METAGENOMICS AND WHOLE GENOME SEQUENCING TO UNDERSTAND DRINKING WATER QUALITY

Brent Gilpin, William Taylor, Iain Haysom, Kirstin Thom, Susan Lin, Paula Scholes, Beth Robson, Paula Scholes, Belinda Cridge and Megan Devane
Institute of Environmental Science & Research Limited
Christchurch

ABSTRACT

Regulatory responsibility in Aotearoa for drinking water has shifted from the Ministry of Health to Taumata Arowai whose expanded remit will encompass an estimated 75,000 drinking water supplies. Traditionally the only tool used by water suppliers for monitoring microbial contamination was testing for the indicator bacteria *E. coli*. Taumata Arowai is developing new drinking water quality standards and rules, with increased testing requirements in the source water, at the treatment plant and within the reticulation network. This includes adding another indicator bacteria - Total Coliforms. The key issue is that while detection of *E. coli* and Total Coliforms suggests there is a problem, it provides little guidance as to the source of that problem, or what to do about it. This has led to inaction by some water suppliers, and default boil water notices by others. Taumati Arowai is emphasising a "Find & Fix" approach, which while increased testing of more supplies will certainly find more issues, fixing those requires new science and new understanding.

Technological advances in DNA sequencing have now reached the point of maturity that they can be applied to better understanding and management of drinking water. Whole genome sequencing of isolates, and microbial community analysis using metagenomic approaches provide new tools for evaluating drinking water. There are two key applications:

1) Understanding source water (rivers, lakes and ground water aquifers) and contamination pathways. Metagenomic approaches provide opportunities to understand the sources affecting water supplies and the microbial community present including non-faecal pathogens such as Legionella.

2) Investigating the causes of detections of Total Coliforms and *E. coli* within a drinking water supply. Is the detection due to contamination events affecting the source water, treatment failure, network contamination (breaks in pipes or backflows, storage tank issues), network biofilm sloughing, or in fact sampler or laboratory errors?

This paper will explore these revolutionary technologies with examples of their actual or theoretical application to the 2016 Havelock North campylobacteriosis outbreak ill.

KEYWORDS

Drinking water, public health, infectious disease, water quality

PRESENTER PROFILE

Brent is a Senior Science Leader in the Environmental Science team at ESR where he has an applied focus that interfaces between research activities, outbreak investigations, and commercial work. His key focus is the use of technology and science to better understand microbial contamination of water, food and our environment.

INTRODUCTION

Regulatory responsibility in Aotearoa for drinking water has shifted from the Ministry of Health to Taumata Arowai whose expanded remit will encompass an estimated 75,000 drinking water supplies. Traditionally the only tool used by water suppliers for monitoring microbial contamination was testing for the indicator bacteria *E. coli* with the 486 registered drinking water supplies undertaking about 90,000 tests annually. Taumata Arowai is developing new drinking water quality standards and rules, with increased testing requirements in the source water, at the treatment plant and within the reticulation network. This includes adding another indicator bacteria - Total Coliforms. The key issue is that while detection of *E. coli* and Total Coliforms suggests there is a problem, it provides little guidance as to the source of that problem, or what to do about it.

This has led to inaction by some water suppliers, and default boil water notices by others. Taumati Arowai is emphasising a "Find & Fix" approach, which while increased testing of more supplies will certainly find more issues, fixing those requires new science and new understanding.

Next generation sequencing, massively parallel sequencing, and second generation sequencing are all largely interchangeable terms that describe technologies able to sequence millions, if not billions of DNA sequences at the same time (van Dijk et al., 2018, Zhang et al., 2021). These have revolutionised many facets of biology, notably in sequencing of the human genome, sequencing of genomes of all other types of organisms, and in sequencing whole communities of eurkaryotes and prokaryotes in specific ecological niches. Third generation sequencing technologies take this further by increasing the length of sequences determined.

This paper will explore the application of two types of next generation sequencing.

1. The whole genome sequencing (WGS) of individual bacteria isolated from a water sample.
2. Metagenomic analysis or profiling of the community of organisms in a sample.

For drinking water these tools these tools have a number of applications including:

1) Whole genome sequencing can be use to:
    a. Evaluate whether *E. coli* found in a water sample is from a laboratory source
    b. Where pathogens such as *Campylobacter* are found in a water sample, evaluation of the potential source of those pathogens, and where a source is found confirmation that the sources are linked.
2) Metagenomic analysis can be used to:
    a. Better understand source waters (rivers, lakes and ground water aquifers) and contamination pathways. Metagenomic approaches provide opportunities to understand the sources affecting water supplies and the microbial community present including non-faecal pathogens such as Legionella.
    b. Investigating the causes of detections of Total Coliforms and *E. coli* within a drinking water supply. Is the detection due to contamination events affecting the source water, treatment failure, network contamination (breaks in pipes or backflows, storage tank issues), or network biofilm sloughing.

This paper will explore these revolutionary technologies with examples of their actual or theoretical application to the 2016 Havelock North campylobacteriosis outbreak.

DISCUSSION

WHOLE GENOME SEQUENCING

HAVELOCK NORTH 2016

In August 2016, an outbreak of campylobacteriosis occurred in Havelock North with more than 8,300 people estimated to have become ill (Gilpin et al., 2022). *Campylobacter jejuni* isolates were recovered from water samples collected from the Havelock North water supply on Friday 12th August 2016. Clearly the presence of pathogenic bacteria in drinking water is not acceptable. But having isolated these bacteria a number of questions were raised which whole genome sequencing was able to assist with.

> Question 1: Are these bacteria in the water were the same as those isolated from infected and sick people?
>
> Question 2: Where did these bacteria had come from? Initial potential sources including chicken manure, cattle, sheep and human sewage.

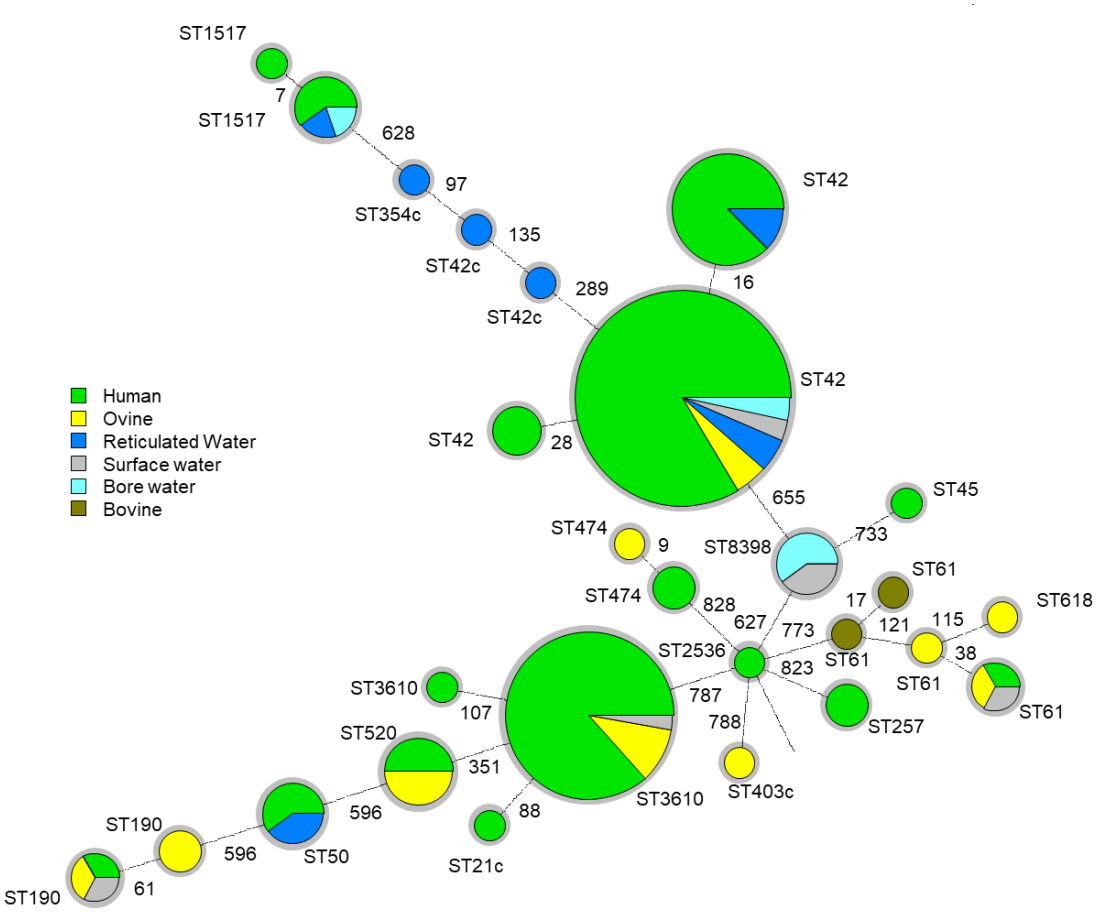
Whole genome sequencing was able to assist with both of these questions. The *Campylobacter jejuni* isolates were purified, and DNA extracted from each isolate. These were sequenced using an Illumina MiSeq and the 1.6 million base

pair genomes assembled. Core genome multi-locus sequence typing was undertaken to compare the isolates (Figure 1).

This was complex outbreak with more than a dozen different strains of *C. jejuni* found among the clinical isolates. But amongst these over 80% were either sequence types ST42 or ST3610.

Amongst the drinking water isolates there were eight distinct strains or types of *C. jejuni,* with four of these also found in clinical samples. including the largest cluster of isolates. Therefore, for question 1, we could confirm that isolates in the water were indistinguishable from clinical isolates, and therefore a likely cause. Not all the water isolates were found in clinical cases, and not all clinical isolates were found in water samples. There are several explanations for this. First, only a small proportion (<5%) of likely clinical cases were genome sequenced, so more sequencing would likely have identified more genotypes of *C. jejuni*. Conversely the earliest water samples obtained were only collected on the 12 August 2016. It is likely the water was contaminated for up to six days before this, so sampling on days prior to the 12 August, would very likely have recovered additional isolates from the water of different genotypes, which may have matched more of the clinical isolates.

Figure 1: Core genome (cgMLST) minimum spanning tree (up to 1,394 loci) of Campylobacter jejuni isolates. Numbers on the branches are the number of cgMLST differences. Size of circles reflects the number of isolates with that same cgMLST profile (0-3 cgMLST differences).



It should also be noted that some of the clinical cases will have been unrelated to the water source. Campylobacteriosis is a common illness, usually from foodborne sources. Therefore a few of the cases in Havelock North during the outbreak period are likely to be from the "normal" sources such as poultry. The clinical isolate labelled ST45 is a possible example of this.

There is one other possibility for bore water samples ST8398. This is a new genotype of *C. jejuni*, which is most closely related to isolates previously recovered from wildfowl such as pukeko. There were pukeko in the paddocks around the bores. These raise the intriguing possibility that these strains of *C. jejuni* may lack the ability to cause illness in people, even if consumed in drinking water.

The second question regarding the source of the isolates could be guided by several observations.

- First analysis of the drinking water samples using faecal source tracking qPCR methodology (Devane et al. 2020) identified ruminant DNA markers in all of the drinking water samples.
- The major campylobacter MLST genotypes were also those typically associated with ruminants such as cows and sheep.
- Even more conclusive was the isolation of campylobacter from likely animals in the vicinity of the bores. While there were only a few isolates from bovine sources, these were not the same as any of the water or clinical isolates. In contrast sheep faecal material sampled had five genotypes of campylobacter indistinguishable from the clinical isolates with one of the sheep isolates indistinguishable from the water isolates. Again, only limited sampling of sheep was possible so overlap in only some of the genotypes is understandable. It does reinforce the diversity of genotypes present and the need in any outbreak investigation to isolate and genotype as many isolates as possible.

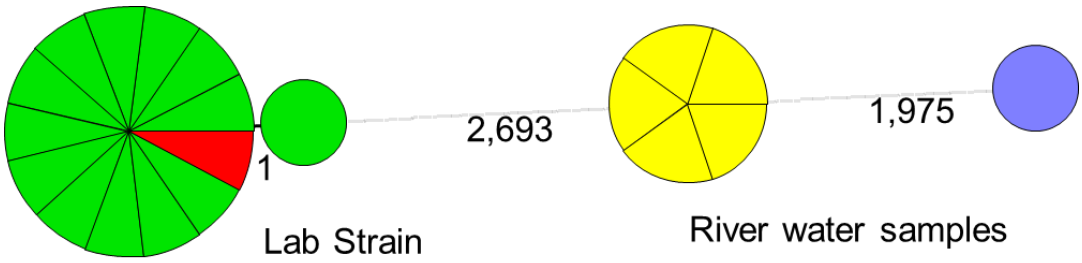
WHOLE GENOME SEQUENCING

ESCHERCHIA COLI

WGS of pathogens is very powerful, but it is not often that samples are obtained and tested for pathogens. In contrast the indicator organism *Escherchia coli*, is the basis of microbial water guidelines. In New Zealand each year tens of thousands of samples are tested for *E. coli*, and detections do occur. WGS can be applied to provide some insight. When these detections are unexpected, they are sometimes attributed to laboratory error or cross contamination. Laboratory controls minimise the possibility of these, but WGS does allow elimination of contamination from a laboratory positive control. From a plate or colilert sample, *E. coli* can be isolated and sequenced. Many laboratories use as an *E. coli* media quality control strain, ATCC 25922 which was imported to New Zealand in 1974. This strain was originally isolated from a clinical sample in Seattle, Washington, USA in 1946. It therefore should be very different to any

environmental strains in New Zealand. Figure 2 illustrates scenarios of matches with laboratory strain and also where multiple different genotypes of *E. coli* would support a more environmental source.

Figure 2: Whole genome (wgMLST) minimum spanning tree (up to 5,000 loci) of Escherichia coli isolates. Numbers on the branches are the number of cgMLST differences. Size of circles reflects the number of isolates with that same cgMLST profile. The lab strain (red) is indistinguishable from the green isolates indicating laboratory contamination of a sample. In contrast the yellow and purple E. coli isolates are thousands of alleles different suggesting environmental source.



METAGENOMICS

Assuming that E. coli detected in a water sample is not from a laboratory positive control then when *E. coli* or total coliforms are detected, there will also be many more other microorganisms present, and potentially DNA from mammals or others sources of interest. This is the power and the promise of metagenomics to allow insight to be gained from assessment of the community of organisms whose cells and/or DNA is present.

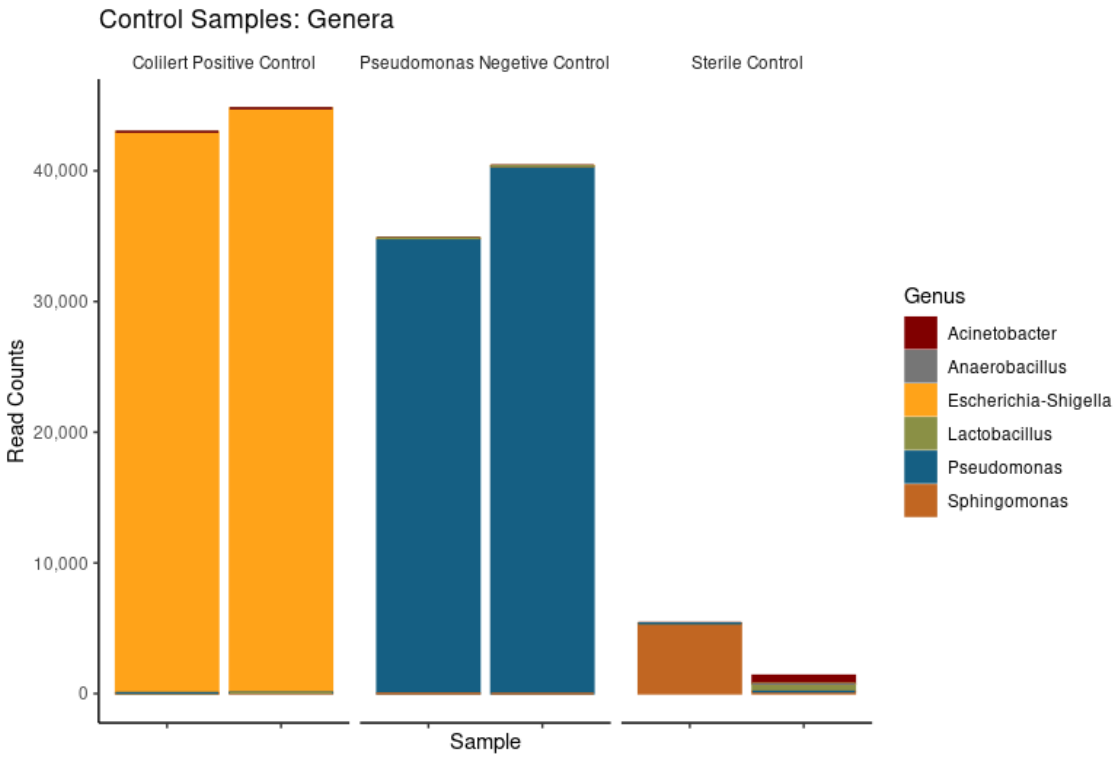
COLILERT TRAYS

From a water sample a colilert tray is often used to enumerate levels of Total coliforms and *E. coli* (Figure 3). We undertook metagenomic analysis of the individual wells from colilert trays with positive, negative and sterility controls and with river water samples. DNA was extracted from the liquid from individual wells and analysed using 16s metagenomics.

Figure 3: Colilert tray with yellow wells indicating Total Coliforms and clear wells no total coliforms. The yellow wells when exposed to UV light then fluoresce if E. coli are present.

The first experiments were to review the metagenomic profile of the colilert trays inoculated with positive control *E. coli,* negative control Pseudomonas, and with sterile water only (Figure 4). Less than 5,500 reads were obtained from the sterility control compared with over 35,000 from the positive or negative control samples. The reads in the sterility controls will be from the sterile water or from some of the analysis reagents. While there were a small number of non-target reads in the positive and negative control samples, each was dominated by the *E. coli* or Pseudomonas reads. A colilert sample contaminated with the lab strain would be expected to look similar to this inoculated control.

Figure 4 The range of genera of bacteria detected in colilert wells for positive, negative and sterility controls.



In contrast river water samples (Figure 5) containing between 50 and 110 *E. coli*/100 mL, had in *E. coli* positive wells not just *Escherchia*, but also other bacteria. The other wells with only total coliforms or negative for total coliforms had a range of genera present that differed between each river sample. While this demonstrates that metagenomics can be applied to colilert wells, and evaluation of the genera present may provide some guidance, the enrichment process in testing does bias and restrict the diversity of reads obtained.

In contrast the right hand panel of Figure 5 is the same river samples, but with metagenomics directly on extracted water samples. While there are less reads, there is more diversity without the enrichment of the colilert growth media.

Escherchia are still detected, but with a range of other bacterial genera. If this was drinking water then further analysis could provide insight to sources of contamination and potential health risks.

Metagenomic analysis also allows detection of non-microbial DNA. As illustrated in Figure 6, DNA from animals, birds, fish and other animal groups can be detected. This detection can drill down to identify species of fish or animals whose DNA is present.

Figure 5. The range of genera of bacteria detected from three different river samples from either colilert wells (left hand side) and filtered water samples without enrichment (right hand side).

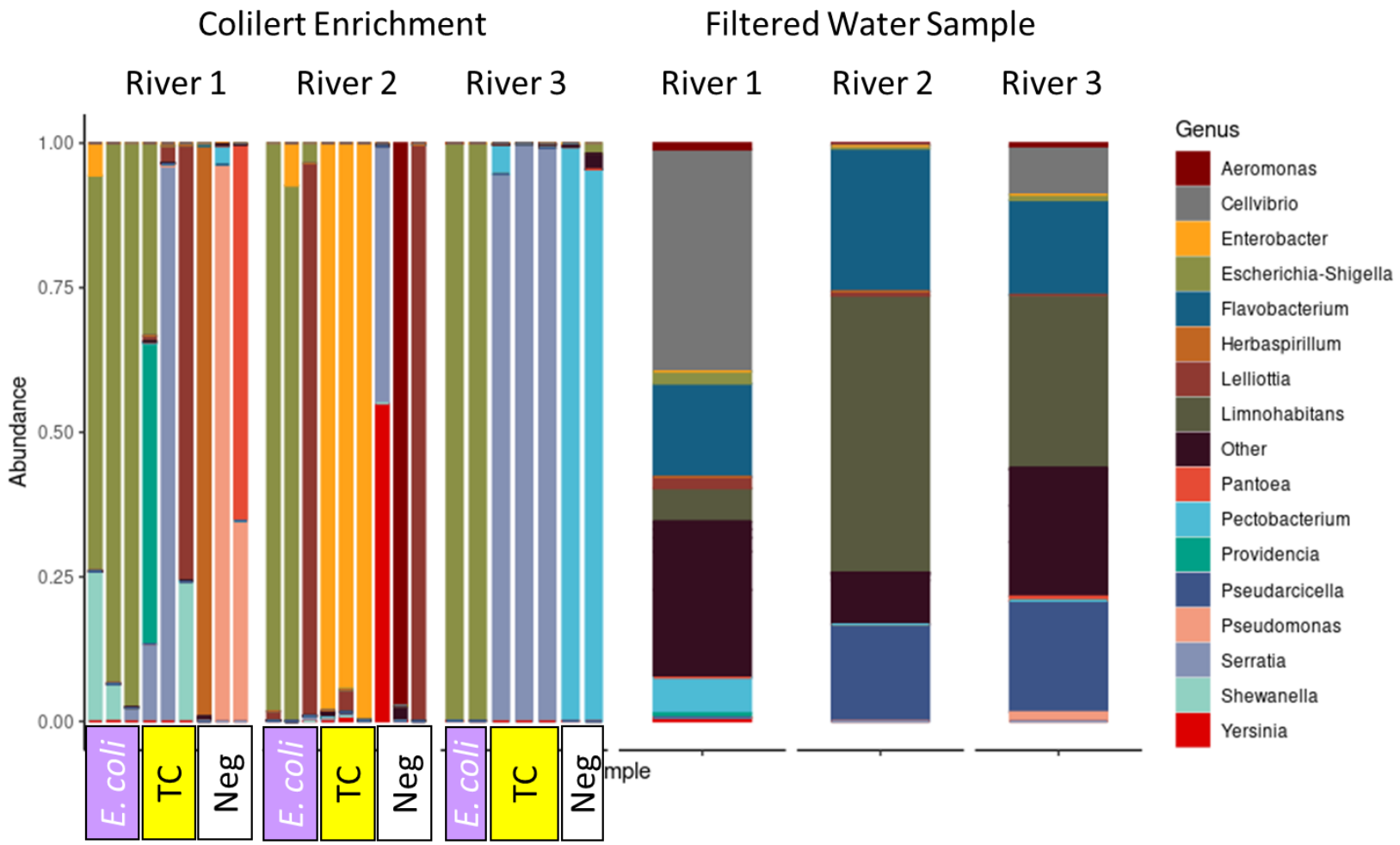
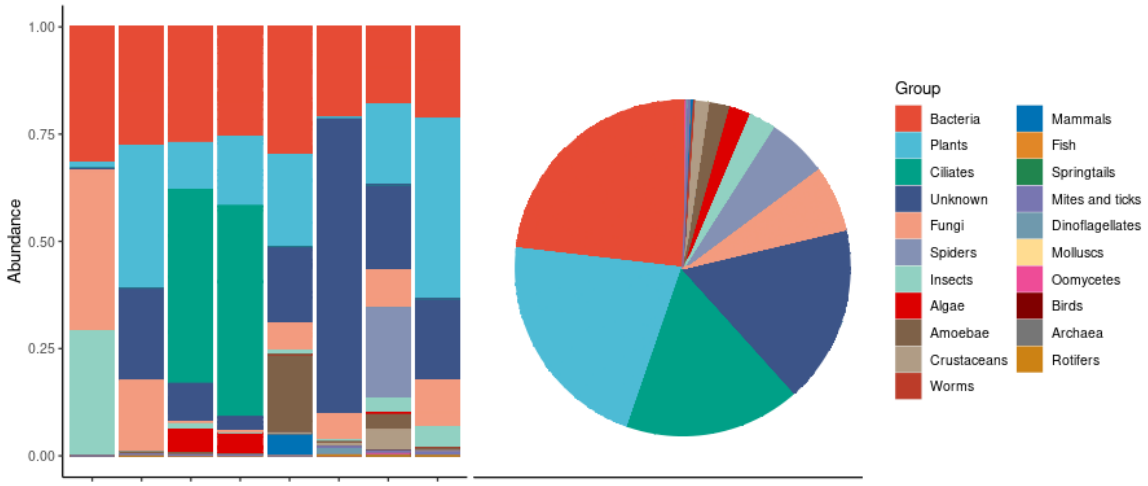


Figure 6: Metagenomic profiles from eight different drinking water samples are shown in the barchart and the piechart represents a total composite of eDNA from all sites.



CONCLUSIONS

The tools and approaches outlined in this paper could have, if available and if applied, been used to confirm the contamination risk in the Havelock North water BEFORE the August 2016 outbreak. Whole genome sequencing of any *E. coli* could have confirmed they were not laboratory contaminants.

Metagenomic analysis of the bore water, provided sampling was done at the right times, could have confirmed that the water was not as old as thought. Micro-organisms consistent with surface water inputs and faecal contamination are likely to have been detectable. In the case of the Brookvale Road bores in Havelock North sheep were only occasionally present. If sampling was done during periods when they were present, and when rainfall was occurring then it is likely sheep DNA or bacteria indicative of sheep would have been detected. Pukeko and other birds were present year-round, so it is possible that markers indicative of pukeko could have been detected. These would have challenged the notion that this was a secure bore.

Of even greater promise is the application of these tools to understanding and managing drinking water supplies. More thorough source water characterization, particularly assessment of the sources of contamination, will assist with decisions around the level of water treatment required, and with actions that could be taken to reduce the contamination risks.

Application of metagenomics to reticulated water offers the opportunity to also explore questions around what changes are occurring in the water. Whether biofilms are impacting on the water. When total coliforms are detected, what the health risk might be, and what the source might be. The growth in the number and scope of publications exploring drinking water and metagenomics highlights that this will be in increasingly important component of drinking water safety (see reference list).

ACKNOWLEDGEMENTS

We would like to acknowledge support for this work from ESR Sustainable Investment Fund, and from the Health Research Council.

REFERENCES

Bowyer, R., Schillereff, D. N., Jackson, M. A., Le Roy, C., Wells, P. M., Spector, T. D., & Steves, C. J. (2020). Associations between UK tap water and gut microbiota composition suggest the gut microbiome as a potential mediator of health differences linked to water quality. *The Science of the total environment*, *739*, 139697. https://doi.org/10.1016/j.scitotenv.2020.139697

Brumfield, K. D., Hasan, N. A., Leddy, M. B., Cotruvo, J. A., Rashed, S. M., Colwell, R. R., & Huq, A. (2020). A comparative analysis of drinking water employing metagenomics. *PloS one*, *15*(4), e0231210. https://doi.org/10.1371/journal.pone.0231210

Bruno, A., Agostinetto, G., Fumagalli, S., Ghisleni, G., & Sandionigi, A. (2022). It's a Long Way to the Tap: Microbiome and DNA-Based Omics at the Core of Drinking Water Quality. *International journal of environmental research and public health*, *19*(13), 7940. https://doi.org/10.3390/ijerph19137940

Devane, M. L., 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. https://doi.org/10.1016/j.watres.2020.116204

Hegarty, B., Dai, Z., Raskin, L., Pinto, A., Wigginton, K., & Duhaime, M. (2022). A snapshot of the global drinking water virome: Diversity and metabolic potential vary with residual disinfectant use. *Water research*, *218*, 118484. https://doi.org/10.1016/j.watres.2022.118484

Gilpin, B. J., Walker, T., Paine, S., Sherwood, J., Mackereth, G., Wood, T., Hambling, T., Hewison, C., Brounts, A., Wilson, M., Scholes, P., Robson, B., Lin, S., Cornelius, A., Rivas, L., Hayman, D., French, N. P., Zhang, J., Wilkinson, D. A., Midwinter, A. C., … Jones, N. (2020). A large scale waterborne Campylobacteriosis outbreak, Havelock North, New Zealand. *The Journal of infection*, *81*(3), 390–395. https://doi.org/10.1016/j.jinf.2020.06.065

Mahajna, A., Dinkla, I., Euverink, G., Keesman, K. J., & Jayawardhana, B. (2022). Clean and Safe Drinking Water Systems *via* Metagenomics Data and Artificial Intelligence: State-of-the-Art and Future Perspective. *Frontiers in microbiology*, *13*, 832452. https://doi.org/10.3389/fmicb.2022.832452

Li, Q., Yu, S., Yang, S., Yang, W., Que, S., Li, W., Qin, Y., Yu, W., Jiang, H., & Zhao, D. (2021). Eukaryotic community diversity and pathogenic eukaryotes in a full-scale drinking water treatment plant determined by 18S rRNA and metagenomic sequencing. *Environmental science and pollution research*

international, *28*(14), 17417–17430. https://doi.org/10.1007/s11356-020-12079-y

Li, Q., Yu, S., Li, L., Liu, G., Gu, Z., Liu, M., Liu, Z., Ye, Y., Xia, Q., & Ren, L. (2017). Microbial Communities Shaped by Treatment Processes in a Drinking Water Treatment Plant and Their Contribution and Threat to Drinking Water Safety. *Frontiers in microbiology*, *8*, 2465. https://doi.org/10.3389/fmicb.2017.02465

Liguori, K., Keenum, I., Davis, B. C., Calarco, J., Milligan, E., Harwood, V. J., & Pruden, A. (2022). Antimicrobial Resistance Monitoring of Water Environments: A Framework for Standardized Methods and Quality Control. *Environmental science & technology*, *56*(13), 9149–9160. https://doi.org/10.1021/acs.est.1c08918

Oh, S., Hammes, F., & Liu, W. T. (2018). Metagenomic characterization of biofilter microbial communities in a full-scale drinking water treatment plant. *Water research*, *128*, 278–285. https://doi.org/10.1016/j.watres.2017.10.054

Perrin, Y., Bouchon, D., Delafont, V., Moulin, L., & Héchard, Y. (2019). Microbiome of drinking water: A full-scale spatio-temporal study to monitor water quality in the Paris distribution system. *Water research*, *149*, 375–385. https://doi.org/10.1016/j.watres.2018.11.013

Rahmatika, I., Kurisu, F., Furumai, H., & Kasuga, I. (2022). Dynamics of the Microbial Community and Opportunistic Pathogens after Water Stagnation in the Premise Plumbing of a Building. *Microbes and environments*, *37*(1), ME21065. https://doi.org/10.1264/jsme2.ME21065

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., & Thermes, C. (2018). The Third Revolution in Sequencing Technology. *Trends in genetics : TIG*, *34*(9), 666–681. https://doi.org/10.1016/j.tig.2018.05.008

Vanhaecke, T., Bretin, O., Poirel, M., & Tap, J. (2022). Drinking Water Source and Intake Are Associated with Distinct Gut Microbiota Signatures in US and UK Populations. *The Journal of nutrition*, *152*(1), 171–182. https://doi.org/10.1093/jn/nxab312

Zhang, L., Chen, F., Zeng, Z., Xu, M., Sun, F., Yang, L., Bi, X., Lin, Y., Gao, Y., Hao, H., Yi, W., Li, M., & Xie, Y. (2021). Advances in Metagenomics and Its Application in Environmental Microorganisms. *Frontiers in microbiology*, *12*, 766364. https://doi.org/10.3389/fmicb.2021.766364

Zhou, Z., Xu, L., Zhu, L., Liu, Y., Shuai, X., Lin, Z., & Chen, H. (2021). Metagenomic analysis of microbiota and antibiotic resistome in household activated carbon drinking water purifiers. *Environment international*, *148*, 106394. https://doi.org/10.1016/j.envint.2021.106394