Novel discriminatory tests for *E. coli* to improve water quality assessments

Summary report – Year 2: Commentary on the impacts/significance of research findings re. New Zealand's current reliance on *E. coli* as indicators of pathogenic infection risk

Adrian Cookson

September 2021

Report for Dairy New Zealand Limited

Contract Number: A25871

Output ID Number: 11527

agresearch
āta mātai, mātai whetū

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, Massey University, 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.

This project brings together scientific researchers, Regional Council staff, the dairy industry, iwi and local community groups to undertake sampling to discover whether naturalised *E. coli*-like bacteria are more common in undisturbed predator-free mainland island sites, and whether dilution with faecal *E. coli* occurs as waterways pass through farmland and urban point source wastewater sites. Potentially, the naturalised *E. coli*-like bacteria could be biological markers of ecological health, so we will isolate them from our different sampling sites and use their DNA sequence information and specific growth activities to develop tests which permit their identification from environmental samples. These tests will assist Regional Councils, the pastoral industry, and other similar regulatory bodies overseas, to undertake more accurate water quality assessments and agricultural mitigation strategies.

This report is a summary of Year 2 of the MBIE-SI 'Novel discriminatory tests for *E. coli* to improve water quality assessments'. It includes some brief commentary of the current research landscape, and New Zealand's current reliance on *E. coli* as indicators of pathogenic infection risk. We also describe continued efforts to close bacterial genomes by combining long and short read sequence data to generate hybrid assemblies.

We also include brief summaries of two Our Land and Water National Science Challenge (OLW-NSC) projects that align with this work; 'Faecal source tracking and the identification of naturalised *Escherichia coli* to assist with establishing water quality and faecal contamination levels' and 'Faecal source tracking to understand the role of introduced predators and avian species on water quality assessments in the Mākirikiri Reserve, Dannevirke'.

2. Background

2.1 Background 1

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 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 thought to be able to survive and grow in the environment (Berthe et al 2013, Byappanahalli et al 2006, Walk et al 2009) where they may be found in >90% of water samples (Cookson et al 2017b), and comprising up to 48% of total '*E. coli*' (Devane and 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*.

New Zealand habitats, intensively controlled to limit predation by introduced mammalian pests and protect endemic wildlife, offer a unique opportunity to study environmental *E. coli* in relatively untouched settings. These environments offer suitable sites from which baseline levels of *E. coli* can be measured to provide information on the distribution and prevalence of both cryptic *Escherichia* clades and faecal *E. coli*. Our previous work has utilised high throughput metabarcoded amplicon sequencing methods targeting the hypervariable *gnd* gene (Barcak and Wolf 1988, Nelson and Selander 1994) to provide high-resolution data of *E. coli* population variation from complex samples (Cookson et al 2017a). The use of such methods in this programme will allow the differentiation of faecal *E. coli* and cryptic *Escherichia* clades and allow us to establish community diversity of *Escherichia* populations from environmental samples including deposited faeces, water, soil, sediment, and periphyton (biofilm material attached to submerged surfaces), collected in native forest, and contrasting sites with associated agricultural land-use. We will also analyse whole genome sequence (WGS) data from cultured isolates to examine the spatio-temporal distribution of *E. coli* and to identify traits associated with environmental persistence of faecal *E. coli*/ cryptic *Escherichia* clades as potential targets for future genetic or phenotypic tests for their delineation.

2.1.1 Summary of Year 2: 1 October 2020 to 30 September 2021

NZ mammal-free environments (closed-canopy bush where waterfowl are absent) provide a unique opportunity to investigate microbiological water quality and the disconnect between *Escherichia* and human health. Thus, samples obtained from pest-free mainland islands will provide important baseline data on the prevalence of naturalised strains as a potential marker for ecological health and for comparison with *Escherichia* populations from agricultural/urban sites in the same catchment with contrasting land use.

The necessary permissions/concessions were obtained to sample at the following mainland islands/reserves representing sites of minimal anthropogenic impact

- Sanctuary Mountain Maungatautari (Waikato)
- Bushy Park Tarapuruhi (Whanganui)
- Pūkaha Mount Bruce (Tararua)
- Brook Waimārama Sanctuary (Nelson)
- Orokonui Ecosanctuary (Otago)

Co-ordinated sampling at parallel State of the Environment sites representing those impacted by pastoral farming or urban activity has occurred with

- Waikato Regional Council (Maungatautari)
- Horizons Regional Council (Pūkaha Mount Bruce and Bushy Park Tarapuruhi)
- Nelson City Council (Brook Waimārama)

Our sample analysis strategy employs both culture-dependent (microbiological media) and culture-independent (molecular analysis of environmental DNA extractions) techniques for the identification of faecal and environmental *Escherichia* species. Methods have successfully been developed, implemented and validated for the collection, DNA extraction and isolation of *E. coli* from

- Water
- Soil
- Sediment
- Periphyton
- Faeces

To 30 September 2021, all thirty of the anticipated sampling visits were undertaken encompassing a mainland island/fenced reserve site and a corresponding site impacted by pastoral farming.

- 414 environmental samples (water, soil, sediment, periphyton and faeces) were obtained
- 1351 isolates were stored for subsequent analysis, including 1048 *E. coli*, 289 *Escherichia marmotae*, 10 *Escherichia ruysiae* and 4 others (3 *Hafnia* and 1 *Serratia*).
- 1342 isolates were subtyped by targeting the hypervariable gene *gnd*, and all 1351 using *E. coli* phylotyping
- 299 environmental *Escherichia* have been identified; faeces (188, of which 155 from avian faeces), water (34), periphyton (20), sediment (26), and soil (31)
- Summary data sheets have been prepared which will be provided to stakeholders associated with each sample site, where data for the bush reserve and downstream agricultural/urban site are summarised

A constructed wetland intercepting tile drain flows from dairy cattle grazed pastures in the Waikato (Toenepi catchment) has been chosen as a site to investigate the survival and movement of faecal indicator bacteria. Here historical data indicates an increase of *E. coli* concentration through the wetland. i.e., more *E. coli* are measured exiting the wetland compared to numbers entering. Our hypothesis is that the increased numbers of *E. coli* exiting the wetland does not represent a human

health risk as it is attributable to the growth of non-faecal *E. coli*/ *Escherichia* species. Therefore, wetland sampling methods have been developed to measure

- hydraulic retention time (using a rhodamine tracer)
- survival of faecal and environmental *E. coli*, and cryptic *Escherichia* clade V (in-situ field mesocosms)
- inoculation and measurement (using real-time PCR methods) of the same bacterial isolates into the wetland to evaluate whether increased concentrations throughout the wetland are due to the growth of non-faecal environmental strains indicating that the human health risk is not increased.

Three *Escherichia* isolates originally isolated from the wetland have been selected for inoculation. One *E. coli* is an isolate from bovine faeces, another *E. coli* was isolated from wetland sediment, and the last is an environmental *Escherichia marmotae* (clade V) isolate. The development and validation of real time PCR methods to determine the sensitivity and specificity of probe/primer combinations for each isolate are currently in progress. Sampling of the wetland has already been undertaken in a 'dry' state (31 May 2021) before autumn/winter rains occurred and further sampling will occur in October 2021 now that heavy rainfall events have provided sufficient water flow through the wetland.

3. Material and Methods

3.1 Long-read minION sequencing summary

The genetic analysis of *E. coli* using whole genome sequencing forms an integral part of this project for the identification of traits that may be important in the environmental persistence of some strains. Our current sequencing methods generate incomplete genomes often in several hundred fragments due to the presence of multi-copy genes or mobile genetic elements such as plasmids and transposons. However, the use of Oxford Nanopore sequencing technology (https://nanoporetech.com/products/minion) allows real-time sequencing of long-read lengths to generate closed genomes. Closed genomes provide greater resolution of missing genetic micro-diversity which may be functionally important and thus, the identification of traits in environmental *Escherichia* clades that may prove discriminatory in differentiating them from faecal *E. coli*.

The current programme activities are concentrated on the isolation of faecal *E. coli* and environmental *Escherichia* clades from samples sites of contrasting anthropogenic impact, i.e., native bush remnants and sites impacted by urban storm/wastewater or pastoral farming. Whole genome sequencing has commenced but for the purposes of method development, four *Escherichia* clade V isolates (Table 1) obtained as part of a previous Our Land & Water National Science Challenge project were chosen as targets for preliminary long-read minION sequencing (Oxford Nanopore) and the generation of closed genomes (described in the 2020 Year 1 summary report). Short read sequencing of the four isolates was performed in 2018 using the Illumina MiSeq platform, and assembled using SKESA (Souvorov et al 2018) to provide draft genomes of between 4.58Mb and 5.11Mb .

3.1.1 DNA extraction and sequencing summary

Approximately 8 colonies of each isolate were removed from a sheep blood agar plate, mixed in 350µl TE (10mM Tris-HCl containing 1mM EDTA) buffer, and lysed using 50µl lysozyme (10mg/ml), 20µl proteinase K (20mg/ml) and 50µl 10 x SDS. Lysed cellular material was precipitated by adding 160µl protein precipitation solution from the Promega Wizard® Genomic DNA Purification kit and removed by centrifugation. DNA was precipitated using isopropanol, washed with 70% ethanol and after briefly drying, the DNA pellet was resuspended in 10mM Tris-HCl. The concentration and purity of purified DNA was determined using a NanoDrop spectrophotometer.

Separate genomic DNA libraries of each of the *Escherichia* isolates were generated using the Rapid Barcoding Sequencing kit (Oxford Nanopore, SQK-RBK004) to allow for multiplex sequencing according to Oxford Nanopore protocols. Libraries were pooled and loaded onto the flow cell and sequenced with 'offline' base calling carried out using GUPPY (Wick et al 2019).

The assembly pipeline Unicycler (Wick et al 2017) was used to generate bacterial genomes using both short reads (MiSeq data) where available, and long-reads (minION) to generate a hybrid *de novo* assemblies. Reads were first demultiplexed (sequences separated according to individual genomic DNA preparations/barcodes) before running a Snakemake script incorporating the Unicycler pipeline, to generate the hybrid genome assembly. *De novo* genome assemblies (GFA format) were visualised using Bandage, a Bioinformatics Application for Navigating De novo Assembly Graphs Easily (Wick et al 2015). Where only long-read sequences were available the

Flye assembler (version 2.9) was used (Kolmogorov et al 2019). *De novo* genome assemblies (GFA format) were visualised using Bandage, a Bioinformatics Application for Navigating De novo Assembly Graphs Easily (Wick et al 2015).

We will undertake short-read Illumina whole genome sequencing of approximately 500 isolates from this study to be completed April 2022. A total of 81 isolates, including 25 *E. marmotae* and 1 *E. ruysiae*, underwent short-read Illumina sequencing in July 2021. Isolates will be chosen for WGS analysis according to identification of *gnd* sequence types (gSTs) common to different samples, sites, or site visits, those from animal faecal samples for faecal source tracking purposes, and environmental *Escherichia* species. For our initial method development, four *Escherichia marmotae* (clade V) isolates (AGR4045, AGR4162, AGR4167 and AGR4200) obtained as part of a previous Our Land & Water National Science Challenge project were chosen as targets for preliminary long-read minION sequencing (Oxford Nanopore Technology) and the generation of closed genomes. A further 25 *Escherichia* isolates (Table 1) were analysed using minION sequencing, 13 to generate assemblies using hybrid methods and 12 using long-read sequencing only (Table 2), with generation of short-read data in progress with subsequent hybrid assemblies.

Table 1: Sample metadata associated with *Escherichia* species included in this study.

ID	*Site*	*Source*	*Phylotype*	*gST*	*Species*
AGR4045	Makirikiri Stream	soil	clade V	gST546	*E. marmotae*
AGR4111	Toenepi wetland	water	clade IV	gST549	*E. ruysiae*
AGR4162	Makirikiri Stream	sediment	clade V	gST536	*E. marmotae*
AGR4167	Mangatera River	sediment	clade V	gST537	*E. marmotae*
AGR4200	Toenepi wetland	water	clade V	gST539	*E. marmotae*
AGR4316	Pukaha Mount Bruce	faeces	D	gST504	*E. coli*
AGR4587	Tapuata Stream Stream	water	B2	gST252	*E. coli*
AGR4608	Bushy Park	water	clade V	gST546 97.5%	*E. marmotae*
AGR4741	Brook Waimarama	biofilm	clade V	gST587	*E. marmotae*
AGR4791	Orokonui Ecosanctuary	water	clade V	gST543	*E. marmotae*
AGR4793	Orokonui Ecosanctuary	biofilm	clade V	gST538	*E. marmotae*
AGR4797	Orokonui Ecosanctuary	faeces	clade V	gST539	*E. marmotae*
AGR4801	Orokonui Ecosanctuary	faeces	B2	gST258	*E. coli*
AGR4808	Orokonui Ecosanctuary	faeces	clade IV	gST592	*E. ruysiae*
AGR4810	Orokonui Ecosanctuary	sediment	clade V	gST542	*E. marmotae*
AGR4814	Orokonui Ecosanctuary	faeces	clade V	gST546	*E. marmotae*
AGR4816	Orokonui Ecosanctuary	faeces	clade V	gST545	*E. marmotae*
AGR4818	Orokonui Creek	water	clade V	gST547	*E. marmotae*
AGR4848	Pukaha Mount Bruce	biofilm	B2	gST258	*E. coli*
AGR4886	Bushy Park	water	clade V	gST593	*E. marmotae*
AGR4974	Pukaha Mount Bruce	faeces	clade V	gST548	*E. marmotae*
AGR4990	Makakahi River	biofilm	B1	gST130	*E. coli*
AGR5038	Brook Stream, Manuka St	biofilm	clade V	gST594	*E. marmotae*
AGR5076	Orokonui Creek	water	B1	gST041	*E. coli*
AGR5151	Tapuata Stream	water	B2	gST252	*E. coli*
AGR5548	Orokonui Creek	sediment	B1	gST041	*E. coli*
AGR5733	Pukaha Mount Bruce	water	D	gST010	*E. coli*
AGR6128	Tapuata Stream	water	B2	gST252	*E. coli*
AGR6137	Tapuata Stream	water	B2	gST266	*E. coli*

4. Results and Discussion

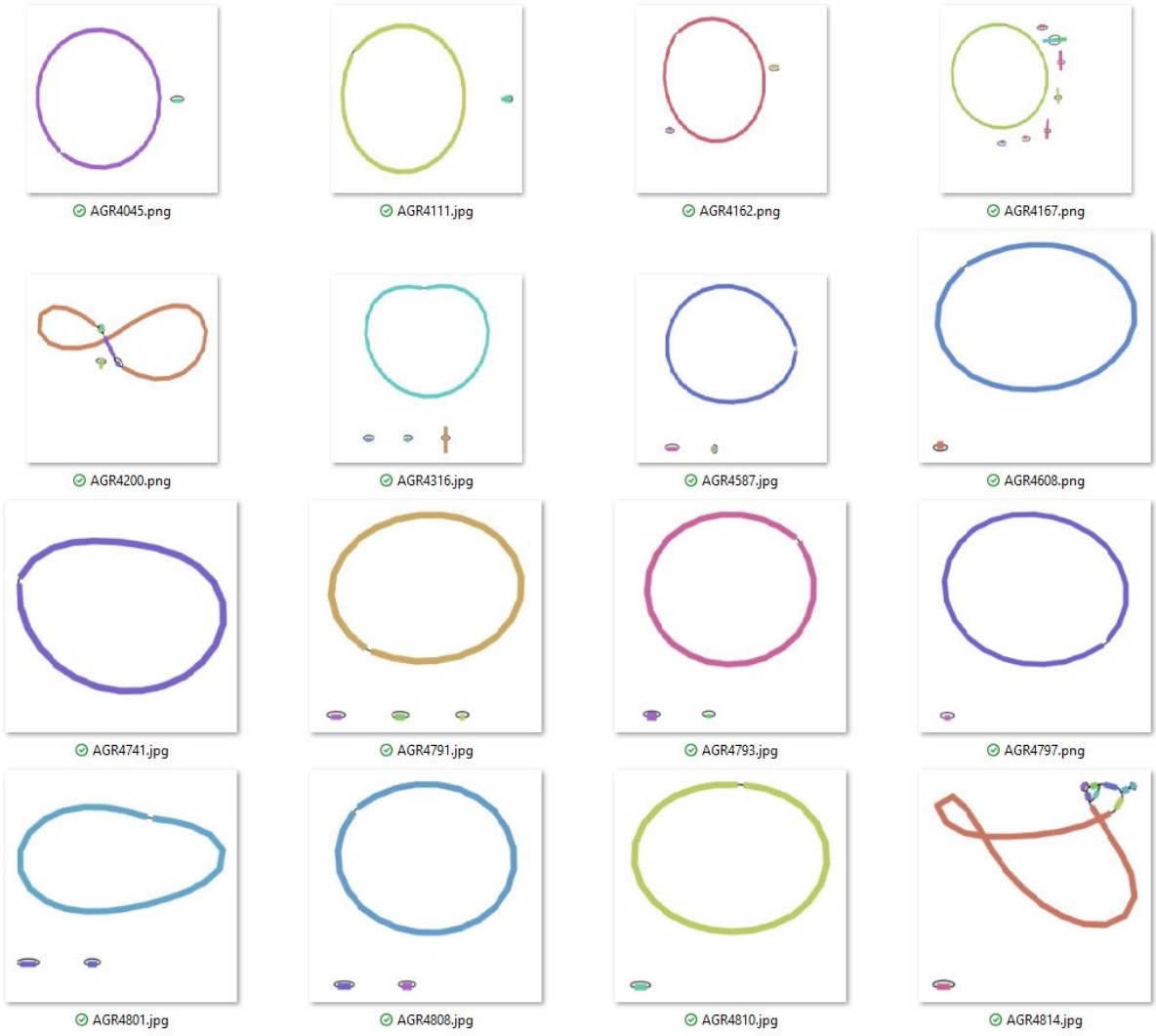
In total, DNA from 25 *Escherichia* isolates has been sequenced, 13 with both long and short-read platforms to generate assemblies using hybrid methods and 12 using minION long-read data only (Table 2). Short-read Illumina analysis is currently underway with DNA extracts from these 12 isolates to polish the long-read assemblies. Short read sequencing of AGR4045, AGR4162, AGR4167 and AGR4200 was performed in 2018 using the Illumina MiSeq platform, and assembled using SKESA (Souvorov et al 2018) to provide draft genomes. Preliminary long-read sequencing provided data that improved genome assemblies, but the genomes were not closed (see Year 1 2020 summary report). Subsequent optimisation of the methods was undertaken, and three of the four draft genomes were closed using hybrid assemblies of short- and long-read sequencing as outlined in the sequencing summary for *de novo* assemblies (Table 2).

Table 2: *De novo* genome assemblies metadata of *Escherichia* species included in this study.

ID	*Hybrid assembly*	*No. of contigs*	*Genome closed*	*Genome size (bp)*	*Contig max length (bp)*	*plasmid length (bp)*
AGR4045	Y	2	Y	4,726,397	4,626,179	100,218
AGR4111	Y	2	Y	4,719,904	4,529,317	194,450
AGR4162	Y	3	Y	4,892,461	4,745,911	87,795, 58,755
AGR4167	Y	8	N	5,279,296	5,008,151	116,793, 70,471, 63,988
AGR4200	Y	5	N	4,630,648	4,404,128	179,817, 46,457
AGR4316	Y	4	N	5,199,453	5,067,977	88,962, 36,415
AGR4587	Y	4	Y	5,228,112	5,076,384	145,705
AGR4608	N	2	Y	5,210,874	5,151,802	59,072
AGR4741	Y	1	Y	4,579,446	4,579,446	none
AGR4791	Y	4	Y	4,776,898	4,573,181	86,662, 84,475, 32,583
AGR4793	Y	3	Y	4,959,455	4,828,563	92,134, 38,760
AGR4797	Y	2	Y	4,613,934	4,565,022	48,912
AGR4801	Y	3	Y	4,993,445	4,747,681	150,082, 95,684
AGR4808	N	3	Y	4,823,754	4,623,726	114,876, 85,154
AGR4810	N	2	Y	5,000,391	4,890,917	109,475
AGR4814	Y	7	N	4,817,229	4,510,978	111,759, 100,176, 92,311
AGR4816	N	4	Y	5,062,678	4,849,683	144,417
AGR4818	N	4	N	5,123,822	4,685,621	232,839, 124,622, 80,743
AGR4848	Y	1	N	5,190,651	5,190,651	none
AGR4886	N	1	Y	4,660,634	4,660,634	none
AGR4974	N	2	Y	4,769,367	4,677,827	91,541
AGR4990	N	2	Y	4,825,910	4,726,947	98,963
AGR5038	N	1	Y	4,589,114	4,589,114	none
AGR5076	N	3	Y	4,982,932	4,800,098	145,787, 37,047
AGR5151	Y	3	Y	5,027,195	4,855,042	137,735, 34,418
AGR5548	N	2	Y	5,132,795	5,057,694	75,101
AGR5733	N	2	N	5,178,899	5,086,712	92,187
AGR6128	Y	11	N	4,998,316	4,761,444	122,672, 95,638
AGR6137	y	4	Y	5,425,109	4,911,512	269,521, 158,294, 86,294

4.1 Bandage plots

Data from Bandage plots indicates that at least 21 of the genomes have been closed, of which three (AGR4741, AGR4848, and AGR5038) occur as a single chromosome only. The remaining 18 genomes contain up to three plasmids ranging in size from 32,583 to 269,521bp.



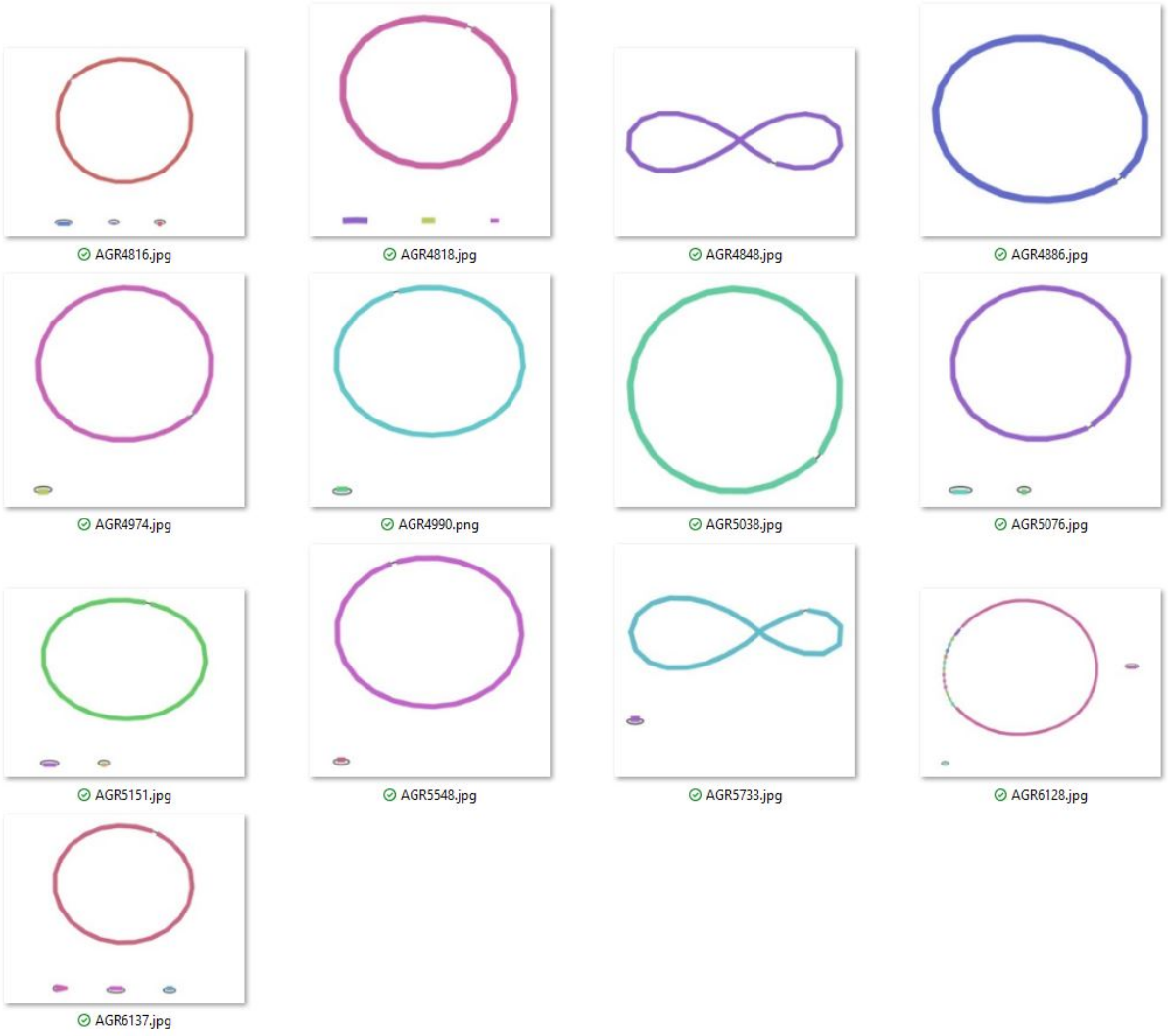


Figure 1: *Bandage plots of de novo genome assemblies of Escherichia species included in this study. A circular chromosome and plasmids associated with each assembly are indicative of closed genomes.*

5. Future Work

During the last year of the 3-year programme, we will undertake high resolution analysis of closed genomes to understand the phylogenetic relationship between mobile genetic elements such as plasmids, prophage and insertion sequences to determine their role in the divergent evolution of *E. coli*, and newly described *E. marmotae* and *E. ruysiae* species. For the wider project, we will undertake the *in-situ* mesocosm analysis in the Toenepi wetland to provide data on the contrasting fitness and persistence of *E. coli* and *Escherichia* species during prolonged incubation. These resilience data will be utilised to develop models that fit contrasting persistence of the naturalised and faecal strains respectively. The on-going development of rapid methods to differentiate the *E. coli* and *Escherichia* species will be applied to wetland samples to provide insight into any apparent dilution of environmental *Escherichia* species by faecal *E. coli* associated with contrasting land-use.

6. Acknowledgements

The Smart Idea that is underpinning this work is funded by the Ministry of Business, Innovation and Employment. We are grateful to our collaborators from Massey University, National Institute of Water and Atmospheric Research, and the Institute of Environmental Science and Research for their contributions to the overall programme. We are also appreciative of the staff that have allowed access to reserves for sample collection, and Regional/City Council staff that are providing additional water quality assessment data. Special thanks go to Dr Marie Moinet for leading much of the lab analyses. Finally, we acknowledge the expertise of Dr Sara Burgess (Massey University) in the leadership, development and training of staff on the use of the minION sequencing technology.

7. Appendices

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

OLW-NSC collaboration between AgResearch and ESR.

Recent work has identified *E. coli*-like bacteria that are not able to be distinguished from faecal *E. coli* using the standard water quality monitoring tests. These 'naturalised' *E. coli*-like bacteria grow and multiply in soil, water and sediment, but are rarely found in faeces. 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.

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.

‘Faecal source tracking to understand the role of introduced predators and avian species on water quality assessments’

OLW-NSC collaboration between AgResearch and Massey University

Contrary to livestock and human activities, wildlife has not thoroughly been studied as a source of faecal contamination in waterways but could contribute too. Very few studies have looked at *E. coli* present in NZ wildlife (Devane et al 2019), and none have undertaken detailed characterization of the different *E. coli* strains found (Moriarty et al 2011, Murphy et al 2005, Phiri 2015, Sumner et al 1977). Genetic markers were developed to detect gull, Canada goose, and duck faecal contamination in water (Green et al 2012) but data is lacking on invasive mammals, that can be present at high densities and could also contribute to misinterpreting water quality monitoring test results.

The aims of this study were to target faecal material from bird and introduced predator species for the examination of *E. coli* populations in these faecal specimens and to determine whether avian and/or invasive mammal species contribute faecal bacteria, such as *E. coli* and *E. coli*-like naturalised *Escherichia* species, that impact water quality assessments.

We examined the profile of *E. coli* populations in gut contents and faeces from bird and introduced predator species in the Mākirikiri Reserve, Dannevirke, and compared to water, soil, sediment and biofilm samples taken within the reserve to determine whether avian and/or invasive mammal species contribute faecal bacteria in the environment.

E. coli (n=420) were recovered from animal and environmental samples (n=106). Initial characterisation of *E. coli* was using real time PCR targeting the *uid*A gene, and a subset were further typed by sequencing a region of the hypervariable *gnd* gene to generate a specific *gnd* sequence type (gST). These data informed which isolates underwent further phylogenetic analysis using whole genome sequencing (n=100). *E. coli* populations from sample enrichments were analysed using metabarcoding and *gnd* amplicon sequencing.

Analyses showed a significantly lower α-diversity of *Escherichia* gSTs in animal compared with environmental samples, and that some gSTs were present in both sample types, *e.g.,* gST535 (85% of samples) and gST258 (71%). Core-genome analysis showed limited variation between several animal and environmental isolates (<10 SNPs). Cryptic *Escherichia* species, phenotypically similar to *E. coli*, were isolated and detected by metabarcoding, especially in birds, but only in low abundance.

Our data show at an unprecedented scale that *E. coli* and cryptic *Escherichia* clones are shared between wildlife, water and the wider environment. Phylogenetic analysis of isolates and profiling of *Escherichia* populations can provide further useful information on the source(s) of faecal contamination to assist with management decisions where microbial water quality is compromised.

8. References

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.

Barcak GJ, Wolf REJ (1988). Comparative nucleotide sequence analysis of growth-rate-regulated *gnd* alleles from natural isolates of *Escherichia coli* and from *Salmonella typhimurium* LT-2. *Journal of Bacteriology* **170:** 372-379.

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.

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

Cookson AL, Biggs PJ, Marshall JC, Devane M, Stott R (2017b). 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.

Devane M, Gilpin BJ (2019). Analysis of environmental water and sediment samples for the presence of naturalised *Escherichia* including *E. coli*. Prepared for Northland Regional Council edn. ESR Ltd.

Devane ML, Gilpin B, Moriarty E (2019). The sources of “natural” microorganisms in streams. The Institute of Environmental Science and Research Ltd.: Christchurch, New Zealand.

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

Green HC, Dick LK, Gilpin B, Samadpour M, Field KG (2012). Genetic markers for rapid PCR-based identification of gull, Canada goose, duck, and chicken fecal contamination in water. *Applied and Environmental Microbiology* **78:** 503-510.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* **37:** 540-546.

Leonard M, Gilpin B, Horn B, Coxon S, Armstrong B, Hewitt J *et al* (2020). Quantitative Microbial Risk Assessment Pilot Study. In: Environment Mft (ed): Porirua, New Zealand.

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

Moriarty EM, Karki N, MacKenzie M, Sinton LW, Wood DR, Gilpin BJ (2011). Faecal indicators and pathogens in selected New Zealand waterfowl. *New Zealand Journal of Marine and Freshwater Research* **45:** 679-688.

Murphy J, Devane ML, Robson B, Gilpin BJ (2005). Genotypic characterization of bacteria cultured from duck faeces. *Journal of Applied Microbiology* **99:** 301-309.

Nelson K, Selander R (1994). Intergeneric transfer and recombination of the 6-phosphogluconate dehydrogenase gene (*gnd*) in enteric bacteria. *Proceedings of the National Academy of Sciences of the United States of America* **91:** 10227-10231.

Phiri BJ (2015). Estimating the public health risk associated with drinking water in New Zealand, Massey University.

Souvorov A, Agarwala R, Lipman DJ (2018). SKESA: strategic k-mer extension for scrupulous assemblies. *Genome Biology* **19:** 153.

Sumner JL, Perry IR, Reay CA (1977). Microbiology of New Zealand feral venison. *Journal of the Science of Food and Agriculture* **28:** 829-832.

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 *et al* (2020). Phylogenetic background and habitat drive the genetic diversification of *Escherichia coli*. *PLOS Genetics* **16:** e1008866.

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 *et al* (2009). Cryptic lineages of the genus *Escherichia*. *Applied and Environmental Microbiology* **75:** 6534-6544.

Wick RR, Schultz MB, Zobel J, Holt KE (2015). Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31:** 3350-3352.

Wick RR, Judd LM, Gorrie CL, Holt KE (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology* **13:** e1005595.

Wick RR, Judd LM, Holt KE (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology* **20:** 129.