Faecal source tracking and the identification of naturalised *Escherichia coli* to assist with establishing water quality and faecal contamination levels

Marie Moinet, Lauren Gadd, Megan Devane[a], David Wood[a], Brent Gilpin[a], Adrian Cookson

[a] Institute of Environmental Science Research Limited. Christchurch, New Zealand

June 2021

Report for Our Land & Water National Science Challenge

Contract Number: A26554

Output ID Number: 1808

agresearch
āta mātai, mātai whetū

ESR
Science for Communities

Inquiries or requests to:

Adrian Cookson
Adrian.Cookson@agresearch.co.nz
Consumer Interface Innovation Centre of Excellence,
AgResearch Ltd, Hopkirk Research Institute,
Cnr University Ave & Library Rd, Private Bag 11008, Palmerston North, New Zealand

Dr Gale Brightwell
Science Team Leader
Food System Integrity Team
Consumer Interface Innovation Centre of Excellence

Contents

1. Executive Summary

E. coli are routinely measured as faecal indicator bacteria (FIB) to provide indications of microbial water quality in parallel with other physico-chemical parameters. Recent evidence indicates that some 'naturalised' *Escherichia* species indistinguishable from *E. coli*, are widespread in the environment but are rarely associated with humans or ruminants. The omnipresence of these cryptic *Escherichia* species within the environment has led water managers to consider whether they confound microbial water quality assessments as they are phenotypically identical to *E. coli* in current standard detection methods. Likewise, elevated numbers of *E. coli* phylogroups B1 and B2 dominating a sample are potentially indicative of non-recent faecal sources as they may also persist in the environment.

This work used water samples collected from sites (observed land uses of dairy, urban and sheep & beef) with historically elevated numbers of *E. coli* as part of the Ministry for the Environment Quantitative Microbial Risk Assessment (QMRA) Pilot Study (Leonard *et al.*, 2020). The study aims were to determine the prevalence of *E. coli* phylogroups B1 and B2, and naturalised *Escherichia* species (*Escherichia marmotae* and *Escherichia ruysiae*), and determine any correlation with pathogen detection. Twenty separate isolates identified as *E. coli* on selective media, were recovered from each of 42 water samples (post-incubation colilert trays) and the respective phylogroup identified from the 840 isolates was identified using multiplex PCR. B1 and B2 were the most abundant phylogroups (B1, n= 475, 56.5%; B2, n=113, 13.5%). Cryptic *Escherichia* clades were rare (*E. ruysiae*/Clade IV, n=2, 0.24%; *E. marmotae*/Clade V, n=3, 0.36%).

The hypothesis was tested that identification of high numbers of *E. coli* B1 and/or B2 phylogroups in a water sample (as indicators of non-recent faecal pollution) would be associated with lower prevalence of pathogens. Pathogen data (presence/absence) from the 2020 QMRA Pilot Study were integrated with overall *E. coli* MPN/100ml water samples, and the prevalence of phylogroup B1 and/or B2 isolates from water sample enrichments. Pathogens were present in 93.1% (27 of 29) and 88.2% of water samples where B1 and/or B2 were present at 10 or greater, or greater than 15 isolates per sample respectively. Using logistic regression analysis, higher levels of generic *E. coli* appeared to be predictive for *Salmonella*, Norovirus GI, Norovirus GII, and viruses. High numbers of isolates from phylogroup B1 and/or B2 were significantly associated with the lower detection of the pathogens *Cryptosporidium* and *Salmonella*.

By targeting *gnd*, a hypervariable allele found in many *Enterobacteriaceae*, *E. coli* metabarcoding and population analysis of DNA recovered from water sample enrichments indicated that there were no significant variations of *gnd* diversity between urban, dairy and sheep and beef samples at the *gnd* sequence type (gST) level. Principle component analysis broadly grouped *E. coli* populations from each sample according to observed land use and faecal source marker (human, ruminant, and wildfowl). Cryptic *Escherichia* clades were rare at a relative abundance of <1% of reads per sample.

E. coli phylogroups B1 and B2 were identified frequently in the water samples from sites with historically high *E. coli* levels. B1 and B2 phylogroups of *E. coli* are derived from faecal material and are known to persist in waterways, and therefore, when identified as the dominant *E. coli* group(s) in a water sample, they are potentially naturalised *E. coli* and indicative of aged faecal sources. An important finding from this study was that where these naturalised *E. coli* phylogroup B1 and B2 were found to be the dominant *E. coli* in a water sample, it was in association with one or more pathogens. This indicates that naturalised faecal *E. coli* present in a waterway are still likely to represent a significant health risk.

Naturalised cryptic *Escherichia* species are non-*E. coli* species and are not highly prevalent in animal and human faeces. These naturalised non-faecal *Escherichia* have been identified as consistently contributing to the low concentrations of *E. coli* in environmental samples (water, soil, sediment, periphyton) from pristine sites and those with low anthropogenic influences. Furthermore, these naturalised *Escherichia* species were identified infrequently (0.6%, 5 of 840 isolates) in the 42 samples examined in this study. This latter finding suggests that naturalised non-*E. coli Escherichia* species do not confound microbial water quality monitoring at sites where *E. coli* monitoring and faecal source markers indicate faecal contamination.

2. Background

Worldwide, naturally occurring *Escherichia coli* (*E. coli*) from the gut of warm-blooded animals (including birds) (Tenaillon *et al.*, 2010) remains the preferred indicator of faecal contamination for water quality monitoring (Anon, 2003, Anon, 2017). However, current culture-based methods, used to enumerate *E. coli* as a proxy for faecal contamination and pathogenic microorganisms, cannot distinguish between naturalised *Escherichia* and faecal strains. Recent studies have demonstrated that at least four benign 'cryptic' *Escherichia* clades; with three given species designation; Clades III and IV *Escherichia ruysiae* (van der Putten *et al.*, 2021), and Clade V *Eschericia marmotae* (Liu *et al.*, 2015). These new species are indistinguishable from generic *E. coli* using diagnostic biochemical reactions and are able to survive and grow in the environment (Byappanahalli *et al.*, 2006, Walk *et al.*, 2009, Berthe *et al.*, 2013) where they were found in >90% of water samples (Cookson *et al.*, 2017), and comprising up to 48% of total '*E. coli*' (Devane & Gilpin, 2019). Under the current testing regimen, this could cause faecal *E. coli* contamination to be overestimated considerably and new tools are urgently required to distinguish benign, naturalised *Escherichia* from faecal *E. coli*.

Sequencing of housekeeping genes within *E. coli* has provided important phylogenetic insights into intra-species variation with the identification of at least seven distinct phylotypes (Clermont *et al.*, 2013). Paradoxically, some *E. coli*, particularly phylotypes B1 and B2, have also been associated with prolonged survival in environmental samples (e.g., water and sediment), where no obvious faecal contamination event has been noted (Touchon *et al.*, 2020). Thus, there appear to be two groups of naturalised *Escherichia*; from environmental, and enteric sources, but the extent to which either or both confound microbial water quality assessments is unknown (Devane *et al.*, 2020).

In this work we provide detailed analysis of *E. coli* obtained from a subset of water samples obtained as part of a Ministry for the Environment Quantitative Microbial Risk Assessment (QMRA) Pilot Study (Leonard *et al.*, 2020) undertaken to validate methods for the design of a large-scale replacement study of the 1998-2000 QMRA. The sixteen pilot study sample sites were selected due to their historically elevated numbers of *E. coli*, and were representative of three different observed land uses (dairy, urban, and sheep and beef). The sites were geographically distributed around New Zealand (nine North Island, seven South Island). The 2020 Pilot Study (Leonard *et al.*, 2020) enabled a selection of new laboratory methodologies to be trialled, and data, including the presence/absence of defined waterborne human pathogens (including bacteria, protozoa and viruses), and potential faecal source, are used within this work.

Detailed analysis of the bacterial isolates from this study and the potential role of cryptic *Escherichia* clades and naturalised *E. coli* in confounding microbial water quality assessments will contribute to updated guidelines for regional councils and water managers. Implementation of new guidelines will provide an improved understanding of human health risk and assist decision making processes for appropriate mitigations to reduce FIB.

3. Method

3.1 Study sites and sample collection

The original water samples were collected as part of the Ministry for the Environment QMRA Pilot Study (Leonard *et al.*, 2020) from 16 sites around New Zealand as part of a repeated cross-sectional pilot study (Figure 1). Collection sites encompassed three different observed land uses (urban, sheep and beef, and dairy) having historically elevated levels of *E. coli*. Between 13 Feb 2020 and 19 March 2020, 42 pre-incubated (1 in 10 dilution of water sample, incubated for 18 hours at 35°C) Colilert 2000 trays (IDEXX, New Zealand) were transported (maintained at 4°C during overnight transit) in single weekly batches to the Hopkirk Research Institute (Palmerston North) from ESR Christchurch.

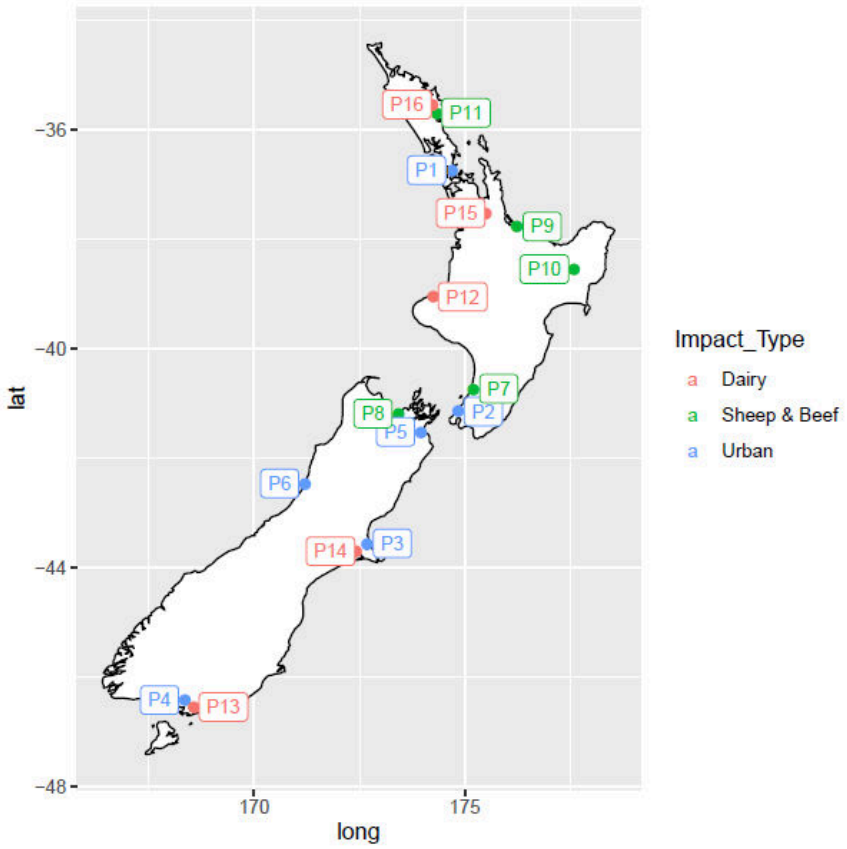


Figure 1: Geographical distribution of sample sites across New Zealand and three observed land uses

3.2 *E. coli* culture, isolation and storage

UV-positive Colilert wells indicative of the presence of *E. coli* were identified and 150µl (small wells) and/or 200µl (large wells) of growth removed from individual Colilert wells using a sterile needle. Bacterial culture from the fluorescent wells of each individual Colilert plate was pooled and centrifuged at 5,000 x g at 4°C (Multifuge X3R) if there was greater than 1.5ml pooled culture or 10,200 x g at room temperature in a bench-top centrifuge if there was less than 1.5ml pooled culture. Bacterial pellets were resuspended in 1ml EC broth (Fort Richard, New Zealand), added to 450µl glycerol in sterile cryovials. The cryovials were gently inverted until the glycerol and bacterial culture were well-mixed, and stored at -80°C.

Twenty *E. coli* isolates were obtained from each glycerol stock by plating onto three separate ECC CHROMagar (Fort Richard, New Zealand) plates. Plates were incubated at 35°C for 18 to 21 hours. Separate, well-spaced blue colonies (n=20) were subcultured onto MacConkey agar plates (Fort Richard, New Zealand) and incubated as before. A single well-spaced colony was resuspended in 400µl sterile milliQ water, heated at 100°C for 10mins and stored at -20°C to provide a boiled lysate preparation for subsequent DNA analysis. Further growth from each isolation plate was resuspended in cryovials containing 1.5ml Brain Heart Infusion broth (Fort Richard, New Zealand) containing glycerol (33% w/v) and stored at -80°C. In total 840 presumptive *E. coli* isolates were stored for further analysis.

3.3 Multiplex PCR to identify *E. coli* phylogroup

Each of the 840 boiled lysate preparations was used as DNA template in separate PCR reactions to identify the *E. coli* phylogroup or *Escherichia* cryptic Clades (II to V) using an extended quadruplex PCR phylogroup assignment method (Clermont *et al.*, 2013). The identity of Cryptic *Escherichia* Clade V as *E. marmotae* (Liu *et al.*, 2015) was confirmed using a MALDI Biotyper (Bruker) and a custom-made database that included spectra from an isolate confirmed as *E. marmotae* using whole genome sequence analysis.

3.4 *E. coli* phylogroup pathogen association and logistic regression

Two separate *E. coli* phylogroup B1 and B2 abundances (10 or higher, and greater than 15) were used to understand their potential role as predictors of pathogens and human health risk incorporating the presence/absence of pathogens (*Campylobacter*, *Salmonella*, Shiga toxin-producing *E. coli*, *Cryptosporidium*, *Giardia*, Human adenovirus, Norovirus, and Enterovirus) determined using quantitative PCR from the QMRA Pilot Study data (Leonard *et al.*, 2020). Additionally, these same data were included in an exploratory analysis, where logistic regression was undertaken to investigate the association of phylogroups B1 and B2 with pathogens.

3.5 *gnd* metabarcoding for *E. coli* population analysis

A frozen slurry (400µl) from each glycerol preparation of pooled Colilert well growth (n=42) was inoculated onto 10ml pre-warmed EC broth and incubated at 37°C for 3 hours with shaking. Broth supernatant was removed by centrifugation (5,000 x g at 4°C) and DNA extracted from the cell pellet using the Geneaid Presto Mini gDNA Bacteria kit (DNAture, New Zealand). For metabarcoding PCR, individually barcoded PCR amplicons were generated using methods described previously (Cookson *et al.*, 2017). For *E. coli* community analysis, the hypervariable 284bp region of the *gnd* gene from DNA samples was amplified using indexed 2*gnd*F and 2*gnd*R PCR primers (Cookson *et al.*, 2017). The individually indexed PCR products were pooled to create a mixture of libraries for sequencing (Illumina MiSeq (version 2 chemistry) 2 x 250 base pair paired end analysis, Massey Genome Service, Massey University, Palmerston North, New Zealand). Sequence reads were processed using dada2 (Callaghan *et al.*, 2016) and then mapped against *gnd*Db (Cookson *et al.*, 2019), a database containing almost 600 unique 284bp *gnd* sequences, to provide information on the *E. coli* community profile associated with individual sample enrichments. Principal Component Analysis (PCA) was used to simplify the population data of each sample to identify trends, and links associated with sample, land use and faecal source tracking qPCR data obtained from ESR as part of their contribution to the QMRA Pilot Study.

3.6 Whole genome sequencing

DNA extractions and library preparations for whole genome sequencing (WGS) were undertaken using standard methods. WGS was undertaken by Novagene Limited (Beijing, China) using the illumina HiSeq paired end v4 platform (2 x 125 bp). Ten isolates (5 phylogroup F, 1 phylogroup B1, 1 *Escherichia ruysiae* [Clade IV] and 3 *Escherichia marmotae* [Clade V]) underwent WGS. WGS read data is anticipated from the sequence provider in July 2021 whereupon read assemblies and subsequent genomic epidemiological analysis will be undertaken.

4. Results and Discussion

4.1 Faecal source and phylogroup analysis

Metadata associated with the 42 Colilert trays received are included in Table 1. There were 11 water samples from sheep and beef, 14 from dairy and 17 from urban land uses. Wildfowl sources were identified at all sites but usually as a secondary source to the dominant faecal sources outlined below. Where the land use was designated as sheep and beef, ruminant-associated faecal source markers were identified from 8 of 11 (72.7%) water samples either as the dominant source (n=6) or together with human faecal source markers (n=2). Wildfowl faecal source markers were identified as the dominant source in 2 (18.2%) water samples and the remaining sample was not tested for faecal source markers. For observed dairy land use, ruminant-associated faecal source markers were identified from 10 of 14 (71.4%) water samples as the dominant source (n=7), or with human faecal source markers (n=2), or with wildfowl (n=1). Wildfowl faecal source markers were the sole faecal source marker in the remaining 4 water samples but occurred in 3 samples at low levels. Human faecal source markers were found in all 17 urban water samples as the dominant faecal source (n=14) or in conjunction with ruminant faecal source markers (n=3).

For each Colilert sample, 20 putative *E. coli* isolates were characterised from positive wells according to the quadruplex PCR phylogroup method (Clermont *et al.*, 2013) and assigned to one of the seven *E. coli* phylogroups or to members of the cryptic *Escherichia* clades. Overall, B1 was the most abundant *E. coli* phylogroup isolated from water samples obtained across all three observed land use sample sites (Figure 2). B1 dominated all sites particularly dairy (193/280, 68.93%), but urban sites were associated with an increased abundance of phylogroup B2 (68/340, 20.0%). In the developed world, B2 has been recognised as the dominant *E. coli* phylogroup in human faeces, whereas, in general, animal intestinal microbiota are dominated by phylogroup B1 (Vadnov *et al.*, 2017, Ambrosi *et al.*, 2019). Overall samples from urban sites had a greater variety of phylogroups (average phylogroups per 20 isolates = 4.35) compared to sheep and beef (3.82), and dairy (3.29). Cryptic *Escherichia* clades were rarely identified in these 42 water samples (*E. ruysiae*/clade IV, n=2, 0.24%; *E. marmotae*/clade V, n=3, 0.36%).

Table 1: Dominant faecal source(s), MPN per 100ml phylogroup and associated metadata from 42 water samples included in this study.

Sample	*Site*	*Landuse*	*Faecal Source*	*E. coli MPN*	*A*	*B1*	*B2*	*C*	*D*	*E*	*F*	*U*[b]	*Clade I*	*Clade IV*	*Clade V*	*Total*[c]
CMB200071	P13	Dairy	Human&Ruminant	340	0	20	0	0	0	0	0	0	0	0	0	1
CMB200072	P15	Dairy	Ruminant	1800	0	16	0	1	0	0	0	3	0	0	0	3
CMB200095	P14	Dairy	Ruminant	630	0	16	0	0	0	3	0	1	0	0	0	3
CMB200098	P16	Dairy	Wildfowl	290	0	17	2	0	0	0	1	0	0	0	0	3
CMB200099	P12	Dairy	Ruminant&Wildfowl	10	0	20	0	0	0	0	0	0	0	0	0	1
CMB200194	P12	Dairy	Ruminant	860	2	11	0	0	3	2	0	0	0	2	0	5
CMB200195	P13	Dairy	Ruminant	300	0	18	0	2	0	0	0	0	0	0	0	2
CMB200196	P15	Dairy	Ruminant	1000	3	7	0	6	3	1	0	0	0	0	0	5
CMB200217	P14	Dairy	Ruminant	540	1	8	5	0	4	2	0	0	0	0	0	5
CMB200220	P16	Dairy	Wildfowl	75	0	0	0	9	0	11	0	0	0	0	0	2
CMB200221	P12	Dairy	Ruminant	120	0	20	0	0	0	0	0	0	0	0	0	1
CMB200236	P13	Dairy	Human&Ruminant	2500	1	16	0	1	1	1	0	0	0	0	0	5
CMB200237	P15	Dairy	Wildfowl	1700	1	10	0	6	0	3	0	0	0	0	0	4
CMB200269	P14	Dairy	Wildfowl	310	2	11	2	0	1	3	0	0	1	0	0	6
CMB200067	P8	Sheep&Beef	Ruminant	63	0	10	0	10	0	0	0	0	0	0	0	2
CMB200068	P9	Sheep&Beef	Human&Ruminant	1000	4	1	11	0	0	3	0	0	0	0	1	5
CMB200097	P11	Sheep&Beef	Wildfowl	200	0	16	2	0	0	2	0	0	0	0	0	3
CMB200189	P9	Sheep&Beef	Ruminant	880	0	18	1	0	0	0	0	0	1	0	0	3
CMB200193	P8	Sheep&Beef	ND[a]	52	0	11	0	0	0	9	0	0	0	0	0	2
CMB200218	P10	Sheep&Beef	Ruminant	85	0	19	1	0	0	0	0	0	0	0	0	2
CMB200219	P11	Sheep&Beef	Wildfowl	410	0	17	1	0	1	1	0	0	0	0	0	4
CMB200233	P8	Sheep&Beef	Ruminant	250	0	14	2	3	0	0	0	0	1	0	0	4
CMB200234	P9	Sheep&Beef	Ruminant	1100	0	14	0	2	1	1	0	2	0	0	0	5
CMB200238	P10	Sheep&Beef	Ruminant	130	0	1	17	0	1	1	0	0	0	0	0	4
CMB200267	P7	Sheep&Beef	Human&Ruminant	570	3	5	1	1	3	3	1	0	3	0	0	8
CMB200065	P2	Urban	Human	550	6	10	4	0	0	0	0	0	0	0	0	3
CMB200066	P5	Urban	Human	97	5	11	0	2	0	2	0	0	0	0	0	4
CMB200070	P4	Urban	Human	5200	1	12	1	0	1	5	0	0	0	0	0	5
CMB200091	P1	Urban	Human	160	0	0	20	0	0	0	0	0	0	0	0	1
CMB200092	P6	Urban	Human&Ruminant	3400	2	7	3	5	1	1	0	1	0	0	0	7
CMB200094	P3	Urban	Human	200	0	18	0	2	0	0	0	0	0	0	0	2
CMB200190	P4	Urban	Human	460	0	9	2	3	4	1	1	0	0	0	0	6
CMB200191	P5	Urban	Human	510	1	15	1	3	0	0	0	0	0	0	0	4
CMB200192	P2	Urban	Human	4900	3	11	1	1	2	1	0	0	1	0	0	7
CMB200214	P6	Urban	Human&Ruminant	1300	1	12	5	0	0	1	0	0	0	0	1	5
CMB200216	P3	Urban	Human	410	2	5	11	0	0	0	1	1	0	0	0	5
CMB200231	P2	Urban	Human	7700	0	7	3	5	3	0	2	0	0	0	0	5
CMB200232	P5	Urban	Human	1000	4	7	0	8	0	1	0	0	0	0	0	4
CMB200235	P4	Urban	Human&Ruminant	12000	1	17	0	1	0	1	0	0	0	0	0	4
CMB200265	P1	Urban	Human	190	0	10	9	0	1	0	0	0	0	0	0	3
CMB200266	P6	Urban	Human	6100	12	2	0	5	0	1	0	0	0	0	0	4
CMB200268	P3	Urban	Human	530	0	6	8	4	1	0	0	0	0	0	1	5

[a] ND, No faecal source detected; [b] U, untyped with no phylogroup identified; [c] Total number of phylogroups identified amongst 20 colonies obtained from each water sample

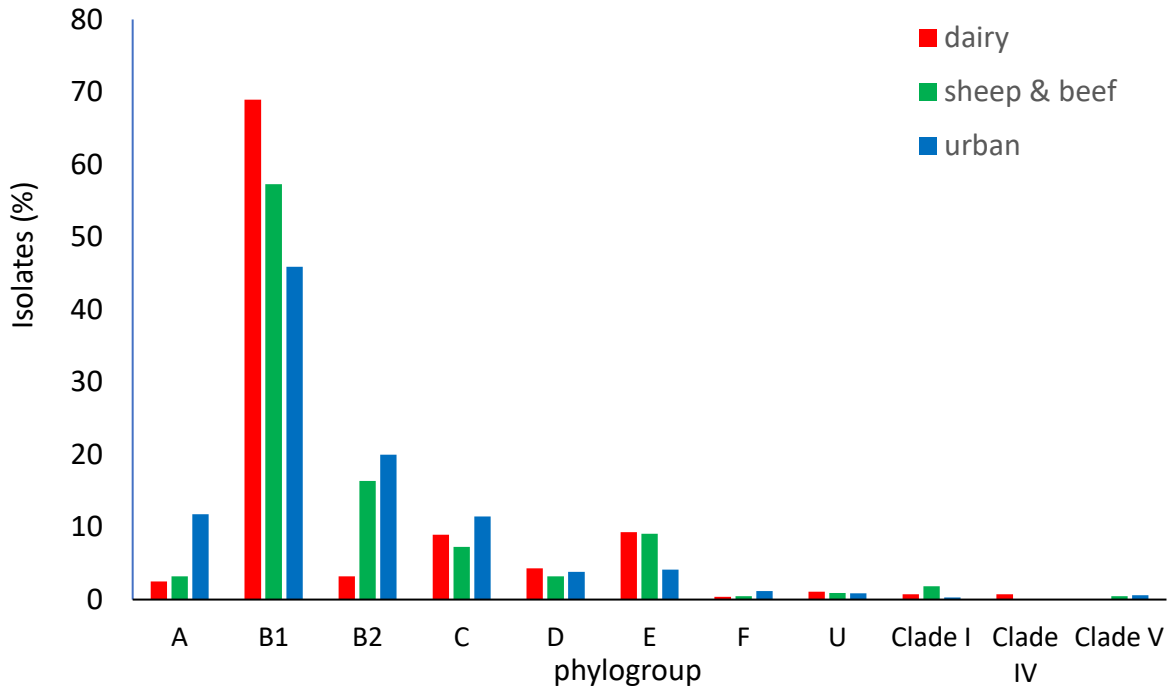


Figure 2: Distribution of isolate phylogroups according to the observed land use.

4.2 Logistic regression analysis of *E. coli* and pathogen data

E. coli subtypes, especially those belonging to phylogroups B1 and B2, can persist in the aquatic environment as reviewed by Devane *et al.* (2020). These faecal *E. coli* may still be associated with persistent pathogens, and as such represent a health risk.

The hypothesis was tested that identification of high numbers of *E. coli* B1 and/or B2 phylogroups in a water sample (as indicators of non-recent faecal pollution) would be associated with lower prevalence of pathogens. Pathogen data (presence/absence) from the 2020 QMRA Pilot Study was integrated with overall *E. coli* MPN per 100ml water samples, and the prevalence of phylogroup B1 and/or B2 isolates from water sample enrichments. Pathogens were detected in 95.2% (n=40) protozoa in 85.7% (36), bacteria in 69% (29) and viruses in 26.2% (11) of water samples (n=42) (Table 2). The protozoan *Giardia* was identified in the highest frequency (81%), followed by *Campylobacter* (69%) and then the protozoan *Cryptosporidium* (43%).

Table 2: Rate and count of pathogens determined as part of QMRA Pilot Study (2020)

Microbe	Rate (%)	Count
Pathogen	95.2	40
Protozoa	85.7	36
Giardia	81.0	34
Bacteria	69.0	29
Campylobacter	69.0	29
Cryptosporidium	42.9	18
Virus	26.2	11
Norovirus GII	16.7	7
Norovirus GI	14.3	6
Salmonella	14.3	6
Enterovirus	4.8	2
Shiga toxin-producing *E. coli*	2.4	1
Human adenovirus	0	0

The average (mean) level of *E. coli* for the 42 water samples was 1426.7 MPN per 100ml water and *E. coli* phylogroup B1 (mean 10.9 per 20 isolates) was detected at a higher rate (per 20 isolates) than phylogroup B2 (mean 2.6/20 isolates) (Table 3).

Table 3: Measures of generic Colilert '*E. coli*' counts, and phylogroups B1 and B2 determined using multiplex PCR

Indicator	Mean	SD	Median
'E. coli'	1426.7	2421.8	520.0
Phylogroup B1	10.9	6.0	11.0
Phylogroup B2	2.6	4.6	1.0

When *E. coli* phylogroups B1 and/or B2 were present at 10 or higher isolates per water sample, one or more pathogens was detected in 93.1% (27 or 29) samples. For the criteria where B1 and or B2 were present at greater than 15 isolates per sample (20 isolates in total), then 88.2% (15 of 17) water samples were positive for one or more pathogens. This criteria of >15/20 B1 or B2 isolates per sample was suggested as indicating that the water body may have contained non-recent faecal inputs. The identification of pathogens in samples with high numbers of B1 or B2 isolates is suggesting that non-recent faecal inputs are still associated with health risks.

Box plots were used to examine the relationship between *E. coli* phylogroups B1, B2, combined B1 and B2, generic *E. coli* and pathogens (Appendix 1 to 4). Zero (0) implies that the pathogen was absent from the water sample and one (1) indicates that it was present. For most of the box plots there is an overlap between the numbers of isolates from each sample of the phylogroup B1 and B2 indicators and the pathogens. Low isolate numbers of B1 and B1 + B2 combined were predictive of *Cryptosporidium* and *Salmonella* (Appendix 1 and 3). In contrast phylogroups B1 + B2 combined are not predictors of reduced risk from other pathogens particularly *Giardia* and norovirus

Genotype I (Appendix 3). Phylogroup B2 were rare and were not predictive of any pathogens (Appendix 2). High levels of generic *E. coli* were predictive of *Salmonella*, norovirus Genotypes I and II, and viruses (Appendix 4).

Logistic regression was also used to analyse any relationship between *E. coli* MPN and the presence/absence of pathogens and between numbers of isolates per sample from phylogroups B1 and/or B2 with pathogens determined previously in the QMRA Pilot Study (2020). Although not true as multiple water samples were collected from the same site, it is assumed that the water samples included in this work were collected randomly. This simplifying assumption is appropriate for this exploratory analysis of these 42 water samples and 840 *E. coli* isolates. Further, more detailed analysis of the hypotheses generated by this data would require reconsideration of the study design, including expansion of the number of samples for analysis.

Tables 4 to 6 use

$$\log(odds) = \log(P/(1-P))$$

for analysis. A positive sign suggests the probability of detecting a pathogen increases as the number of isolates from *E. coli* phylogroups B1 and/or B2 increases in a water sample. A negative sign suggests that the probability of detecting the pathogen decreases with increasing numbers of isolates from *E. coli* phylogroups B1 and/or B2.

$$log\ (P/(1-P)) = beta0 + beta1 \times x1$$

where $x1$ is substituted with the numbers of isolates of B1 or B2 per sample etc. The P value tests the null hypothesis that the slope ($beta1$) is zero. That is, numbers of B1 and/or B2 have no effect on pathogen detection.

Low numbers of phylogroup B1 per sample appeared to be negatively predictive for *Cryptosporidium* (Figure 3 and Table 4). This observation indicated that when high numbers of phylogroup B1 are present in a water sample, as an indicator of aged faecal material, there is a lower likelihood of that sample containing the protozoan *Cryptosporidium*.

B1 model

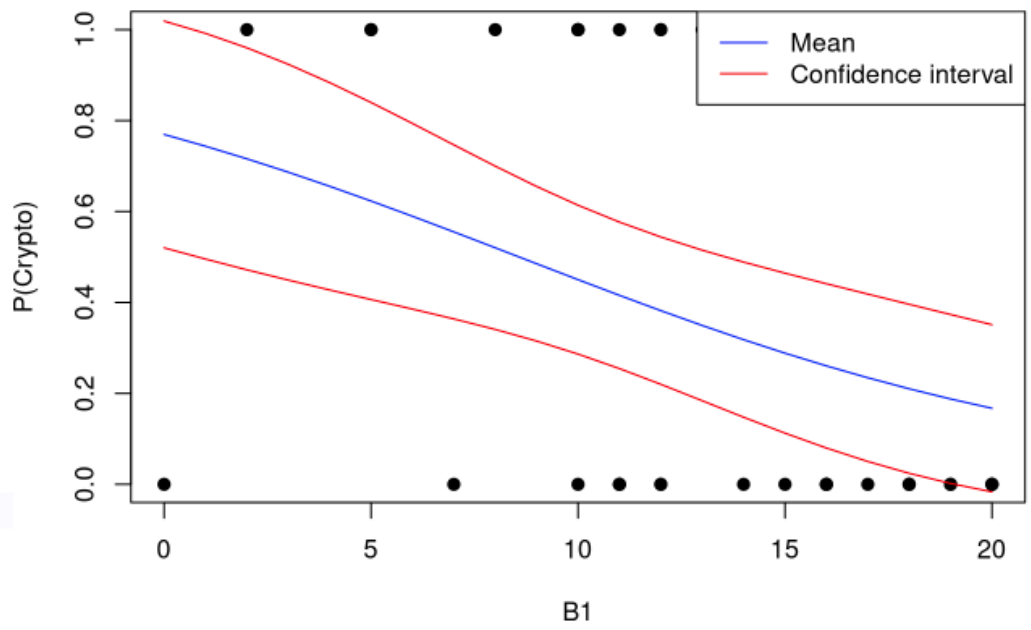


Figure 3: The model shows as B1 increased the probability of detecting *Cryptosporidium* decreases. The odds of detecting *Cryptosporidium* for each additional B1 count is decreased between 0.24 and 0.03.

Table 4: Logistic regression of *E. coli* phylogroup B1 and presence/absence of pathogens determined in QMRA Pilot Study (2020).

Microbe	$beta0$	$beta1$	P sig
Campylobacter	0.845	-0.004	0.944
Salmonella	-0.314	-0.165	0.058
Shiga toxin-producing *E. coli*	-6.413	0.197	0.429
Cryptosporidium	1.205	-0.141	0.021*
Giardia	2.568	-0.094	0.201
Human adenovirus	-25.566	0.000	1.000
Norovirus GI	-2.923	0.094	0.201
Norovirus GII	-1.139	-0.045	0.517
Enterovirus	-3.379	0.034	0.790
Bacteria	0.845	-0.004	0.944
Protozoa	4.205	-0.186	0.069
Virus	-0.924	-0.010	0.861
Pathogen	15.735	-0.762	0.158

In contrast to *E. coli* phylogroup B1, phylogroup B2 was not able to predict the presence or absence of any pathogens (Appendix 5). This is likely due to the scarcity of the *E. coli* phylogroup B2 isolates observed in this study rather than there being a lack of association. Therefore, the data for B1 and B2 phylogroups were combined in the following statistical analysis.

Low numbers of *E. coli* phylogroups B1 and B2 isolates out of the total of 20 *E. coli* isolates per water sample appeared to be negatively predictive for *Salmonella* (Figure 4 and Table 5). As observed with phylogroup B1 and *Cryptosporidium*, this observation indicated that when high numbers of phylogroups B1 and B2 are present in a water sample as an indicator of aged faecal material, there is a lower likelihood of that sample containing *Salmonella*.

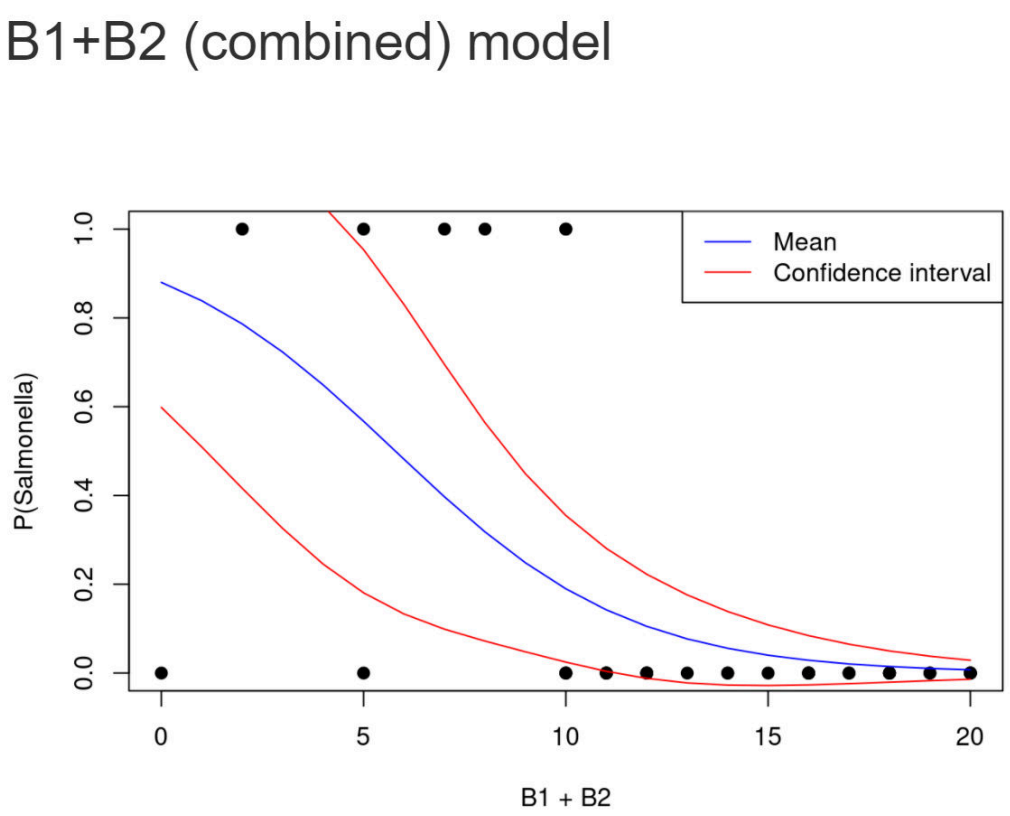


Figure 4: The model shows as B1 and B2 increased the probability of detecting *Salmonella* decreases. The odds of detecting *Salmonella* for each additional count is decreased between 0.49 and 0.12.

The following hypothesis was based on the premise that high numbers of *E. coli* phylogroups B1 and B2 isolates present in a water sample are indicative of non-recent/aged sources of faecal contamination.

Hypothesis: where there are high numbers of *E. coli* phylogroups B1 and/or B2 isolates in a water sample, there is a lower likelihood of identifying *Cryptosporidium* and *Salmonella*. The risk from other pathogens (*Campylobacter*, viruses and *Giardia*), however, remains the same as if the contamination was from fresh sources of faecal pollution. In conclusion, non-recent/aged sources of contamination are likely to be associated with pathogens and represent a health risk to recreational water users.

However, this hypothesis is generated from analysis of a relatively small dataset of 42 samples from sites with historically high levels of *E. coli* where pathogens were detected in 40 of 42 samples (95.2%).

Table 5: Logistic regression of combined *E. coli* phylogroups B1 and B2 and presence/absence of pathogens determined in QMRA Pilot Study (2020)

Microbe	$beta0$	$beta1$	P sig
Campylobacter	2.305	-0.107	0.163
Salmonella	1.992	-0.344	0.009**
Shiga toxin-producing *E. coli*	-5.607	0.128	0.622
Cryptosporidium	1.479	-0.132	0.058
Giardia	1.365	0.006	0.938
Human adenovirus	-25.566	0.000	1.000
Norovirus GI	-2.119	0.024	0.792
Norovirus GII	-1.386	-0.017	0.837
Enterovirus	-2.996	0.000	1.000
Bacteria	2.305	-0.107	0.163
Protozoa	3.201	-0.098	0.342
Virus	-0.874	-0.012	0.861
Pathogen	13.902	-0.633	0.207

The use of the *E. coli* MPN per 100ml water data appeared to be predictive for *Salmonella*, Norovirus GI, Norovirus GII, and viruses where the probability of detecting these pathogens increases with increasing *E. coli* (Table 6). Water managers and Regional Councils are guided by modelling data that links the probability of increased generic *E. coli* in water samples with increased pathogens (especially *Campylobacter*). In contrast to *E. coli* phylogroups B1 and B2, increases in the identification of *Salmonella* in water samples were correlated with generic *E. coli*, potentially linking *Salmonella* presence to more recent faecal sources.

Table 6: Logistic regression of *E. coli* MPN per 100ml and presence/absence of pathogens determined in QMRA Pilot Study (2020)

Microbe	$beta0$	$beta1$	P sig
Campylobacter	-1.662	0.932	0.110
Salmonella	-7.341	1.880	0.030*
Shiga toxin-producing *E. coli*	-4.319	0.220	0.890
Cryptosporidium	-1.127	0.309	0.533
Giardia	2.894	-0.521	0.409
Human adenovirus	-25.566	0.000	1.000
Norovirus GI	-8.966	2.391	0.013*
Norovirus GII	-7.088	1.868	0.023*
Enterovirus	-5.507	0.875	0.456
Bacteria	-1.662	0.932	0.110
Protozoa	2.599	-0.293	0.674
Virus	-5.661	1.625	0.019*
Pathogen	-0.370	1.364	0.247

4.3 *E. coli* metabarcoding and population analysis

The hypervariable gene *gnd* was targeted using a metabarcoding amplicon sequencing method to generate separate amplicon sequence variants (ASV) that could be differentiated by a single nucleotide difference. After filtering and denoising of sequencing reads, on average 73,396 reads per library were used to generate *E. coli* population data for each of the 42 samples. Although there were 992 separate *gnd* sequence types (gSTs) identified using the gndDb, the 100 most abundant gSTs made up 78.9% of the total reads. Generally increasing numbers of *E. coli* (MPN per 100ml) was associated with increasing numbers of gSTs (≥ 10 reads) within a sample (Figure 7). Reads corresponding to cryptic Clade IV and Clade V *Escherichia* gSTs were found in sequence data from 59.5% (25 of 42) water samples but were uncommon with a maximum relative abundance of 0.85%. This result is supported by the low numbers of *Escherichia* cryptic clade isolates (0.6%) identified in the characterisation of putative *E. coli* phylotypes from Colilert water samples using the quadruplex PCR phylogroup assignment method (Clermont *et al.*, 2013).

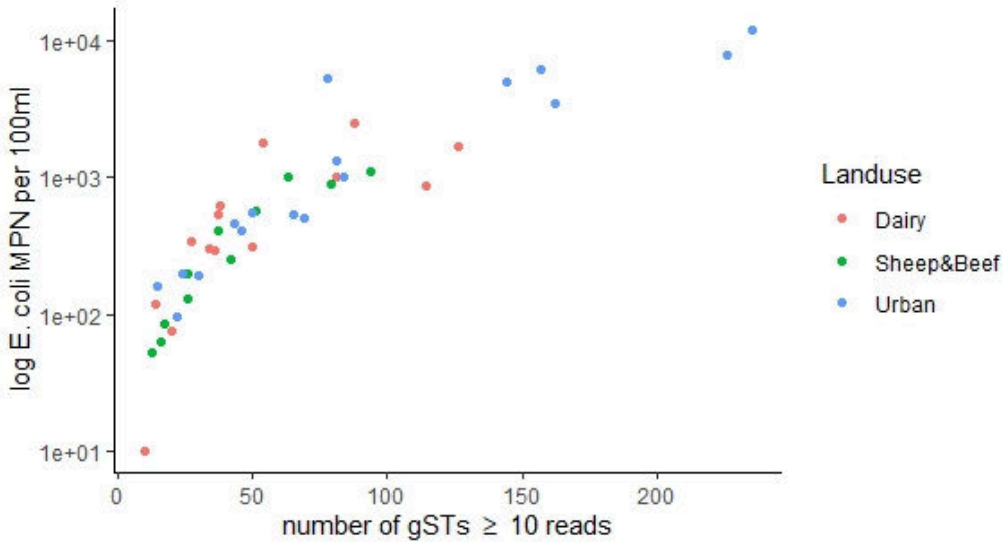


Figure 7: As the number of *E. coli* (MPN per 100ml) in each water sample increased so did the richness or number of *gnd* sequence types (gSTs)

When mapped against the archived *gnd* database, gndDb, only seven of the 100 most abundant sequence types were novel; BLAST analysis (Altschul *et al.*, 1990) indicated that three of the novel gSTs corresponded to *Enterobacter gnd* sequences (100% homology match), and the remaining four novel gSTs were matched to *E. coli gnd* sequences (98.90 to 99.65% homology match).

Shannon and Simpson alpha diversity measures indicated that water samples obtained from observed urban and dairy land use sites appeared more diverse that those from sheep and beef sample sites (Figure 8), but these data were not significant ($p>0.05$).

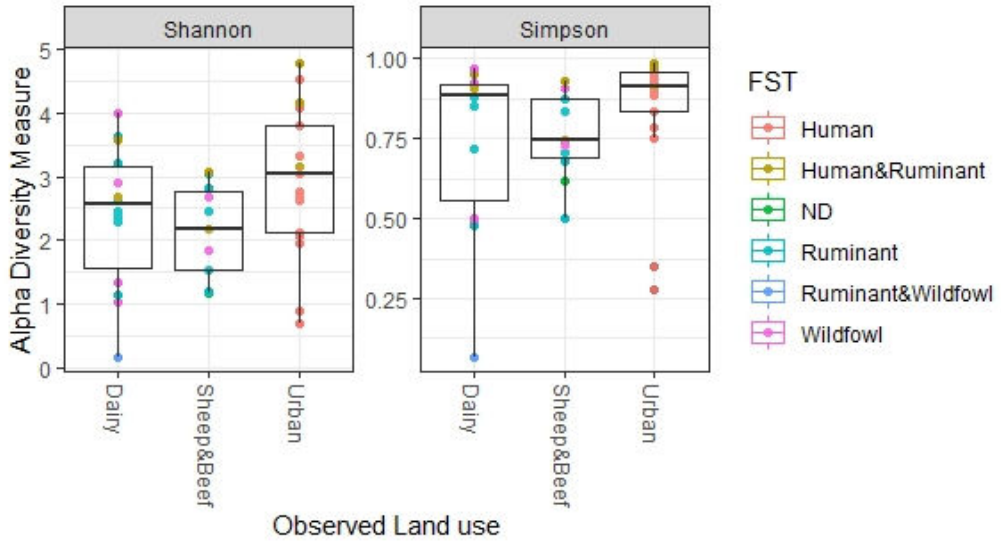


Figure 8: Shannon and Simpson alpha diversity measures indicated that *E. coli* populations from observed urban land use water collection sites were more variable than those from dairy or sheep and beef. ND - no faecal source detected.

Principle component analysis was able to broadly separate out *E. coli* populations from water samples according to land use and faecal source tacking data (Figure 9). These data taken at the *E. coli* population level may indicate that some host specificity occurs with *E. coli*. One data point (marked with an asterisk in Figure 9) was from a water sample obtained from a site where the observed land use was urban and was thought to be an outlier because of a recent heavy rainfall event in the catchment, dominated by pastoral farming.

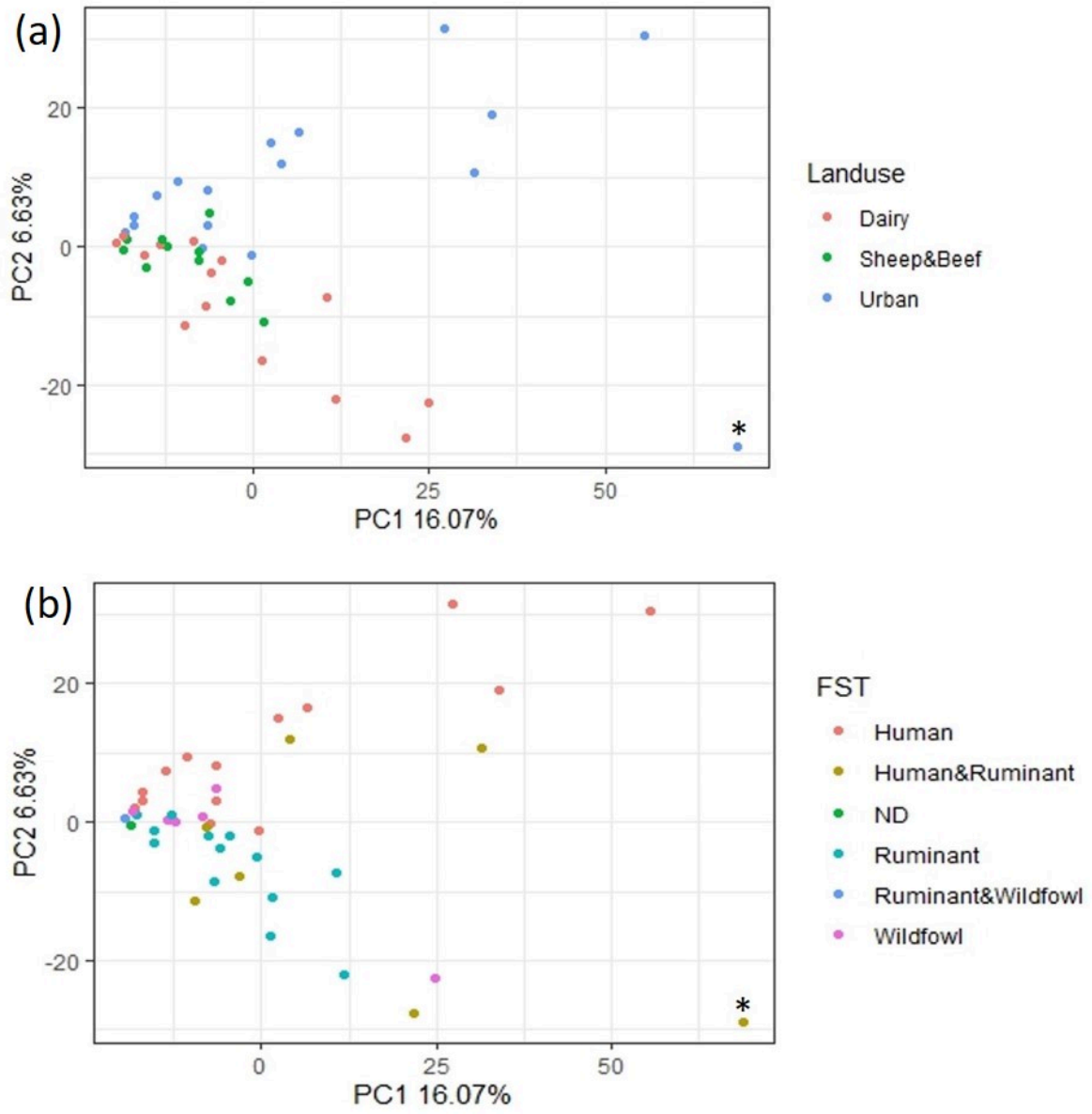


Figure 9: Principal component analysis of *E. coli* populations obtained from water samples were able to be separated based on observed land use and dominant faecal source(s). Asterisk highlights water sample from observed urban land use collection site where the site upper catchment was dominated by pastoral farming.

5. Recommendations

In summary, where faecal source markers are consistently identified in a waterway then it is unlikely to be necessary to investigate naturalised FIB sources as the *E. coli* identified by testing will be derived from faecal inputs and are likely to be associated with pathogens. These FIB exceedances will, therefore, represent a health risk even if the contamination is from aged sources.

It may be valid, however, for water managers to investigate whether naturalised sources of (non-*E. coli*) cryptic *Escherichia* species such as *E. marmotae* and *E. ruysiae*, are contributing to *E. coli* exceedances where there is a consistent non-identification of faecal source markers and faecal contamination pathways into a waterway**.**

If naturalised *Escherichia* clades are identified as the dominant/sole source of elevated *E. coli* levels then this exceedance may represent a lower likelihood of health risk to recreational users of the target water body. A threshold for *E. coli* monitoring data may need to be established for a location where high percentages of cryptic *Escherichia* have been identified as contributing to the *E. coli* exceedance (Devane, 2019). Establishment of this threshold for naturalised cryptic *Escherichia* contributions would require multiple sampling events, which account for seasonal effects. Routine monitoring of this water body would be required to detect spikes in faecal *E. coli* above the threshold, with appropriate tools such as faecal source tracking used to evaluate sources and monitor health risks.

6. Acknowledgements

We acknowledge the assistance of all the Regional Council employees who collected the initial water samples in support of the 2020 QMRA Pilot Study. We are appreciative of ESR collaborators making available the data obtained from the 2020 QMRA Pilot Study for inclusion in this study. Lauren Gadd was supported by a Pūhoro STEM Academy Summer Internship Scholarship (November 2020 to February 2021) funded by the Our Land & Water National Science Challenge. Many thanks to Richard Muirhead (AgResearch), and Massey University and NIWA partners for their helpful input into this work, especially Jonathan Marshall for help with Principal Component Analysis R code. Overall funding for this work was through Our Land & Water National Science Challenge Extension Funding.

7. References

Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403-410.

Ambrosi C, Sarshar M, Aprea MR*, et al.* (2019) Colonic adenoma-associated *Escherichia coli* express specific phenotypes. *Microbes and Infection* **21**: 305-312.

Anon (2003) Microbiological water quality guidelines for marine and freshwater recreational areas. Ministry for the Environment. Wellington, New Zealand.

Anon (2017) National Policy Statement for Freshwater Management 2014 (amended 2017). Ministry for the Environment. Wellington, New Zealand.

Berthe T, Ratajczak M, Clermont O, Denamur E & Petit F (2013) Evidence for coexistence of distinct *Escherichia coli* populations in various aquatic environments and their survival in estuary water. *Applied and Environmental Microbiology* **79**: 4684-4693.

Byappanahalli MN, Whitman RL, Shively DA, Sadowsky MJ & Ishii S (2006) Population structure, persistence, and seasonality of autochthonous *Escherichia coli* in temperate, coastal forest soil from a Great Lakes watershed. *Environmental Microbiology* **8**: 504-513.

Callaghan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA & Holmes SP (2016) DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* **13**: 581-583.

Clermont O, Christenson J, Denamur E & Gordon D (2013) The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environmental Microbiology Reports* **5**: 58-65.

Cookson AL, Biggs PJ, Marshall JC, Devane M & Stott R (2017) Faecal source tracking and the identification of naturalised *Escherichia coli* to assist with establishing water quality and faecal contamination levels. Our Land & Water National Science Challenge.

Cookson AL, Biggs P, Marshall JC, Reynolds A, Collis RM, French NP & Brightwell G (2017) Culture independent analysis using *gnd* as a target gene to assess *Escherichia coli* diversity and community structure. *Scientific Reports* **7**: 841.

Cookson AL, Lacher DW, Scheutz F, Wilkinson DA, Biggs PJ, Marshall J & Brightwell G (2019) gndDb, a database of partial gnd sequences to assist with analysis of *Escherichia coli* communities using high-throughput sequencing. *Microbiology Resource Announcements* **8**: e00476-00419.

Devane M & Gilpin BJ (2019) Analysis of environmental water and sediment samples for the presence of naturalised *Escherichia* including *E. coli*. ESR Ltd.

Devane ML (2019) Transfer of knowledge and implementation of routine analysis of environmental samples for the presence of naturalised *E. coli*. Institute of Environmental Science and Research Ltd. .

Devane ML, Moriarty E, Weaver L, Cookson AL & Gilpin B (2020) Fecal indicator bacteria from environmental sources; strategies for identification to improve water quality monitoring. *Water Research* **185**: 116204.

Leonard M, Gilpin B, Horn B*, et al.* (2020) Quantitative Microbial Risk Assessment Pilot Study. (Ministry for the Environment) Porirua, New Zealand.

Liu S, Jin D, Lan R, Wang Y, Meng Q, Dai H, Lu S, Hu S & Xu J (2015) *Escherichia marmotae* sp. nov., isolated from faeces of *Marmota himalayana*. *International Journal of Systematic and Evolutionary Microbiology* **65**: 2130-2134.

Tenaillon O, Skurnik D, Picard B & Denamur E (2010) The population genetics of *Escherichia coli*. *Nature Reviews Microbiology* **8**: 207-217.

Touchon M, Perrin A, de Sousa JAM, Vangchhia B, Burn S, O'Brien CL, Denamur E, Gordon D & Rocha EPC (2020) Phylogenetic background and habitat drive the genetic diversification of *Escherichia coli*. *PLOS Genetics* **16**: e1008866.

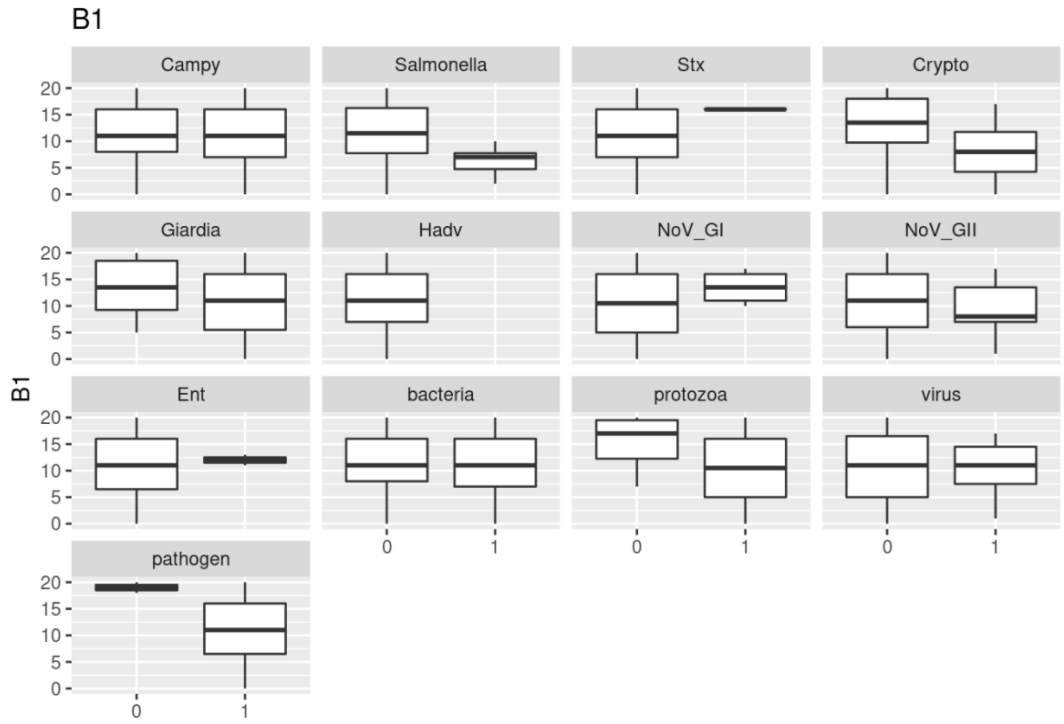
Vadnov M, Barbič D, Žgur-Bertok D & Erjavec MS (2017) *Escherichia coli* isolated from feces of brown bears (*Ursus arctos*) have a lower prevalence of human extraintestinal pathogenic *E. coli* virulence-associated genes. *Canadian Journal of Veterinary Research* **81**: 59-63.

van der Putten BCL, Matamoros S, Mende DR, Scholl ER, consortium C & Schultsz C (2021) *Escherichia ruysiae* sp. nov., a novel Gram-stain-negative bacterium, isolated from a faecal sample of an international traveller. *International Journal of Systematic and Evolutionary Microbiology* **71**: 004609.

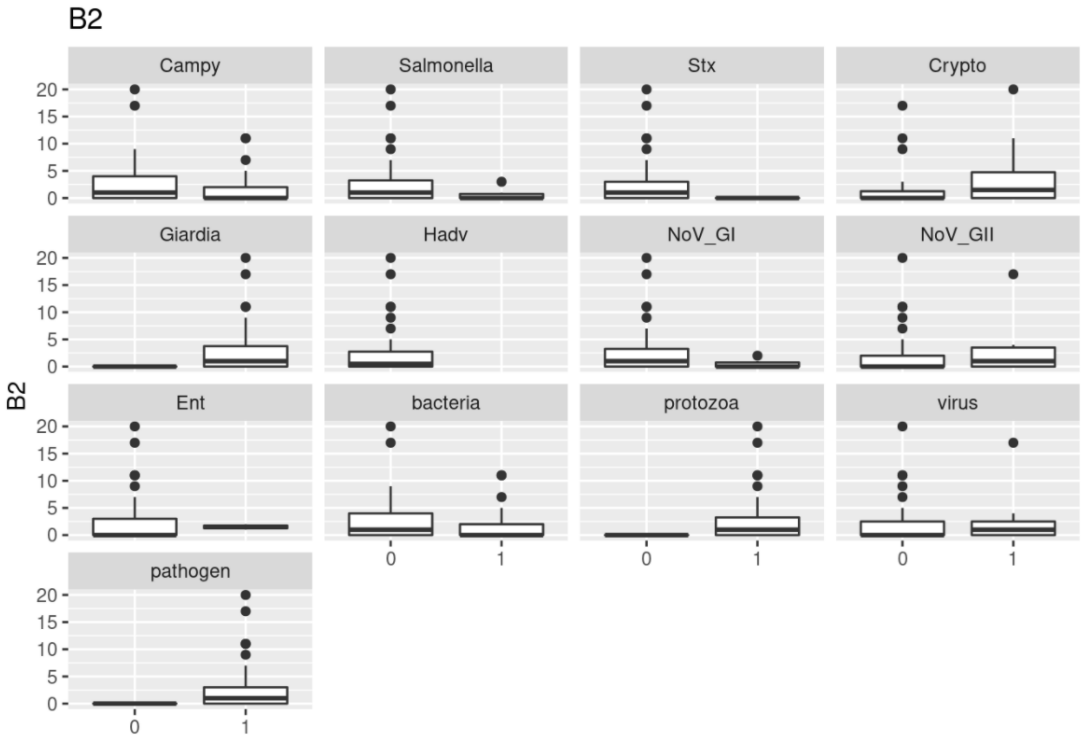
Walk S, Alm E, Gordon D, Ram J, Toranzos G, Tiedje J & Whittam T (2009) Cryptic lineages of the genus *Escherichia*. *Applied and Environmental Microbiology* **75**: 6534-6544.

8. Appendices

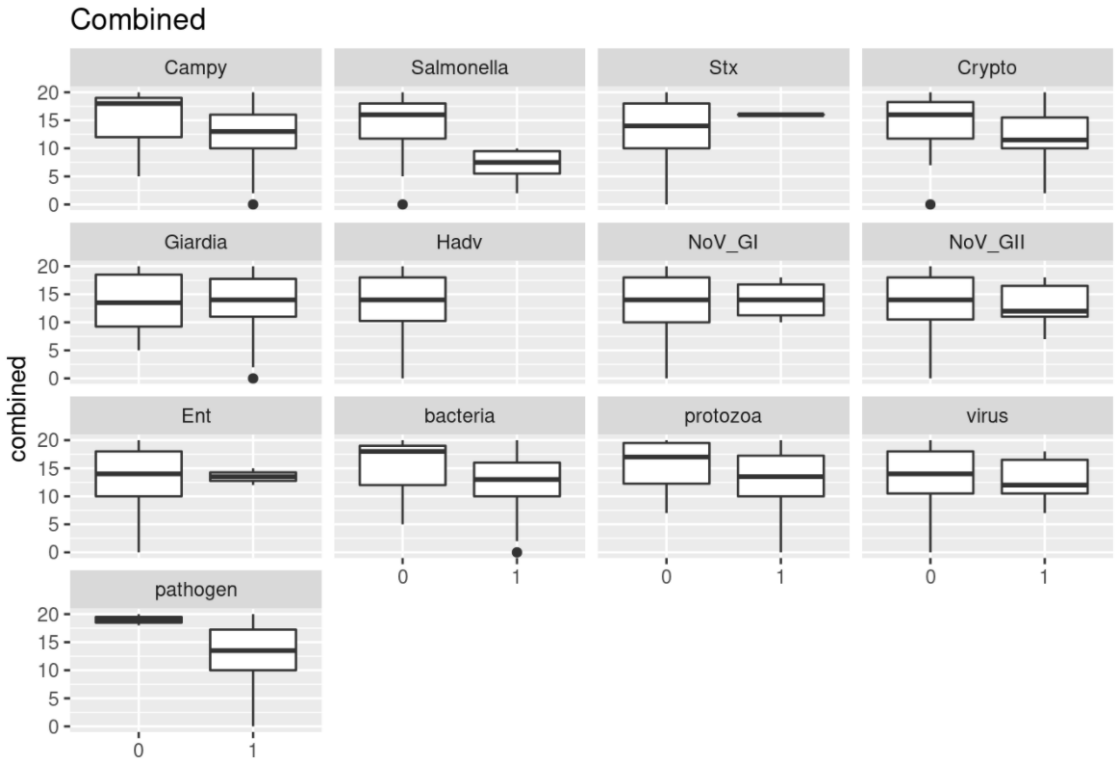
Appendix 1: Box plots were used to examine the relationship between *E. coli* phylogroup B1 and pathogens. Zero (0) implies that the pathogen was absent from the water sample and one (1) indicates that it was present.



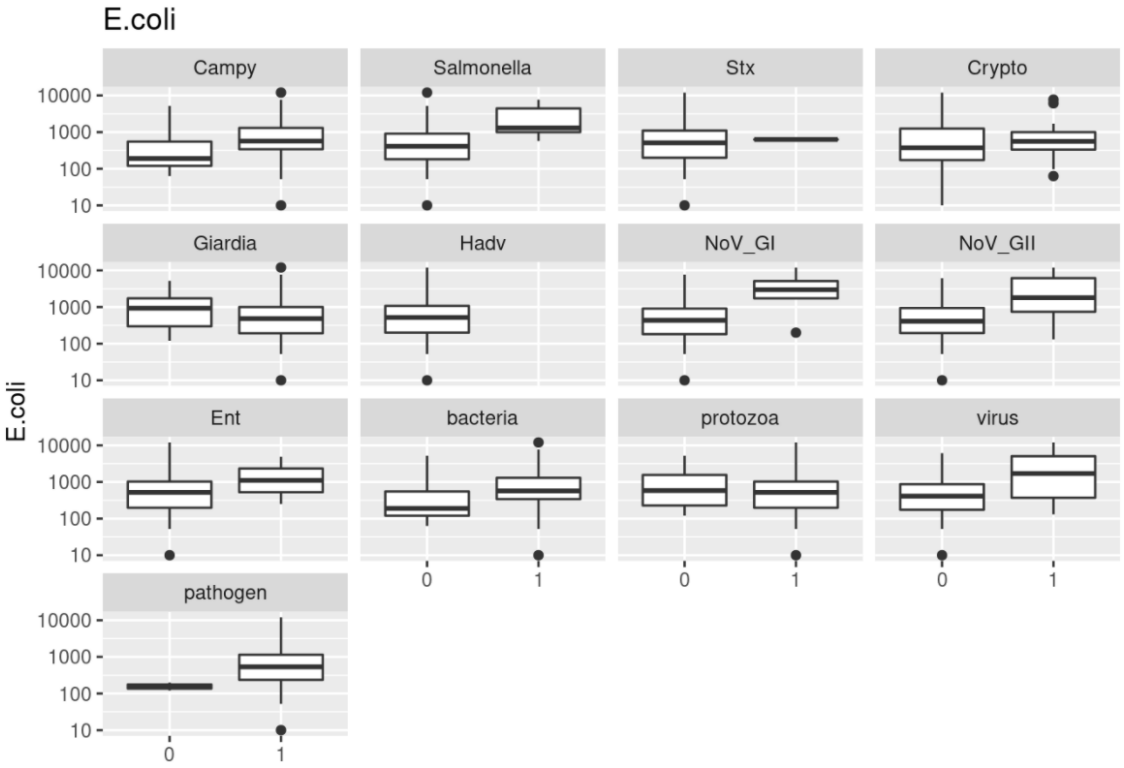
Appendix 2: Box plots were used to examine the relationship between *E. coli* phylogroup B2 and pathogens. Zero (0) implies that the pathogen was absent from the water sample and one (1) indicates that it was present.



Appendix 3: Box plots were used to examine the relationship between *E. coli* phylogroup B1 + B2 and pathogens. Zero (0) implies that the pathogen was absent from the water sample and one (1) indicates that it was present.



Appendix 4: Box plots were used to examine the relationship between generic *E. coli* (MPN per 100ml) and pathogens. Zero (0) implies that the pathogen was absent from the water sample and one (1) indicates that it was present.



Appendix 5: Logistic regression of *E. coli* phylogroup B2 and presence/absence of pathogens determined in QMRA Pilot Study (2020)

Microbe	$beta0$	$beta1$	P sig
Campylobacter	1.085	-0.100	0.167
Salmonella	-1.407	-0.284	0.333
Shiga toxin-producing *E. coli*	-2.996	-16.576	0.997
Cryptosporidium	-0.462	0.067	0.346
Giardia	0.486	17.912	0.994
Human adenovirus	-26.566	0.000	1.000
Norovirus GI	-1.335	-0.398	0.302
Norovirus GII	-1.750	0.048	0.544
Enterovirus	-2.838	-0.079	0.736
Bacteria	1.085	-0.100	0.167
Protozoa	0.916	17.531	0.994
Virus	-1.043	0.003	0.972
Pathogen	2.251	17.296	0.997