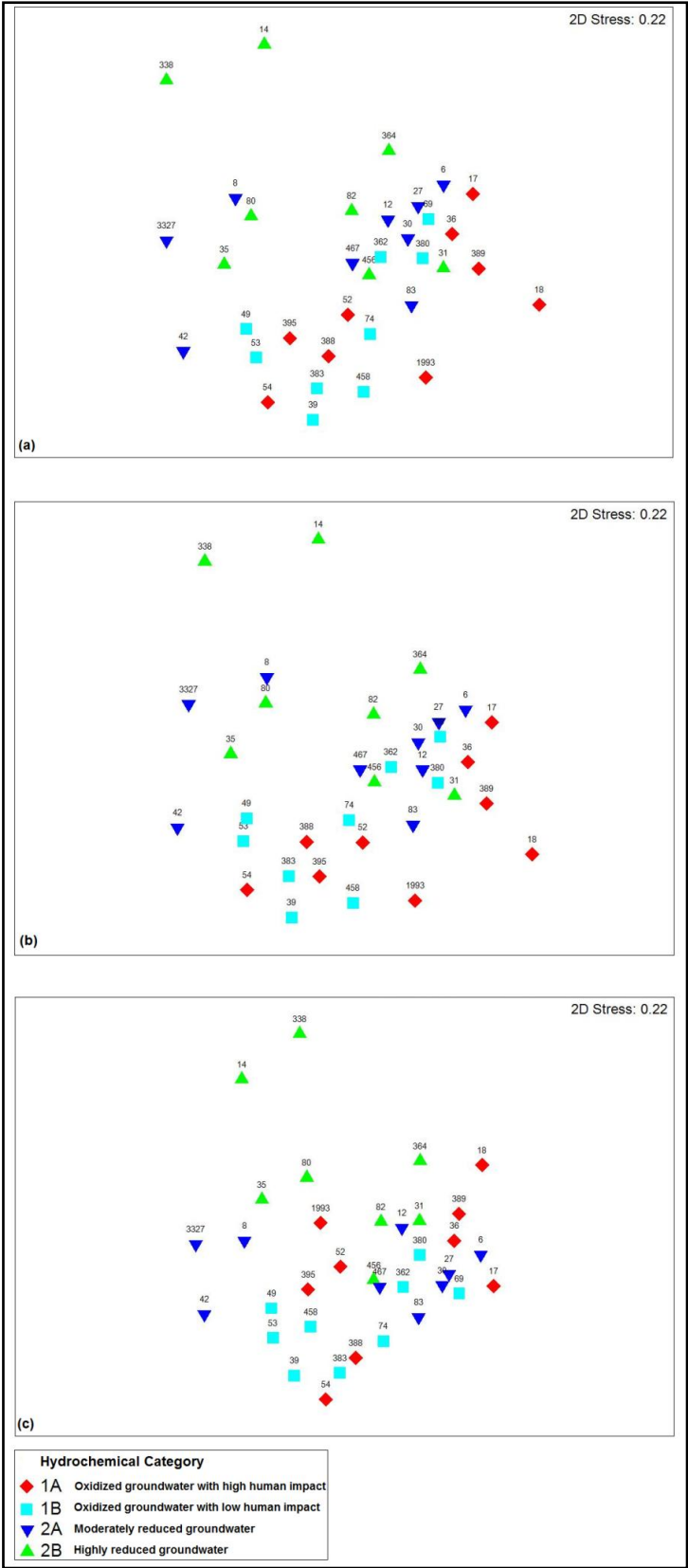


Figure S2. Non-metric multidimensional scaling based on the relative abundances of: (a) all OTUs; (b) all OTUs except singletons; and (c) the 100 most abundant OTUs. Discrimination of samples was based on the four hydrochemical categories and displayed on 2D plots with a final stress of 0.22.

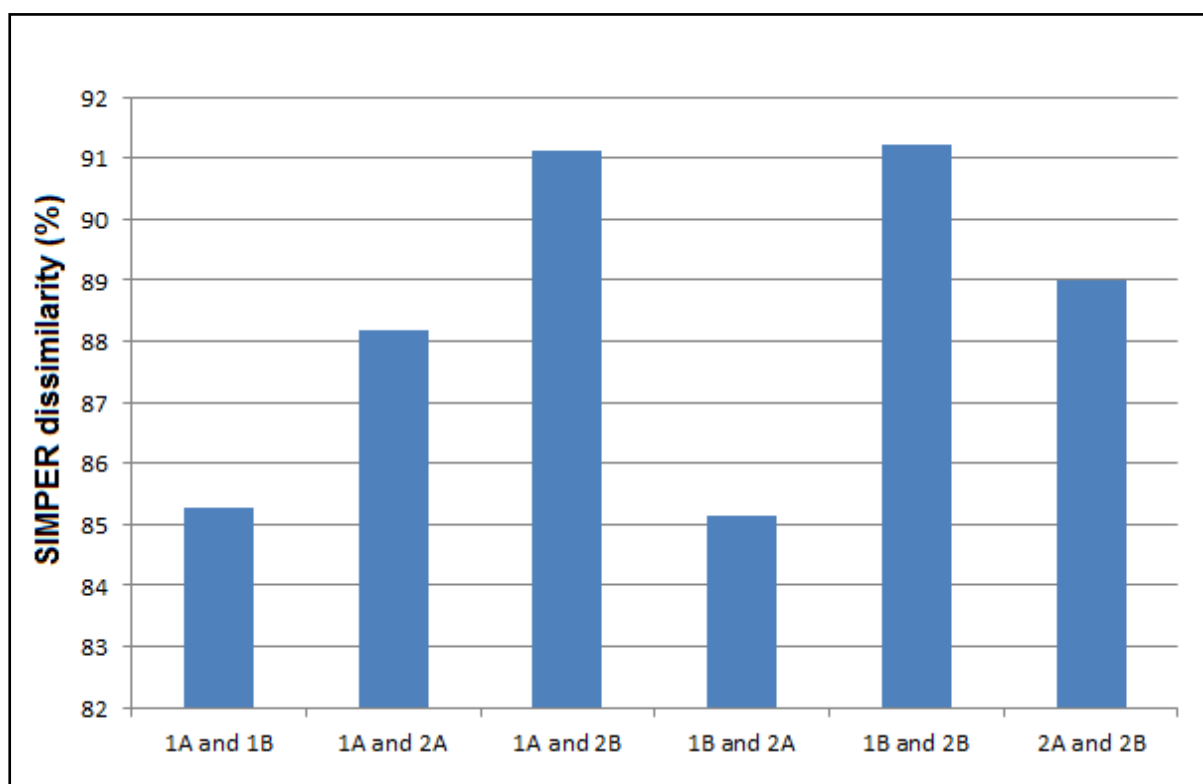


Figure S3. Percentage dissimilarity between each pair of hydrochemical categories.

PRELIMINARY OVERVIEW OF HORIZONTAL GENE TRANSFER IN GROUNDWATER BACTERIA

Abstract

All organisms including bacteria adapt to changing environments. Horizontal gene transfer (HGT) is a key method that facilitates the exchange of genetic materials between bacterial species. Previous studies have revealed that *Pseudomonas spp.* are among the dominant species in New Zealand groundwater, across diverse hydrochemical and environmental conditions, and we propose that *Pseudomonas spp.* have acquired genes from other species. To test this hypothesis, the bacterial metagenomes from six representative groundwater environments were subjected to high throughput DNA sequencing using the Illumina MiSeq[™] platform. The whole genome sequencing results are in accord with previous Roche 454 sequencing data and T-RFLP based bacterial community structure. *De novo* assembly suggests that estimated genome sizes are larger than the expected sizes and this supports our hypothesis. However, further analysis should be conducted to determine whether this size difference is purely due to the samples being mixtures of species or an indication of the HGT between bacterial species. The mapping of short reads into the contigs also implied the possible occurrence of HGT. Further bioinformatics analyses will be conducted to gain a better understanding of the genome structure of these samples and to postulate the underlying biological process that shapes the groundwater bacterial genetic composition.

INTRODUCTION

All organisms including bacteria adapt to changing environments to ensure the survival of the species. As opposed to the normal process of inheritance where genetic information passes from parent cells to offspring, horizontal gene transfer (HGT) has been identified as a key method that facilitates the exchange of genetic material between distinct prokaryotic species, allowing acquisition of new traits and adaptation to different habitats (Eisen 2000; De la Cruz & Davies 2000; Gogarten & Townsend 2005; Pál et al. 2005; McDaniel et al. 2010). Further, operational genes have a higher tendency to transfer horizontally than informational genes that are involved in DNA replication, transcription and translation, and the genes that have higher expression rates are less likely to be subjected to HGT (Jain et al. 1999; Park & Zhang 2012).

The occurrence of HGT means that it is not possible to define boundaries between prokaryotic species. Since the introduction of DNA-DNA hybridization techniques in the late 1970s, two bacterial isolates are considered to belong to the same species if the total DNA of the isolates shows a homology of more than 70% under standard hybridization conditions (Achtman & Wagner 2008). Presently, bacterial isolates that show 99% identity of 16S rRNA gene sequences are considered as a single species with a few exceptions (Medini et al. 2008). However, it is believed that, unlike in eukaryotes, the universal properties of an ideal bacterial species may not be reflected by taxonomically named species, but by the ecotypes that are occupying the same ecological niche. Thus, taxonomically defined bacterial species can be considered more as genera than species (Cohan 2002). Although early microbiological studies date back centuries, the establishment of the definition for bacterial species is hindered due to difficulties in morphological characterization of microorganisms and lack of the availability of advanced microbiological

technologies. However, the recent development of genomic technologies provides immense potential to resolve the current uncertainties about species definitions and to better understand the ecological aspects of the processes such as HGT that help microbiota to adapt into changing environments.

The previous three chapters of this thesis provide an insight into the bacterial community composition of New Zealand groundwater systems and the underlying factors that shape the microbial diversity. The results in chapters 3.1 and 3.2 (Sirisena et al. 2013, 2014) suggested that bacterial diversity is mainly related to groundwater chemistry and not to environmental factors such as aquifer lithology or surrounding land use. The pyrosequencing data in Chapter 3.3 permitted a more precise investigation of the relationships between abiotic and biotic components, and confirmed that hydrochemistry, in particular the redox potential, was the key factor that shaped the groundwater bacterial community structure.

Despite the relationships between bacterial diversity and hydrochemistry, Chapters 3.1, 3.2 and 3.3 also revealed that some bacterial species are present across a wide range of environments, even where the hydrochemistry is radically different. The T-RFLP studies showed that one particular taxon, as represented by a unique combination of terminal restriction fragments, was present with high abundance in many sites across highly diverse hydrochemical and environmental conditions. Chapter 3.3 identified this as *Pseudomonas spp.*

This chapter describes a solid platform to understand the genetic aspects of *Pseudomonas spp.*, in particular why it may be found across a range of hydrochemical conditions in New Zealand groundwater. Here, I suggest that HGT helped the dominant *Pseudomonas spp.* to survive differing hydrochemical and environmental conditions. Therefore, these dominant species may have acquired genetic material from other species in the environment, some of which may not be close relatives. To test this hypothesis, the

bacterial metagenomes from six representative groundwater environments were subjected to high throughput DNA sequencing using the Illumina MiSeq[™] platform. This resulted in an enormous amount of sequence data, and the complete analysis of the entire data set is beyond the scope of my PhD project. Therefore, the main objective of this brief chapter is to provide a preliminary overview of the approach of setting up six genomic databases as a resource to search for hypothetical HGT events. Further analyses will be conducted based on these results after the submission of the thesis to explore the microbial genome changes that occur as a response to the adaptation of microbial communities into diverse habitats.

MATERIALS AND METHODS

Groundwater site selection

For this study, groundwater sampling sites were selected on the basis of two main criteria: 1) sites that confirmed/indicated the presence of *Pseudomonas* as the dominant genus; and 2) sites that are located in different geographical regions with different hydrochemical conditions. Previous work (Chapters 3.2 and 3.3) has indicated sites that are dominated by *Pseudomonas*. The chemical states of the sites were determined according to the hydrochemical facies described by Daughney & Reeves 2005 (Table 1). In this categorization, groundwaters were grouped at three thresholds according to the redox potential of the water and the degree of human impact on the aquifer recharge zone.

Table 1 General characteristics of the hydrochemical categories at the three thresholds. This table is reproduced after Daughney & Reeves 2005.

Cluster at Threshold 1	Facies Description	Cluster at Threshold 2	Facies Description	Cluster at Threshold 3	Facies Description
1	Surface-dominated Oxidised Unconfined aquifer Low to moderate total dissolved solids Ca-Na-Mg-HCO ₃ water	1A	Signs of human impact	1 A-1	Moderate human impact
			Rainfall recharge		Carbonate or calstic aquifer
			Moderate total dissolved solids		Ca-Na-Mg-HCO ₃ -Cl water
		1 B	Na-Ca-Mg-HCO ₃ -Cl water	1 A-2	Most human impact
			Little human impact		Volcanic or volcanoclastic aquifer
			River recharge	1 B-1	Na-Ca-Mg-HCO ₃ -Cl water
2	Groundwater dominated Reduced Higher total dissolved Solids Ca-Na-HCO ₃ water		Low total dissolved solids	1 B-2	Carbonate or classic aquifer
			Ca-Na-HCO ₃ water		Ca-HCO ₃ water
				2 A	Volcanic or volcanoclastic aquifer
					Na-Ca-Mg-HCO ₃ -Cl water
				2 B	Moderately reduced
					Majority unconfined
					High total dissolved solids
					Highly reduced
					Majority confined
					Highest total dissolved solids

After considering these results and information, six groundwater monitoring sites were selected for the study: GGW ID sites 11, 24, 364, 17, 69; and 79 (Fig. 1).

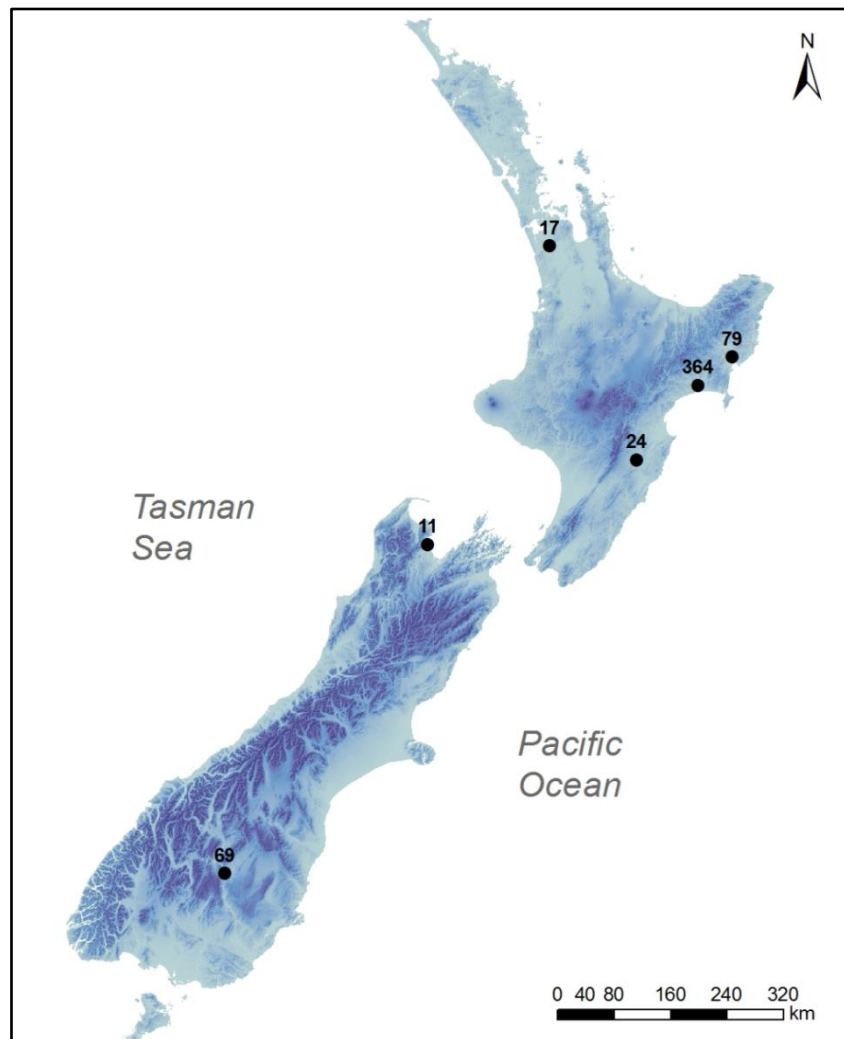


Figure 1 Groundwater sampling sites across New Zealand. The GNS Science Geothermal and Groundwater database ID (GGW ID) numbers are indicated on the map. Further information pertaining to these sites can be obtained from the GGW database website (<http://ggw.gns.cri.nz/ggwdata/>).

The six sites are located in five geographical regions and belonged to five hydrochemical categories at threshold 3 (at which a total of 6 categories are defined by Daughney & Reeves, 2005). The categorisation of the sites was based on the median hydrochemical values determined from samples analysed between 2008 and 2012. Additional information on these sites is available in the GGW database website (<http://ggw.gns.cri.nz/ggwdata/>) and is summarized in Tables 2 and 3.

Table 2 Summary of the site specific information of groundwater sampling sites.

GGW ID	Well Depth(m)	Region	Aquifer Lithology	Aquifer Confinement	Land use	Mean Residence Time (MRT) (y)	Hydrochemical category*
24	22.56	Hawke's Bay	Greywacke	Semi-confined	Agriculture	51	1A-1
17	35	Waikato	Basalt	Unknown	Agriculture	125	1A-2
69	32.3	Otago	Gravel	Unconfined	Agriculture	40	1B-1
11	9.4	Tasman	Gravel	Unconfined	Horticultural	40	2A
364	32	Hawke's Bay	Unknown	Unknown	Agriculture	74	2B
79	51.8	Gisborne	Pumice	Confined	Horticultural	148	2B

* defined as described by Daughney & Reeves (2005) where each category represents a particular combination of hydrochemistry and the degree of human impact.

Table 3 Median hydrochemical values of 19 parameters derived from the actual values measured quarterly from March 2008 to March 2012.

GGW ID	SO ₄	Na	SiO ₂	K	PO ₄	NO ₃ -N	Mn	Mg	Fe	F
24	7.3	13.4	23.5	1.2	0.01	4.40	0.61	7.60	<0.02	0.23
17	2.6	15.9	52.0	1.8	0.08	2.60	<0.01	10.60	<0.02	0.06
69	6.0	7.3	13.5	0.9	0.01	0.49	<0.01	4.40	<0.02	0.12
11	26.0	6.1	25.0	1.0	0.01	<0.01	2.60	13.85	1.20	0.06
364	0.1	85.5	61.5	6.2	1.30	<0.01	0.51	8.90	2.70	0.34
79	1.0	65.5	27.0	5.0	0.05	<0.01	0.71	11.40	0.46	0.37

Table 3 Continued.

GGW ID	Cl	Ca	Br	HCO ₃	NH ₄ -N	EC	Temperature	DO	pH
24	12.1	11.8	<0.01	66.0	<0.01	196.0	13.9	0.16	6.68
17	21.0	12.2	0.05	87.0	<0.01	208.0	15.3	5.76	ND
69	2.8	32.0	<0.04	124.0	<0.01	203.5	13.8	7.25	7.22
11	8.8	15.6	0.05	93.0	0.07	225.0	13.6	3.59	6.38
364	74.0	83.5	0.40	420.5	3.80	849.0	15.3	0.21	7.12
79	34.5	103.0	0.14	503.5	0.57	790.0	14.8	0.23	7.00

Parameter abbreviations follow standard chemical conventions except for: EC – Electrical conductivity; DO – dissolved oxygen; ND – indicates that the parameter value was not determined. Units are in mg L⁻¹ for all variables except pH which is in pH units, EC which is in $\mu\text{S cm}^{-1}$ at 25 °C, and Temperature which is in °C.

Groundwater sampling and DNA extraction

The groundwater bacterial genomic DNA that was utilized for the studies in chapters 3.1 and 3.3 was also used in this study. Detailed descriptions of groundwater sampling, filtration and DNA extractions can be found in Sirisena et al. (2013). Briefly, a 2 L-groundwater sample was collected from each site in June 2010, according to the National Protocol for State of the Environment Groundwater Sampling (Daughney et al. 2006). Water filtration was conducted using 0.22 µm nitrocellulose membrane filters (Millipore, Australia). The DNA extractions were performed using ZR Fungal/Bacterial DNA kits (Zymo Research, United States) as per the manufacturer's instructions. The extracted DNA was quantified using Quant-iT™ High-Sensitivity DNA Assay kits (Life Technologies, CA, USA).

Illumina sequencing of metagenomes

High throughput metagenome sequencing was conducted on the Illumina MiSeq™ platform at the Massey Genome Service, Palmerston North, New Zealand. One of the technical limitations in sample selection was that the majority of samples did not have adequate amounts of genomic DNA for the sequencing. Therefore, the whole genome amplification was performed by multiple displacement amplification (MDA) (Dean et al. 2002) for the six samples using REPLI-g Mini Kits (Qiagen, Germany) as per the manufacturer's instructions. The amplified genomic DNA fragments were labelled with short DNA barcodes that were unique to each metagenome. The samples were prepared for the *de novo* sequencing using Illumina TruSeq DNA Library Preparation kits with 1 µg of genomic DNA obtained from whole genome amplification for each sample according to the manufacturer's instructions. The DNA quantifications were performed using Quant-iT™ ds DNA HS assay kits and the

contamination checks were conducted with Quant-iT™ RNA assay and Quant-iT™ Protein assay kits on Quibit® fluorometer (Life Technologies, CA, USA).

Metagenome sequence analysis strategy

The bioinformatics analyses were also conducted at Massey Genome Service, Palmerston North, New Zealand. The resulting 2×250 base paired end reads were separated according to the DNA barcode assigned for each sample followed by the quality check analysis and pre-processing using three software packages: SolexaQA (Cox et al. 2010); FastQC version 0.10.1 (Andrews 2012); and FastQ Screen version 0.4.1 (Andrews 2013) with the standard parameters.

As the previous studies suggested that *Pseudomonas spp.* dominated the microbial compositions in the samples, the short reads were mapped against randomly selected *Pseudomonas* genomes (Table 4) using the Burrows-Wheeler Aligner (BWA) software package (Li & Durbin 2009).

Table 4 Summary of the reference *Pseudomonas* genomes used in short read mapping.

GenBank ID	Genome size (b)	Species
NC_004129	7074893	<i>Pseudomonas protegens</i> Pf-5
NC_005773	5928787	<i>Pseudomonas syringae</i> pv. <i>phaseolicola</i> 1448A
NC_008027	5888780	<i>Pseudomonas entomophila</i> L48
NC_009439	5072807	<i>Pseudomonas mendocina</i> ymp
NC_010322	6078430	<i>Pseudomonas putida</i> GB-1
NC_012660	6722539	<i>Pseudomonas fluorescens</i> SBW25
NC_015379	6843248	<i>Pseudomonas brassicacearum</i> subsp. <i>brassicacearum</i> NFM421
NC_015556	4920769	<i>Pseudomonas fulva</i> 12-X

An alternative mapping approach was used to map the sequences into genomes of an array of species. Here, the sequences from the six metagenomes were mapped into approximately 2300 complete bacterial genomes in GenBank with an ultra-fast BLAST like analysis using the PAUDA (Protein Alignment Using a DNA Aligner) programme (Huson & Xie 2014). The BLAST output from PAUDA was analysed using the MEtaGenome ANalyzer (MEGAN) programme (Huson et al 2011; Huson & Mitra 2012) in order to obtain a comparative overview of the relative abundances of the bacterial species in each metagenome.

The *de novo* sequence assembly of six genomes was performed using four different assemblers: Velvet (Zerbino & Birney 2008; Zerbino et al. 2009); IDBA-UD (Peng et al. 2012); SOAPdenovo (Li et al. 2010); and Edena (Hernandez et al. 2008). Considering the heterogeneity of samples and the way each assembler works, six different assembly modes were used in this analysis. Here, the sequences were analysed as paired end reads because it allows the use of overlapping regions to construct large contigs: relatively long DNA sequences generated by assembling several overlapping short sequences. Hereafter, only contigs longer than 20 kb were considered for further analyses.

The genes in the contigs of six genomes were then predicted using the Glimmer (Gene Locator and Interpolated Markov ModelER) version 3.02 prokaryotic gene prediction tool (Salzberg et al. 1998; Delcher et al. 1999; Delcher et al. 2007). There are two prediction types: 1) normal – only the whole genes are predicted; and 2) extended – in addition to the whole genes, the open reading frames that run off at the end of contig are also counted. The Glimmer gene predictions (GGPs) resulted in the number of genes, the predicted length of the gene and the amino acid sequence.

For the assessment of horizontal gene transfer, GGPs were subjected to Standard Protein BLAST (BLASTp) analysis (Altschul et al. 1997). Here, approximately 7.3 million

amino acid sequences were downloaded from GenBank to use as the database for BLAST analysis. In BLASTp, nucleotide sequences of the contigs were converted into amino acid sequences and aligned with the sequences in the protein database. This is a more sensitive method to find genes at greater taxonomic distances because a single mismatch in the nucleotide sequence will have higher impact on protein-protein alignment than nucleotide-nucleotide comparison. Although a large number of BLAST hits were received for each contig, only the top five matches were considered for further analyses. Finally, the contigs were analysed by mapping the short reads obtained from the isolate back to the contig itself by using the Geneious version 5.4 sequence analysis software (Drummond et al. 2010). If the mapping coverage was not completed, it reflects the presence of any chimeric sequences on the contig which may have been acquired via horizontal gene transfer. Further bioinformatics analyses are underway and a brief summary of the results are described in this chapter.

RESULTS

High throughput sequencing of six metagenomes provided a total of 22,362,881 reads. Each sample equally contributed (between 14 – 18%) to the total reads (Table 5).

Table 5 Summary of the number of reads resulted from each metagenome.

Sample Name	Number of reads	% of total
GGW ID 11	3,931,655	17.58%
GGW ID 17	4,133,682	18.48%
GGW ID 24	3,700,797	16.55%
GGW ID 364	3,214,663	14.37%
GGW ID 69	3,283,354	14.68%
GGW ID 79	3,703,915	16.56%
Undetermined*	394,815	1.77%
Total	22,362,881	

* Undetermined reads were not considered for further analyses

For the six samples, the T-RFLP profiles obtained as described in chapter 3.1 suggested that there could be the same single dominant species in the bacterial communities (Fig. 2).

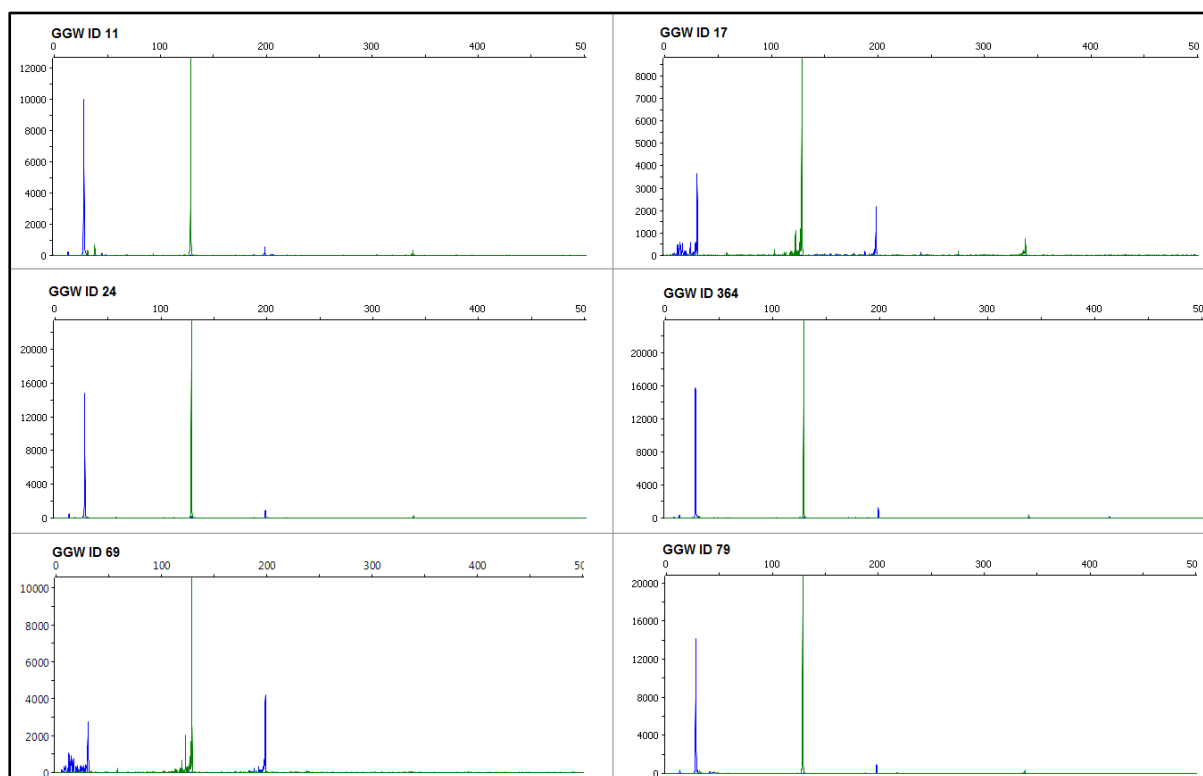


Figure 2 The T-RFLP profiles for the six selected samples. The experimental details are explained in chapter 3.1.

Roche 454 pyrosequencing of three samples confirmed that *Pseudomonas* was the dominant species in these microbial communities (Fig. 3). It was not possible to display the pyrosequencing profiles for all six samples as the sample selection for this study was limited by the amount of available bacterial genomic DNA for the whole genome sequencing.

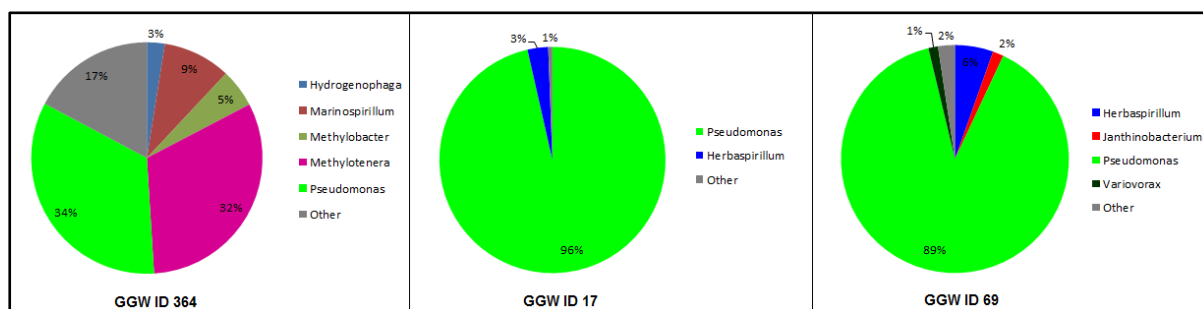


Figure 3 Bacterial community compositions of three samples obtained by Roche 454 pyrosequencing approach. The experimental details are explained in chapter 3.3.

However, the mapping of short reads into selected *Pseudomonas* genomes using BWA indicated that the six metagenomes are more heterogeneous than expected, as a relatively small percentage of reads from each genome was mapped into the eight reference genomes (Table 6). The highest mapping percentage was only 0.0388%, which is a relatively low value. It was noted that it is important to have a high degree of homology between a short read and the reference sequence for the mapping to be recorded in this type of software.

Table 6 Summary of the short read mapping into reference *Pseudomonas* genomes. Number of reads and the percentage of reads mapped are indicated.

GenBank ID	GGW ID 11		GGW ID 17		GGW ID 24		GGW ID 364		GGW ID 69		GGW ID 79	
	# of reads	%	# of reads	%	# of reads	%	# of reads	%	# of reads	%	# of reads	%
NC_004129	4093	0.0010	41543	0.0100	19219	0.0052	4417	0.0014	15487	0.0047	24085	0.0065
NC_005773	2763	0.0007	26725	0.0065	13061	0.0035	3965	0.0012	10795	0.0033	20155	0.0054
NC_008027	2293	0.0006	21607	0.0052	9785	0.0026	3245	0.0010	8437	0.0026	15433	0.0042
NC_009439	1995	0.0005	16811	0.0041	7777	0.0021	2965	0.0009	6755	0.0021	13953	0.0038
NC_010322	3125	0.0008	22433	0.0054	10353	0.0028	3519	0.0011	8781	0.0027	24489	0.0066
NC_012660	17155	0.0044	53351	0.0129	143675	0.0388	6991	0.0022	34519	0.0105	52313	0.0141
NC_015379	4161	0.0011	48871	0.0118	27097	0.0073	4397	0.0014	19327	0.0059	25343	0.0068
NC_015556	2107	0.0005	17329	0.0042	10007	0.0027	3355	0.0010	7245	0.0022	15203	0.0041

The analysis conducted using the PAUDA and MEGAN programmes revealed that the whole genome sequencing approach was more or less consistent with the Roche 454 pyrosequencing and T-RFLP molecular techniques. Of the six metagenomes, *Pseudomonas* appeared to be the sole dominant genus in four samples (Fig. 4).



Figure 4 Word-cloud summary of taxonomic identities in the six metagenomes using the PAUDA: A) based on absolute read counts mapped; and B) based on normalized read counts mapped on the square root scale. The sizes of the letters are proportional to the relative abundance of the genus in the sample.

However, the other two samples had different dominant genera: *Gallionella* in GW ID 11 metagenome and *Methylobacter* in GW ID 364, although *Pseudomonas* still remained as a key component.

The *de novo* assembly of short sequences using two different assemblers with various combinations of modes indicated that, in some cases, the estimated metagenome sizes are at least an order of magnitude larger than the expected dominant genome sizes i.e.: ~ 6 Mb for the genus *Pseudomonas*, reflecting the complexity of the samples (Table 7).

Table 7 Summary of the maximum genome sizes for each sample.

Sample	Assembler	Total Number of bases [†]	Bases on contigs over 20kb*
GW ID 11	Velvet	102037772	26356214
GW ID 17	Velvet	91667183	51986560
GW ID 24	Edena	67215005	12608139
GW ID 364	Edena	98794223	29177983
GW ID 69	Velvet	66978817	12891052
GW ID 79	Edena	129238829	26540677

[†] the genome sizes determined by all the contigs

* the genome sizes determined by the contigs larger than 20kb

As the contig sizes varied significantly, only contigs over 20kb were considered for further analyses. In addition and for a given sample, if a sequence was found under a set of assembly conditions, and an exactly identical sequence, its reverse complement or exact subsequence of that contig was found with another set of assembly conditions for that same sample, only the longest sequence was considered for further analyses.

The basic gene prediction with the Glimmer programme using the contigs over 20kb indicated the number of genes in each metagenome (Table 8). The sample GW ID 69 had the lowest number of Glimmer gene predictions (GGPs) for both prediction types: normal – 469 whole genes in all contigs over 20kb and extended – 472 whole genes and open reading

frames. This could be due to the low number of contigs larger than 20 kb obtained for the sample.

Table 8 Summary of the Glimmer gene predictions.

Sample	Prediction type	Number of genes
GGW ID 11	Normal	17512
GGW ID 11	Extended	17899
GGW ID 17	Normal	8137
GGW ID 17	Extended	8275
GGW ID 24	Normal	2544
GGW ID 24	Extended	2611
GGW ID 364	Normal	73577
GGW ID 364	Extended	74777
GGW ID 69	Normal	469
GGW ID 69	Extended	472
GGW ID 79	Normal	6705
GGW ID 79	Extended	6857

Table 9 provides a summary of the number of BLAST hits for each sample. All isolates produced more or less a similar number of hits except GGW ID 69 which generated only 5103 matches. Again, the number of contigs larger than 20kb might have influenced these numbers.

Table 9 Summary of the BLAST hits for each genome.

Genome	Prediction type	Number of BLAST hits
GGW ID 11	Normal	28528
GGW ID 17	Normal	21956
GGW ID 24	Normal	22861
GGW ID 364	Normal	28099
GGW ID 69	Normal	5103
GGW ID 79	Normal	26742

It is important to note that the number of BLAST hits is greater than the number of genes in the metagenome because different datasets were taken into consideration for the two analyses. In general, the BLAST hits were partially in accord with the PAUDA analysis. As expected, the top hits for isolates GGW ID 17, GGW ID 24, GGW ID 69 and GGW ID 79 were members of *Pseudomonas* genus. However, *Sulfuricurvum kujiense* DSM 16994 was the top match for GGW ID 11 instead of *Gallionella*, whereas *Methylobacterium alcaliphilum* 20Z members gave more hits for GGW ID 364 instead of *Methylobacterium*.

A selected set of contigs for each genome was separately analysed by mapping the short reads back to the contig itself. As a demonstration, a preliminary attempt for contig 1026 (56680 bp with 76 GGP) from the sample GGW ID 11 was used (Fig. 5).

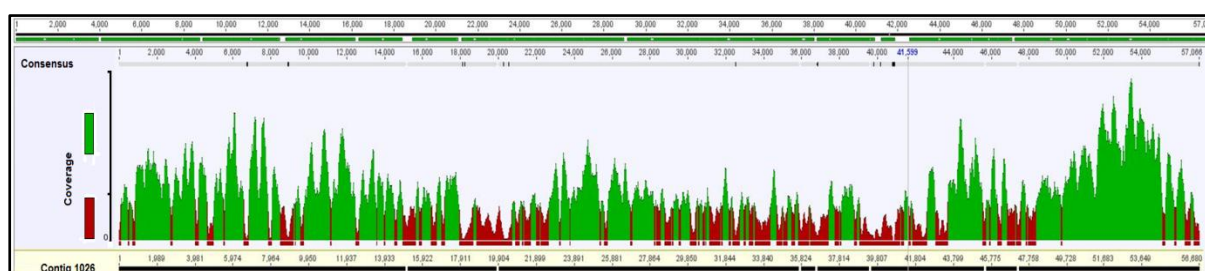


Figure 5 Illustration of the contig coverage for contig 1026 from the sample GGW ID 11. Green areas represent the nucleotide coverage of 20 or over whereas red areas indicate coverage less than 20.

The results indicated that 8453 reads were mapped to this contig with average nucleotide coverage between 22.4 and 37.29 bases. The comparison of these results to the BLAST results suggested that the majority of the contig maps to *Sulfuricurvum kujiense* DSM 16994, but there are some parts to which *Sulfuricurvum kujiense* DSM 16994 is not the perfect match (Figure 6).

11_S3	1026	861	orf00053	15	12423	362	gi 313682928 ref YP_004060666.1	Sulfuricurvum kujiense DSM 16994
11_S3	1026	860	orf00054	15	12408	291	gi 313682938 ref YP_004060676.1	Sulfuricurvum kujiense DSM 16994
11_S3	1026	859	orf00055	11	12397	162	gi 313682937 ref YP_004060675.1	Sulfuricurvum kujiense DSM 16994
11_S3	1026	858	orf00056	15	12382	266	gi 313682935 ref YP_004060673.1	Sulfuricurvum kujiense DSM 16994
11_S3	1026	857	orf00057	15	12367	144	gi 339501264 ref YP_004699299.1	Spirochaeta caldaria DSM 7334
11_S3	1026	848	orf00058	15	12266	90.1	gi 339501263 ref YP_004699298.1	Spirochaeta caldaria DSM 7334
11_S3	1026	849	orf00060	15	12281	227	gi 392979364 ref YP_006477952.1	Enterobacter cloacae subsp. dissolvens SDM
11_S3	1026	850	orf00062	15	12296	102	gi 428219213 ref YP_007103678.1	Pseudanabaena sp. PCC 7367
11_S3	1026	851	orf00063	15	12311	229	gi 384172491 ref YP_005553868.1	Arcobacter sp. L
11_S3	1026	852	orf00064	15	12326	238	gi 390939197 ref YP_006402934.1	Sulfurospirillum barnesii SES-3
11_S3	1026	853	orf00066	15	12341	510	gi 313682934 ref YP_004060672.1	Sulfuricurvum kujiense DSM 16994
11_S3	1026	854	orf00067	11	12356	275	gi 313682933 ref YP_004060671.1	Sulfuricurvum kujiense DSM 16994
11_S3	1026	885	orf00071	15	12783	264	gi 262393973 ref YP_003285827.1	Vibrio sp. Ex25
11_S3	1026	884	orf00074	15	12768	544	gi 296273710 ref YP_003656341.1	Arcobacter nitrofigilis DSM 7299
11_S3	1026	883	orf00076	15	12753	193	gi 390939948 ref YP_006403685.1	Sulfurospirillum barnesii SES-3
11_S3	1026	882	orf00077	15	12738	153	gi 222055152 ref YP_002537514.1	Geobacter daltonii FRC-32
11_S3	1026	881	orf00078	15	12723	424	gi 313682928 ref YP_004060666.1	Sulfuricurvum kujiense DSM 16994
11_S3	1026	880	orf00079	15	12708	117	gi 253700391 ref YP_003021580.1	Geobacter sp. M21
11_S3	1026	879	orf00081	15	12693	339	gi 313682923 ref YP_004060661.1	Sulfuricurvum kujiense DSM 16994
11_S3	1026	878	orf00082	29	12664	357	gi 71909600 ref YP_287187.1	Dechloromonas aromatica RCB
11_S3	1026	877	orf00083	15	12650	275	gi 78778052 ref YP_394367.1	Sulfurimonas denitrificans DSM 1251
11_S3	1026	876	orf00084	15	12634	281	gi 313683120 ref YP_004060858.1	Sulfuricurvum kujiense DSM 16994
11_S3	1026	875	orf00086	15	12619	327	gi 313682920 ref YP_004060658.1	Sulfuricurvum kujiense DSM 16994

Figure 6A The BLAST results for the selected section of the contig 1026 from the sample GGW ID 11.

11_S3	1034	1315	orf00052	15	18590	570	gi 313681358 ref YP_004059096.1	Sulfuricurvum kujiense DSM 16994
11_S3	1034	1314	orf00054	15	18575	423	gi 313681357 ref YP_004059095.1	Sulfuricurvum kujiense DSM 16994
11_S3	1034	1313	orf00055	15	18560	244	gi 313681356 ref YP_004059094.1	Sulfuricurvum kujiense DSM 16994
11_S3	1034	1312	orf00056	15	18545	88.2	gi 78778067 ref YP_394382.1	Sulfurimonas denitrificans DSM 1251
11_S3	1034	1311	orf00057	15	18530	582	gi 313681354 ref YP_004059092.1	Sulfuricurvum kujiense DSM 16994
11_S3	1034	1310	orf00059	15	18515	912	gi 313681353 ref YP_004059091.1	Sulfuricurvum kujiense DSM 16994
11_S3	1034	1309	orf00060	15	18500	496	gi 313681352 ref YP_004059090.1	Sulfuricurvum kujiense DSM 16994
11_S3	1034	1308	orf00061	15	18485	602	gi 313681351 ref YP_004059089.1	Sulfuricurvum kujiense DSM 16994

Figure 6B The BLAST results for the selected section of the contig 1034 from the sample GGW ID 11.

Figure 6A indicates that the majority of the selected contig region is matched to the genus *Sulfuricurvum* although there are some sections that are more closely related to other genera such as *Spirochaeta*, *Enterobacter*, *Arcobacter*, *Vibrio* and *Geobacter*. Similarly, Figure 6B shows that there are some contig regions that are mapped to one genus (*Sulfuricurvum*), but there are some differences at species level. Based on these approaches, further analyses will be conducted to examine other contigs from different genomes especially that are related to *Pseudomonas* to assess the possibility of horizontal gene transfer between bacterial species.

DISCUSSION

The Illumina high throughput sequencing coupled with whole genome amplification was successfully employed to analyse the metagenome compositions of six groundwater samples that can be used as a resource for further explorations of HGT. In general, the whole genome sequencing results are in accord with the previous Roche 454 sequencing data and are comparable to the T-RFLP based bacterial community structure. In GGW ID 364, *Methylothera* was also identified as a dominant genus in addition to *Pseudomonas* with both 454 and Illumina sequencing approaches. However, it is interesting to note that the T-RFLP profile for GGW ID 364 indicates only *Pseudomonas* as the dominant species. This could be due to the terminal restriction fragments with the same lengths for both *Pseudomonas* and *Methylothera* or the PCR primers that were used in T-RFLP may not efficiently recognize *Methylothera*.

Further, *Methylothera*, a new genus within the family *Methylophilaceae*, is an obligate methylotroph, i.e. it is capable of degrading single carbon organic compounds (containing no carbon–carbon bonds), especially methanol and methylamine (Kalyuzhnaya et al. 2006; Lidstrom 2006). Therefore, the degradation of some of these compounds will produce ammonia as a by-product (Bosch et al. 2009). Interestingly, hydrochemical data (Table 3) indicates that GGW ID 364 contained a relatively high ammonia concentration. In GGW ID 11, *Gallionella* appeared to be the dominant species. It is a chemolithotrophic iron oxidizing bacteria that gains its energy by oxidizing dissolved iron. Therefore, this species is normally found in environments that contain dissolved iron concentrations significantly above the detection limit (Wang et al. 2009). Interestingly, the groundwater chemistry data suggests that GGW ID 11 contained a relatively higher iron concentration. Further, *Gallionella* needs microaerophilic conditions, i.e. where there is a small amount of dissolved

oxygen available, in order to oxidize the iron (Anderson & Pedersen 2003). The hydrochemistry of GGW ID 11 partially supports this requirement as the site contains a moderate amount of dissolved oxygen. Overall, these results provide an additional verification to the main conclusions drawn in the previous chapters: the groundwater bacterial diversity is mainly related to the hydrochemistry.

The low percentage values for the short read mapping into the reference *Pseudomonas* genomes (Table 6) indicate that the samples are more complex than expected. This could indicate that there were low abundant species that were not recognized especially with the T-FRLP technique. *De novo* assembly also suggests that estimated metagenome sizes were larger than the expected sizes. This might also be due to the complex nature of these metagenomic samples as the contigs may represent a mixture of species. However, further analysis should be conducted to determine whether this size difference is purely due to the samples being mixtures of species or an indication of the HGT between bacterial species.

Although BLAST results suggested that contig 1026 was mainly related to *Sulfuricurvum kujiense* DSM 16994 (Fig. 6A), the mapping of short reads to the contig was not completed (Fig. 5). In addition, the BLAST result itself also indicated that some of the GGP's of that contig were more closely related to genera such as *Arcobacter*, *Vibrio* and *Geobacter* but not to *Sulfuricurvum*. This might be a sign of the presence of foreign DNA in the genome of *Sulfuricurvum*. However, further studies are required confirm such a relationship.

In these initial analyses, only the contigs over 20kb were considered, to reduce the complexity of the analysis. However, the number of genes and number of BLAST hits for each metagenome is sensitive to this randomly selected 20kb threshold. Therefore, the selection of this threshold should be re-evaluated in the future analyses.

As the analysis of the entire data set generated in this study was beyond the scope of my PhD project, further bioinformatics analyses will be conducted after the submission of my PhD thesis to gain a better understanding of the genome structure of these samples and to postulate the underlying biological process that shapes the genetic composition. The current results and approaches discussed in the chapter will provide a solid foundation in the future analyses of exploring the events of HGT in New Zealand groundwater bacteria.

REFERENCES

- Achtman M, Wagner M 2008. Microbial diversity and the genetic nature of microbial species. *Nature Reviews: Microbiology* 6: 431–440.
- Anderson CR, Pedersen K 2003. In situ growth of *Gallionella* biofilms and partitioning of lanthanides and actinides between biological material and ferric oxyhydroxides. *Geobiology* 1: 169–178.
- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25: 3389–3402.
- Andrews S 2012. FastQC version 0.10.1, Cambridgeshire, United Kingdom, The Babraham Bioinformatics. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- Andrews S 2013. FastQ Screen version 0.4.1, Cambridgeshire, United Kingdom, The Babraham Bioinformatics. http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/
- Bosch G, Wang T, Latypova E, Kalyuzhnaya MG, Hackett M, Chistoserdova L 2009. Insights into the physiology of *Methylobacterium mobilis* as revealed by metagenome-based shotgun proteomic analysis. *Microbiology* 155: 1103–1110.
- Burke C, Steinberg P, Rusch D, Kjelleberg S, Thomas T 2011. Bacterial community assembly based on functional genes rather than species. *Proceedings of the National Academy of Sciences of the United States of America* 108: 14288–14293.
- Cohan FM 2002. What are bacterial species? *Annual Review of Microbiology* 56: 457–87.
- Cox MP, Peterson DA, Biggs PJ 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* 11: 485.

- Daughney CJ, Reeves RR 2005. Definition of hydrochemical facies in the New Zealand groundwater monitoring programme. *Journal of Hydrology (New Zealand)* 44: 105–130.
- Daughney CJ, Jones A, Baker T, Hanson C, Davidson P, Zemansky GM, Reeves RR, Thompson M 2006. A national protocol for state of the environment groundwater sampling in New Zealand: ME 781. Wellington, New Zealand, Ministry for the Environment.
- Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, Lasken RS 2002. Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences*, 99: 5261–5266.
- De la Cruz F, Davies J 2000. Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends in Microbiology* 8: 128–133.
- Delcher AL, Harmon D, Kasif S, White O, Salzberg SL 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Research* 27: 4636–4641.
- Delcher, AL, Bratke KA, Powers EC, Salzberg SL 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 23: 673–679.
- Drummond AJ, Ashton B, Buxton SCM, Cooper A, Heled J et al. 2010. Geneious version 5.4. Available from <http://www.geneious.com/>
- Eisen JA 2000. Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Current Opinion in Genetics & Development* 10: 606–611.
- Gogarten JP, Townsend JP 2005. Horizontal gene transfer, genome innovation and evolution. *Nature Reviews: Microbiology* 3: 679–687.
- Hernandez D, François P, Farinelli L, Østerås M, Schrenzel J 2008. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Research* 18: 802–809.

- Huson DH, Mitra S, Weber N, Ruscheweyh H, Schuster SC 2011. Integrative analysis of environmental sequences using MEGAN4. *Genome Research* 21:1552–1560.
- Huson DH, Mitra S 2012. Introduction to the analysis of environmental sequences: metagenomics with MEGAN. *Methods in Molecular Biology* 856: 415–429.
- Huson DH, Xie C 2014. A poor man’s BLASTX—high-throughput metagenomic protein database search using PAUDA. *Bioinformatics* 30: 38–39.
- Jain R, Rivera MC, Lake JA 1999. Horizontal gene transfer among genomes: the complexity hypothesis. *Proceedings of the National Academy of Sciences of the United States of America* 96: 3801–3806.
- Kalyuzhnaya MG, Bowerman S, Lara JC, Lidstrom ME, Chistoserdova L 2006. *Methylotenera mobilis* gen. nov., sp. nov., an obligately methylamine-utilizing bacterium within the family Methylophilaceae. *International Journal of Systematic and Evolutionary Microbiology* 56: 2819–2823.
- Konstantinidis KT, Ramette A, Tiedje JM 2006. The bacterial species definition in the genomic era. *Philosophical Transactions of the Royal Society of London: Series B, Biological Sciences* 361: 1929–1940.
- Li H, Durbin R 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* 25:1754–1760.
- Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Wang J 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, 20: 265–272.
- Lidstrom ME 2006. Aerobic methylotrophic prokaryotes. *The Prokaryotes*. Edited by Balows A, Truper HG, Dworkin M, Harder W, Schleifer KH. New York, United States, Springer.

- Luo C, Walk ST, Gordon DM, Feldgarden M, Tiedje JM, Konstantinidis KT 2011. Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *Proceedings of the National Academy of Sciences of the United States of America* 108: 7200–7205.
- McDaniel LD, Young E, Delaney J, Ruhnau F, Ritchie KB, Paul JH 2010. High frequency of horizontal gene transfer in the oceans. *Science* 330: 50.
- Medini D, Serruto D, Parkhill J, Relman DA, Donati C, Moxon R, Falkow S, et al. 2008. Microbiology in the post-genomic era. *Nature Reviews: Microbiology* 6: 419–430.
- Pál C, Papp B, Lercher MJ 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature Genetics*, 37: 1372–1375.
- Park C, Zhang J 2012. High expression hampers horizontal gene transfer. *Genome Biology and Evolution* 4: 523–532.
- Peng Y, Leung HC, Yiu SM, Chin FY 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28: 1420–1428.
- Richter M, Rosselló-Móra R 2009. Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences of the United States of America* 106: 19126–19131.
- Salzberg SL, Delcher AL, Kasif S, White O 1998. Microbial gene identification using interpolated Markov models. *Nucleic Acids Research* 26: 544–548.
- Sirisena KA, Daughney CJ, Moreau-Fournier M, Ryan KG, Chambers GK 2013. National survey of molecular bacterial diversity of New Zealand groundwater: relationships between biodiversity, groundwater chemistry and aquifer characteristics. *FEMS Microbiology Ecology* 86: 490–504.

- Sirisena KA, Daughney CJ, Moreau M, Ryan KG, Chambers GK 2014. Relationships between Molecular Bacterial Diversity and Chemistry of Groundwater in the Wairarapa Valley, New Zealand. *New Zealand Journal of Marine and Freshwater Research* (Accepted).
- Stackebrandt E, Frederiksen W, Garrity GM, Grimont AD, Kämpfer P, Maiden, MCJ, Nesme X et al. 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *International Journal of Systematic and Evolutionary Microbiology* 52: 1043–1047.
- Tindall BJ, Rosselló-Móra R, Busse H-J, Ludwig W, Kämpfer P 2010. Notes on the characterization of prokaryote strains for taxonomic purposes. *International Journal of Systematic and Evolutionary Microbiology* 60: 249–266.
- Thomas CM, Nielsen KM 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature Reviews: Microbiology* 3: 711–721.
- Wang J, Muyzer G, Bodelier PL, Laanbroek HJ 2009. Diversity of iron oxidizers in wetland soils revealed by novel 16S rRNA primers targeting *Gallionella*-related bacteria. *The ISME Journal* 3: 715-725.
- Zerbino DR, Birney E 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* 18: 821–829.
- Zerbino DR, McEwen GK, Margulies EH, Birney E 2009. Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler. *PloS one* 4: e8407.

GENERAL DISCUSSION AND CONCLUSION

This study has focused on characterizing the bacterial diversity of New Zealand groundwater at national and local scales using molecular methods and exploring the underlying environmental and geological factors that shape the microbial community composition. Prior to this project, very little information was available on New Zealand groundwater microbiology. Therefore, my work began with an exploratory approach and gradually developed to a hypothesis testing level.

I have used an array of molecular techniques that have unique advantages and drawbacks. In the initial phase, the terminal restriction fragment length polymorphism (T-RFLP) technique was used to determine the bacterial community structure at national and local scales. In the latter stages of the project, two next-generation high throughput sequencing approaches: 1) Roche 454 sequencing; and 2) Illumina MiSeq platform, were also used to obtain a deeper insight into New Zealand groundwater microbial ecology. One of the most interesting outcomes of the project was that all the molecular approaches employed in this work provided more or less comparable results to each other.

In Chapter 3.1, I have investigated the bacterial community structure in New Zealand groundwater at a national scale using the T-RFLP molecular profiling tool. This study provided the first overview of the microbial diversity in New Zealand groundwater ecosystems (Sirisena et al. 2013). This work has been recognized as one of the first studies in the world to evaluate the bacterial diversity across an entire country at large scale (Gregory et al. 2014; Griebler et al. 2014). Overall, the results suggested that the bacterial community structure is mainly related to the groundwater chemistry, whereas the environmental and geological factors such as geographical region, aquifer lithology, land use activities in aquifer recharge zone, well depth and mean residence time (MRT) of groundwater might play a

secondary role in determining the microbial composition [**refer to the section 1.5.1 in the General Introduction**]. To the best of my knowledge, this is the first time the factors that influence groundwater bacterial diversity in New Zealand have been revealed.

On the basis of results presented in Chapter 3.1, I suggest that microbial assessment criteria should be incorporated into the State-of-the-environment (SOE) monitoring programmes regularly conducted in New Zealand. The National Groundwater Monitoring Programme (NGMP) conducted by GNS Science in collaboration with 15 regional authorities has provided a very useful framework for sample collection and retrieving hydrochemical and environmental data regarding sampling sites, which has aided the identification of relationships between the biotic and abiotic factors in the groundwater. Some samples indicated the presence of one, or just a few, dominant species by providing simple T-RFLP profiles with a unique combination of terminal restriction fragment peaks (one peak for each fluorescent dye). Generally, mixtures of bacterial species in varying relative abundances normally occur in environmental samples including groundwater. However, it was not clear from these data whether this special phenomenon was a real biological signal or was due to one of the common drawbacks of the T-RFLP technique: the possibility of providing similar peaks for different species.

The local scale analysis (Chapter 3.2) also revealed that groundwater bacterial diversity is mainly related to the hydrochemistry, in particular the redox potential, where the environmental factors such as aquifer confinement and groundwater bore usage were not influential parameters (Sirisena et al. 2014). This conclusion was supported by the observation that the bacterial community structure for some selected sites remained unchanged where the groundwater chemistry was stable over time [**refer to the section 1.5.2 in the General Introduction**]. However, a comprehensive evaluation of the temporal changes in bacterial communities was beyond the scope of this project. Previous studies have

suggested that the seasonal changes might also influence the microbial composition in groundwater environments (Lin et al. 2012). This could be due to a shift in the hydrochemistry that is coupled with seasonal change or changes in the water table. Therefore, in the future, it would be important to investigate the temporal changes in the microbial community structure of New Zealand groundwater and explore the primary causes that actually influence any potential shifts.

The evaluation of the effectiveness of the T-RFLP technique suggested that it is a robust and reliable tool that can be effectively used especially in rapid exploratory studies, because the results appeared to be independent of the data analysis approach and the experimental procedures. Therefore, the T-RFLP methodology is an ideal choice for regular groundwater monitoring practices conducted in the NGMP and in the SOE programmes operated by regional authorities.

The Roche 454 high throughput pyrosequencing approach provided a deeper insight into the groundwater bacterial diversity and its relationships with abiotic factors [**refer to the section 1.5.3 in the General Introduction**]. In addition, the technology indirectly helped to validate the effectiveness of the T-RFLP methodology as a reliable and robust community profiling tool for groundwater microbiology studies. The 454 sequencing results revealed the taxonomic identities of the bacterial taxa present in groundwater samples, which was not available with the T-RFLP methodology. Therefore, more reliable relationships between microbial community composition and the environmental factors were revealed. The comparison of the results from the two molecular methods indicated that the 454 sequencing approach identified far more bacterial taxa, most of which were represented in low abundances. Interestingly, the groundwater bacterial community composition was still mainly related to the redox potential of the water despite the presence of the high number of low abundance taxa revealed with the 454 sequencing approach. This observation suggests

that highly abundant bacterial species that were reliably detected with the both molecular methods provide a higher contribution to the bacterial diversity-hydrochemistry relationship than the low abundance taxa that were precisely identified with the 454 approach, but not so precisely with the T-RFLP method.

Further, the pyrosequencing methodology confirmed that the presence of single dominant bacterial species in some samples, which was indicated with the T-RFLP approach, is a valid biological phenomenon in New Zealand groundwater ecosystems. Most of the samples contain *Pseudomonas* as the sole dominant taxa despite diverse chemistries and environmental conditions. Therefore, this finding led to development of the hypothesis that *Pseudomonas spp.* may have acquired genetic material from other species present in these environments via the process of horizontal gene transfer (HGT) to sustain under diverse abiotic conditions, especially under various hydrochemistries.

The Illumina MiSeq high throughput sequencing technology was successfully employed to develop six metagenomic databases to explore possible HGT events in those dominant species **[refer to the section 1.5.4 in the General Introduction]**. In addition, this approach provided cross-validations for both T-RFLP and Roche 454 sequencing methodologies used in this project. Overall, the Illumina whole genome sequencing results are approximately in accord with the findings from the other two molecular methods used in this thesis, and further supported the view that the groundwater bacterial diversity is related to hydrochemistry even when there are some instances where taxonomically defined single species, i. e. *Pseudomonas spp.*, is present as the sole dominant taxon across a range of diverse hydrochemistries. Here, I suggest that this species could also follow the general bacterial diversity-hydrochemistry relationship by acting as multiple species if the genetic information is gained from other species in the ecosystem. This hypothesis supports the idea that the universal properties of an ideal bacterial species may not be reflected by

taxonomically named species, but by the ecotypes that are occupying the same ecological niche (Cohan 2002). As the complete analysis of the enormous volume of data generated in the Illumina high throughput whole genome sequencing study was beyond the scope of my PhD project, further bioinformatics analyses will be conducted after the submission of my thesis to explore this hypothesis.

The central theme of the entire project was the application of molecular methods in exploring the microbial diversity of New Zealand groundwater and evaluating the underlying factors that influence groundwater microbiota. Even in this genomics era, the traditional microbiological methods such as culturing techniques could play a crucial role in understanding the microbial diversity in subsurface environments. The extraction of genomic DNA from all the bacterial species is an important factor to the success of almost all the molecular approaches that can be employed in environmental microbiology. But this is not always guaranteed as some species might have rigid cell walls that prevent extracting DNA from those cells. In addition, PCR has become an integral step in most of the molecular approaches. However, the universal primer sequences used in PCR amplifications are universal only to the species that are already identified and stored in microbial databases. Due to these reasons, some bacterial taxa actually present in the environment may not be identified. Interestingly, some studies have revealed that both molecular and culturing methods may overlook certain bacterial species while identifying taxa that the other method could not detect (Donachie et al. 2007). Therefore, as another future direction, I suggest that bacterial community structure in some of these groundwater monitoring sites should be reanalysed using both molecular approaches and culturing techniques to evaluate the missing species with each approach.

This project was limited to the analysis of the groundwater itself and not extended to aquifer sediment materials. However, previous studies have suggested that the microbial

community composition of the liquid groundwater can be different from that of the aquifer itself, and that the latter may be even more closely related to hydrochemistry (Alfreider et al. 1997; Griebler et al. 2002; Flynn et al. 2008; Griebler & Lueders 2009). My results indicate that microbial diversity of groundwater itself is also related to its chemical state. Therefore, in the future, the bacterial community compositions in liquid groundwater and its aquifer components should be compared at least for some of these sites tested in this project. That will answer some of the key questions related to New Zealand groundwater microbial ecology, e.g. how similar are these two microbial communities, is there a shared community between the two environments, and if so are they both related to the hydrochemistry?

Considering the relative scales of sample collection and the broad objectives of my PhD project, the replicates were performed based on hydrochemistries rather the location. In other words, while single samples were collected from diverse geographical regions, replicates existed for each particular hydrochemistry. However, several samples from each location should be analysed to determine the fraction of bias introduced by the sampling strategy (Prosser 2010). Therefore, this would also be an interesting factor to evaluate in the future.

The following section provides the test of hypotheses presented in the General Introduction on p. 21.

Chapter 3.1: that a considerable bacterial diversity is present in New Zealand groundwater at national scale and there are identifiable relationships between bacterial diversity and environmental factors.

Accepted: The national scale study revealed that a considerable bacterial diversity is present in New Zealand groundwater. In addition, it was identified that the bacterial community structure is mainly related to the hydrochemistry whereas the other environmental factors might play a secondary role.

Chapter 3.2: that the relationships among bacterial diversity and environmental factors that are identified at a national scale are consistent and stable at a regional scale.

Accepted: The local scale study also suggested that the microbial diversity-hydrochemistry relationship is identifiable at regional level.

Chapter 3.3: that groundwater bacterial diversity is mainly related to the hydrochemistry, in particular to the redox potential of groundwater.

Accepted: This chapter provided more precise evidence for the relationship between bacterial diversity and hydrochemistry, in particular the redox potential of water. In addition, the study provided the taxonomic identities of the bacterial species present in groundwater environments, which has not been previously available in New Zealand.

*Chapter 3.4: that the Illumina high throughput sequencing methodology is a potential approach that can be used to identify HGT events that may have taken place in dominant *Pseudomonas* spp.*

Partly Accepted: Six groundwater bacterial metagenomes were successfully developed using the Illumina MiSeq sequencing platform. The initial analyses indicated some signs of HGT activities and further analyses will be conducted using these metagenomes as a resource to gain a better understanding of the bacterial genome changes as a response to the adaptation into diverse habitats. However, further analysis is needed to fully evaluate this hypothesis.

In conclusion, the present study has successfully characterized the bacterial diversity in New Zealand groundwater using molecular methods and identified the underlying factors that shape the microbial community composition. Overall, the knowledge generated in this project provides novel insights into New Zealand groundwater ecosystems and creates a scientific basis for the future inclusion of microbial status assessment criteria into regional and national groundwater monitoring programmes and related policies in New Zealand.

REFERENCES

- Alfreider A, Krössbacher M, Psenner R 1997. Groundwater samples do not reflect bacterial densities and activity in subsurface systems. *Water Research* 31: 832-840.
- Donachie SP, Foster JS, Brown MV 2007. Culture clash: challenging the dogma of microbial diversity. *The ISME Journal* 1: 97-102.
- Flynn TM, Sanford RA, Bethke CM 2008. Attached and suspended microbial communities in a pristine confined aquifer. *Water Resources Research* 44: W07425.
- Gregory SP, Maurice LD, West JM, Goody DC 2014. Microbial communities in UK aquifers: current understanding and future research needs. *Quarterly Journal of Engineering Geology and Hydrogeology*. doi: 10.1144/qjegh2013-059
- Griebler C, Lueders T 2009. Microbial biodiversity in groundwater ecosystems. *Freshwater Biology*, 54: 649-677.
- Griebler C, Mindl B, Slezak D, Geiger-Kaiser M 2002. Distribution patterns of attached and suspended bacteria in pristine and contaminated shallow aquifers studied with an in situ sediment exposure microcosm. *Aquatic Microbial Ecology* 28: 117-129.
- Griebler C, Malard F, Lefébure T 2014. Current developments in groundwater ecology—from biodiversity to ecosystem function and services. *Current Opinion in Biotechnology* 27: 159-167.

Lin X, McKinley J, Resch CT, Kaluzny R, Lauber CL, Fredrickson J et al. 2012. Spatial and temporal dynamics of the microbial community in the Hanford unconfined aquifer. *The ISME Journal* 6: 1665-1676.

Prosser JI 2010. Replicate or lie. *Environmental Microbiology* 12: 1806-1810.

Sirisena KA, Daughney CJ, Moreau-Fournier M, Ryan KG, Chambers GK 2013. National survey of molecular bacterial diversity of New Zealand groundwater: relationships between biodiversity, groundwater chemistry and aquifer characteristics. *FEMS Microbiology Ecology* 86: 490-504.

Sirisena KA, Daughney CJ, Moreau M, Ryan KG, Chambers GK 2013. Relationships between Molecular Bacterial Diversity and Chemistry of Groundwater in the Wairarapa Valley, New Zealand. *New Zealand Journal of Marine and Freshwater Research* (Accepted)

