Microbial Source Tracking for the UK Water Industry

Aidan Francis Frederick Robson

Thesis submitted to Newcastle University in partial fulfilment of the requirements for the degree of Doctor of Engineering

17th December 2019

Abstract

This thesis evaluates the suitability of two emerging microbial source tracking (MST) techniques, host-associated *E. coli* biomarkers and community-based MST.

Previous human-associated *E.coli* markers (H8, H12, H14, H24) were evaluated for the first time in the UK; the sensitivity of H8 (10%) was lower than previously reported (50% (Gomi *et al.*, 2014)) and if analysed through regulatory culture-based approaches alone, would have resulted in a high false negative rate (90%). In light of this, the Hu100 marker, with the highest abundance (2.64 x 10^6 gene copies/100 mL) across 14 wastewater treatment plants, was developed through interrogation of 263 *E.coli* genomes. The abundance of Hu100 was not significantly different to other markers, which, could be due to the large variability in the proportion of *E.coli* containing biomarkers. Due to this variation, it is recommend that the total marker abundance is used to compare different sites.

Community-based MST uses high-throughput sequencing to compare bacterial communities of environmental samples, such as sea water, faecal taxon libraries (FTLs) which contain bacterial communities from known sources. Simulated microbial communities were used to evaluate how the composition of FTLs affected the accuracy and sensitivity of community-based MST. The inclusion of local samples appears to be more important than the size of the FTL to the accuracy of community-based MST. Furthermore, the inclusion of a river water sample as a 'background sample', improved method sensitivity from a 5% mixture of the sewage bacterial community in river waste to a 2% contribution of sewage.

Two catchment studies highlighted the ubiquity of urban diffuse pollution, largely from septic tanks and misconnections, in rural and semi-rural catchments. Community-based MST showed a good correlation with human-associated markers and (r_s >0.467, p <4.45x10^{-06}), but only when human sources were dominant. Findings suggest that community-based MST is more useful than marker-based MST to survey catchments for a range of potential pollution sources.

Investing ~£230k to perform MST in-house is the best option for Northumbrian Water, and other water companies, to incorporate qPCR and sequencing into their workflows. While >3000 samples need to be processed to achieve a return on investment, the business risk remains small, and other areas of the business will benefit from this investment.

Acknowledgements

Studying for an engineering doctorate has been an incredibly challenging experience, and I am deeply grateful to the people who have made it possible and enjoyable, and those who helped me make the most of this experience.

There are so many people at Newcastle University to thank, that it is not possible to name them all here. I would like to thank Russell Davenport and Vedrana Kutija who provided the inspiration for this project, I am especially grateful to Russell for his support and for improving my writing no end. The analytical side of this project would not have been possible without training and advice from the technical staff at Newcastle, especially, Sarah Smith, Amy Bell, and Dave Race. I am indebted to Tom Curtis for his constant encouragement and who, although not involved directly in the project, provided many opportunities to develop as an environmental engineer. I cannot begin to thank Justine Easten enough for her constant pastoral support and understanding, and keeping the STREAM program at Newcastle running like clockwork.

Northumbrian Water and EPSRC funded this work, and thanks go to Northumbrian Water staff, particularly my supervisor Rheanne White, for giving their time and advice so freely. Technical advice and support from Allen Snape, Laura Wilkinson, and Brian Lillie was always most welcome, and Andrew Moore gave an incredible amount of time and his honest advice was invaluable to both the project and my personal development..

I am thankful to NERC and the Scottish Environmental Protection Agency for funding and hosting me for a three-month policy internship. I am particularly thankful to Cath Preston, Lindsey Green, Fiona Whyte and Lesley Whyte for their time and guidance during this internship.

This project would not have been possible without contributions from a number of people. I am thankful to Oliver Crudge, an MSc student who I had the pleasure to supervise through his dissertation project, for his contribution to the case study presented in chapter 3. Oliver collected the catchment samples and undertook some initial analysis including *E.coli* plate counts and qPCR for the catchment study in Chapter 3, and helped to write the description of the catchment presented in Appendix B2, which formed part of his MSc thesis. Ben Allen developed the R-code to create mixed samples in Chapter 5, and answered many of my stupid questions with his typical dry wit. The Seaton Sluice case study (Chapter 6) was enhanced with help from the Environment Agency sampling team and the Northumbrian Water scientific support team. I would like to thank Michael Donkin and David Griffiths for giving

the support of the Environment Agency sampling team who collected an additional bathing water sample from Seaton Sluice each week during the bathing water season.

Finally, I would like to thank my family and friends who have made this experience enjoyable. I must thank my mum and dad for always being loving and supportive, and for the endless supply of cake during the last 4 years. I am eternally grateful to Hannah who bore the brunt of my grumpiness, kept me level-headed, and has been unwaveringly supportive, and for which, I will always love her dearly.

Contents

Table of tables

Table of figures

List of abbreviations

BWD – Bathing Waste Directive

CFU – Colony forming unit

CSO – Combined Sewer Overflow

DGGE – Denaturing Gradient Gel Electrophoresis

EA – Environment Agency

EEA – European Environment Agency

FIO – Faecal Indicator Organisms

FRNAPH – F-RNA Specific Phage

ITGR – Intergenic Region

HTS – High Throughput Sequencing

LOD – Limit of Detection

LOQ - Limit of Quantification

MST – Microbial Source Tracking

NCBI – National Centre for Biotechnology Information

NWL – Northumbrian Water Limited

OECD – Organisation for Economic Co-operation and Development

OTU – Operational Taxonomic Unit

PCR – Polymerase Chain Reaction

PE – Population Equivalent

QPCR – Quantitative Polymerase Chain Reaction

PFGE - Pulsed Field Gel Electrophoresis

RBMP – River Basin Management Plan

TRFLP - Terminal Restriction Fragment Length Polymorphism

UWWTD - Urban Wastewater Treatment Directive

VBNC - viable but non-culturable

WFD – Waste Framework Directive

WWTP – Wastewater Treatment Plant

Chapter 1 Introduction

1.1 Environmental water quality

Water quality describes the suitability of water to sustain various uses or processes (Maybeck *et al*., 1996) and is described through a set of distinct parameters that depend on the intended use or process. For example, general inland and estuarine water quality requires consideration of chemical and ecological parameters whereas bathing, or recreational, water quality is described using largely microbiological parameters, since these present an immediate threat to human health.

Across Europe, a number of key pieces of legislation define and drive improvements to environmental water quality. The Urban Wastewater Treatment Directive (UWWTD, 91/271/EC) is concerned with reducing the impact of wastewater on environmental waters, the requirements of which include:

- The prevention of leaks from sewers;
- The limitation of pollution from combined sewer overflows (CSOs); and
- The addition of secondary treatment for all wastewater treatment plants > 2,000 population equivalents (PE) discharging to fresh and estuarine waters.

The Water Framework Directive (WFD, 2000/60/EC) established a systems-based approach to protect and improve the quality of inland surface, transitional (estuarine), coastal, and ground waters. The aim of the WFD was for water bodies to achieve "good status" by 2015. At the end of the first 6-year management cycle, however, the UK have elected to extend this deadline to the end of the third management cycle, 2027. The WFD requires management of systems on a catchment scale (Voulvoulis *et al*., 2017) to achieve 'good ecological status', which is defined as:

> *"The state of a system in the absence of anthropogenic influences"*
>
> *(Voulvoulis et al., 2017)*

As such, there are no absolute standards to define good ecological status, rather, what is the ecological potential of individual water bodies if we removed human influences. There is an assumption then that humans can use these waters for reservoirs or hydroelectricity, for example, as long as the ecology of the water system meets its full potential. It is therefore vital that all unwanted anthropogenic impacts on all waterbodies are removed or at least reduced to a minimum to achieve the WFD’s aim.

The European Bathing Water Directive (BWD, 2006/7/EC), which compliments the WFD (2000/60/EC), serves two main purposes:

1) To provide a framework for monitoring, reporting and regulating microbial water quality; and
2) To reduce the public health risk from microbial contamination of recreational waters (Oliver *et al*., 2014).

These policy drivers are responsible, at least in part, for the improvements and drive to continuously improve the quality of environmental water in the UK.

1.1.1 Current state of water quality in the UK

The quality of surface and recreational water quality has improved significantly since the implementation of the UWWTD (91/271/EC), WFD (2000/60/EC) and BWD (2006/7/EC), although further mitigations are required to improve water quality. Of the 626 designated bathing water sites in the UK, only 62% are classified as ‘Excellent’, significantly less than the European average of 85% (EEA, 2015).

Improvements to recreational water quality have largely been achieved through investment in infrastructure assets. The most notable improvements in water quality came after the year 2000 when secondary treatment became a requirement for wastewater treatment plants with a population equivalent (PE) > 2,000 according to the UWWTD (91/271/EC) (Figure 1.1).

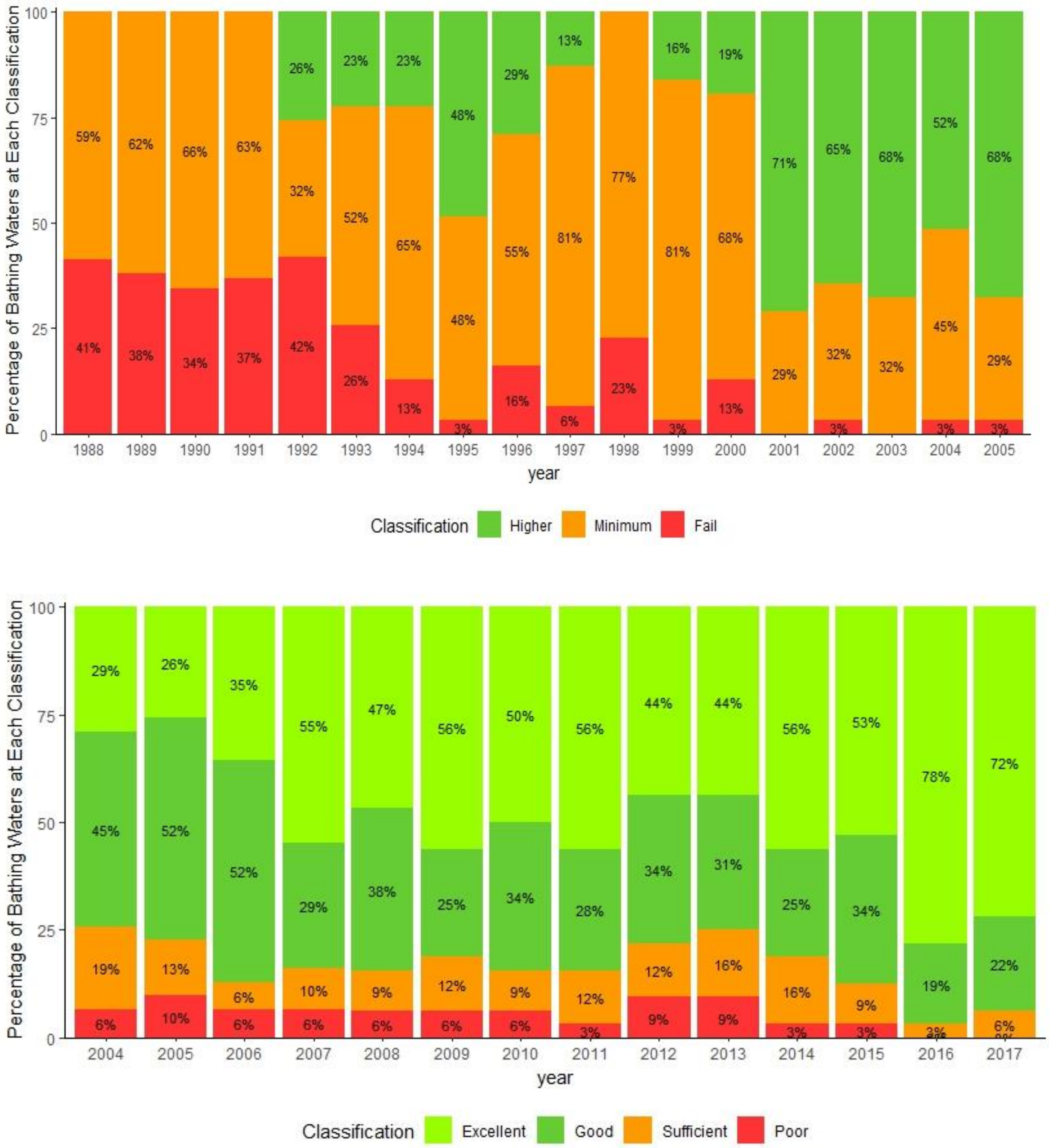


Figure 1.1 Percentage of North East bathing waters which achieved each classification according to the 1976 (Top) and 2006 (Bottom) Bathing Water Directives. Data from Environment Agency bathing water data explorer (EA, 2018) taking all bathing waters in the areas supplied for water and wastewater by Northumbrian Water

However, diminishing returns on investment in terms of water quality are expected as pollution from obvious point-source wastewater discharges is mitigated. The impacts of diffuse pollution sources may increase pressures on water bodies through: continued urbanisation; an increase in the frequency and intensity of rainfall events; population growth; and intensification of agriculture (Jalliffier-Verne *et al*., 2016). While there is a desire to move all bathing waters to excellent water quality, the impact of diffuse pollution (See 1.3 Threats to water quality and public health) means that moving from good to excellent is likely to be much more difficult than from poor to sufficient.

The quality of UK surface waters, graded by the WFD (2000/60/EC), remains concerning since only 36% of UK rivers achieve good ecological status (Priestley, 2015), a figure which has remained stagnant for around nine years. Although, the ‘one out, all out rule’, whereby a failure on a single parameter results in an overall failure of the water body, does not reflect improvements in other quality elements such as those observed in chemical water quality parameters (Voulvoulis *et al.,* 2017).

New methods to inform investment decisions are required to cost-effectively improve the quality of environmental waters in the UK. While an estimated £26 billion of investment is required to improve 80% of England’s surface water to good status (Environment Agency, 2014), the EA maintains an aspiration of having 60% of English surface waters achieve good status by 2021. Unfortunately, the latest river basin management plans (RBMPs) predict that only 27% and 25% of surface waters would achieve good status by 2021, for the Northumbria region (Environment Agency, 2016) and whole of England (Salvidge, 2016), respectively. For the Northumbria river-basin area alone, the required investment in surface-water is predicted to be around £820m over the next 37 years, with the Northumbrian Water taking £440m of this financial burden. It is, therefore, important that these investments are made into mitigation efforts that will have the largest impact on water quality and are economically justified. Methods to direct investment decisions towards the largest sources of pollution are, therefore, becoming increasingly important and paramount to delivering cost-effective water quality improvements.

1.1.2 Benefits of improving water quality

Determining the benefits of improving water quality for individuals is difficult, though most studies attempt to monetize the reduced risks of poor water quality on bathers. Recreational use of faecal contaminated water is most commonly associated with risk of gastrointestinal illnesses, ear, eye and upper respiratory tract infections (Prüss, 1998; Napier *et al.*, 2017). While few observational studies exist, the health burden of these swimming-related illnesses could be large. At two California beaches this health burden was estimated to cost > $3.3 million per year (Dwight *et al.*, 2005). The numbers of users and the cost of healthcare are likely to be less in the North East of the UK, however, this burden will be felt by local economies and the National Health Service. Indeed, the global-annual cost of thalassogenic diseases (those associated with wastewater in the marine environment) was estimated at $12bn or 3-million disability-adjusted life years (DALYs) (Shuval, 2003).

There are also a number of positive health benefits for users of recreational waters such as: the physiological and psychological benefits of exercise; physical and psychological restorative-ness and calming (Straughan, 2012; Phillips *et al*., 2018); and the alleviation of symptoms of chronic conditions such as depression (Denton, 2017); although these benefits are difficult to quantify in monetary terms.

Recreational water quality also brings a number of benefits to the local economy with increased revenue from tourism and marketing opportunities. An economic assessment estimated visits to five Scottish beaches at bathing water sites to be worth between £0.8million and £4 million (Phillips *et al*., 2018). While the quality of bathing water did not seem to affect the frequency of beach visits, visitors reported that it did diminish the quality of their visit. The Blue Flag award is desired by many local authorities as a marketing tool and assurance of quality to bolster tourism (Phillips *et al*., 2018). Blue Flag (2014) report that 61% of people across Europe check bathing water quality before visiting a beach. While such statistics seem unlikely, Blue Flag status is only available to sites with excellent water quality and, therefore, provides an economic driver to improve water quality.

Improving recreational water quality reduces the reputational risk to local authorities and water and sewerage companies, who are often looked to first when there is a water quality issue. Mass participation events, such as swimming or triathlons, present reputational risks especially where water quality is not excellent or subject to rapid decline. For example, epidemiology studies at some events have reported gastrointestinal illness in around half of all participants (Hall *et al*., 2017; Parkkali *et al*., 2017; Van Asperen *et al*., 1998) and often make the local and national media leading to reputational damage. Non-governmental organisations (NGOs) take advantage of media campaigns, using reputational risk as levers to drive water quality improvements. For example, Surfers Against Sewage (SAS) saw success with media campaigns to install secondary treatment for all wastewater discharges by The States of Guernsey in 2009 (Surfers Against

Sewage, 2009). Ensuring that water quality remains excellent is also therefore important to maintaining and improving corporate reputations.

1.2 Monitoring the quality of recreational waters

In the UK, weekly samples are collected from the 623 designated bathing waters between May and September, inclusive, as stipulated in the BWD (2006/7/EC). Regular analysis for two faecal indicator organisms (FIO), *E.coli* and enterococci, takes place on all samples, which are used to generate classifications for each bathing water and inform the public of water quality issues. In England, all regulatory testing occurs at a central laboratory facility, Starcross National Laboratory Service, in Exeter, UK. This is approximately 430 miles from the furthest northerly bathing water (Spittal, Berwick-Upon-Tweed, UK) and, as such, bathing waters are currently analysed up to 24 hours after collection (Oliver *et al.*, 2016). Due to the nature of the analytical methods (see section 2.4 Methods to enumerate faecal indicator organisms) which take 24-48 hours to produce a result, the public could potentially use poor quality water for up to three days before they are informed. The emphasis for managing bathing waters and health risk should therefore be on the long-term improvements of water quality, rather than reliance on a single sample.

1.2.1 Faecal indicator organisms (FIO)

FIO are used as a proxy for the presence of pathogens to indicate faecal contamination and a risk to public health. Ideally, monitoring would be in the form of routine monitoring for all pathogens of concern, however, this is currently unfeasible given: the diversity of potential pathogens; the episodic nature and low environmental concentrations of pathogens; the difficulty in culturing some pathogens; and the diverse analytical methods required to detect and quantify them (Field & Samadpour, 2007). The concept of using faeces-associated organisms to identify contamination of water was first proposed in 1891 by Mr and Mrs Frankland, six years after Theodor Escherichia (1885) described *Bacillus coli* (later renamed *Escherichia coli* (Castellani and Chalmers, 1919), who sought to provide evidence of sewage pollution (Hutchinson and Ridgway, 1977, Ashbolt *et al*,. 2001).

While the detection of FIOs is preferable to pathogen detection, there is an on-going debate about the suitability of current FIOs and the methods of their detection to determine the quality of recreational waters (Oliver *et al*., 2014). The ideal FIO should: be present in the intestinal tracts of warm-blooded mammals; reflect the presence and absence of pathogens; be present in greater concentrations than pathogens; have similar environmental survival profiles as pathogens (Field & Samadpour, 2007); be incapable of regrowth in the environment; be easily, rapidly and inexpensively detected in the environment; and be non-pathogenic (Ishii & Sadowsky, 2008). In Europe the BWD (2006/7/EC) uses both *Escherichia coli* and intestinal enterococci, to monitor and assess the quality of recreational water quality.

1.2.2 E.coli as a faecal indicator

E.coli are Gram-negative, lactose-fermenting, rod-shaped gamma proteobacteria which are facultative anaerobes, have high growth rates and are ubiquitous in the gastrointestinal tracts of most vertebrates (Clermont *et al*., 2008). Since 1891, when the concept of using sewage-associated organisms to identify dangerous contamination was first proposed (Hutchinson and Ridgway, 1977, Ashbolt *et al*,. 2001), these characteristics have made environmental monitoring of *E.coli* relatively quick, easy, inexpensive (McLellan & Eren, 2014) and a useful indicator of recent faecal contamination. However, due to monitoring limitations, the total coliform group, which includes *E.coli*, were originally used as regulatory indicators. The realisation that many total coliforms are common environmental inhabitants (Edberg *et al*., 2000) resulted in the use of faecal coliforms, and subsequently *E.coli*, as an FIO.

Once *E.coli* leave the nutrient-rich gut environment they may: die-off (or decay) through nutrient deficiency, dessecation, or predation (Wanjugi *et al*., 2016); enter a viable but non-culturable (VBNC) state (Ding *et al*., 2017; Oliver, 2010); or persist and grow in the environment (Solo-Gabriele *et al*., 2000; Ishii & Sadowsky, 2008). While most studies report *E.coli* die-off in the natural environment, some studies have observed *E.coli* to be persistent in a range of natural environments. These so-called ‘naturalised strains’ of *E.coli* are phenotypically and taxonomically indistinguishable from enteric strains (Walk *et al*., 2009; Deng *et al*., 2014). This may limit the efficacy of *E.coli* as a FIO by complicating our understanding of pollution sources and health risk. These naturalized strains, have been observed in soils (M N Byappanahalli *et al*., 2012), sands (Beversdorf *et al*., 2007; Vogel *et al*., 2016; Staley *et al*., 2016; Ishii *et al*., 2007), *cladophora* (a green algae) mats (B D Badgley et al., 2011; Verhougstraete et al., 2010), and surface waters (Tymensen *et al*., 2015a). Stress tolerant strains have also recently been noted in wastewater treatment plants (Zhi, *et al.*, 2016a) which may persist after UV or chlorination treatment. Some *Escherichia* strains have been assigned to five cryptic *Escherichia* lineages (CI to CV) which may be more prevalent in environments than mammalian guts (Walk *et al*., 2009). Whilst phenotypically indistinguishable, naturalized strains do differ genetically from enteric strains (Luo *et al*., 2011; M N Byappanahalli *et al*., 2012; Oh *et al*., 2012; Deng *et al*., 2014; Tymensen *et al*., 2015b), Luo *et al.* (2011) identified 120 and 84 genes highly associated with either enteric or environmental stains, respectively, although naturalized and enteric strains likely share a common ancestry (Tymensen, 2016). Interestingly, although clade V are most commonly isolated from environmental sources (Walk *et al*., 2009; Vignaroli *et al*., 2015), they also possess genes and adhesion properties associated with host gut persistence and virulence suggesting a potential for growth in both enteric and environmental systems (Vignaroli *et al*., 2015). It is worth noting that not all *E.coli* considered to be naturalized belong to a cryptic clade. Non-cryptic isolates possessing an environmentally associated gene have been noted, although further research is required to establish whether they are persistent in the environment (Deng *et al*. 2014). The effect of these naturalised strains on the efficacy of *E.coli* as an FIO is unclear, differentiation between these strains may be important, although long-term water quality monitoring largely overcomes this issue since increased numbers of *E.coli*, above baseline concentrations, can be used to indicate faecal contamination.

1.2.3 Enterococci as faecal indicators

The term ‘enterococci’ is used interchangeably with ‘intestinal enterococci’, the latter describing FIOs used in the European Union for water quality testing and defined by biochemical characteristics outlined by ISO 7899-1 (International Organisation for Standardisation (ISO), 1999). Enterococci were previously classified in group D of the genus *Streptococcus* (Lancefield, 1933) based upon physiological characteristics and later given a separate genus (*Entrococcus*) based on genetic evidence; DNA-DNA and DNA rRNA hybridisation studies showed that many species in group D were only distantly related to other groups (Schleifer & Kilpper-Balz, 1984).

Enterococci are Gram-positive, facultative anaerobes and are suggested to be better FIO than *E.coli*, particular in marine waters with enterococci showing a stronger relationship with swimmer gastrointestinal illness (Ostrolenk *et al.*, 1947; Wade *et al.*, 2003a). Enterococci are largely commensal, although some common strains such as *E. faecalis* and *E. faecium* are opportunistic pathogens which increasingly harbour antibiotic resistance mechanisms (Murray, 1990; Fisher & Phillips, 2009) and are a leading cause of nosocomial infections (Wilson *et al*., 2018). Ostrolenk, *et al.* (1947) were among the first to suggest enterococci as a better FIO than to *E.coli*. While *E.coli* was present in greater concentrations in 63% of faecal samples, the authors argue that lower concentrations could make enterococci a more reliable indicator of faecal contamination. In Europe, enterococci concentrations are used alongside *E.coli* to monitor both fresh and sea-water quality (BWD, 2006/7/EC), whereas in the US, *E.coli* is used solely for freshwater and enterococci for seawater due to the differential responses of these FIOs in environmental waters (Brooks & Field, 2016).

In environmental waters enterococci populations typically decrease over time (Byappanahalli *et al.*, 2012) due to the actions of environmental stressors, such as sunlight (Fujioka & Narikawa, 1982), although persistent populations have been observed. Enterococci generally have a greater salt tolerance than *E.coli* (Anderson et al., 2005a; Sinton et al., 2002), which likely leads to a better performance as an FIO in seawater. A range of other abiotic factors appear to contribute to the degradation of enterococci populations in environmental waters and are summarized in a recent review (Byappanahalli *et al.*, 2012). While many abiotic factors contribute to the decrease enteric enterococci populations, a number of studies have shown the persistence and potential growth of enterococci in extra-enteric environments. Whitman *et al.* (2003) identified both *E.coli* and enterococci in 97% of samples of *Cladophora* (a genus of filamentous green algae) from 10 beaches across four states, suggesting that C*ladophora* mats may act as a protective reservoir for FIO. There is also evidence that *Cladophora* provide enough nutrients to enable enterococci growth, although the evidence to date is limited to an experiment at 35 ^{o}C using algal leachate (Byappanahalli *et al*. 2003). This growth may therefore be limited to tropical climates. A range of other environmental reservoirs have been noted in studies including:

- Submerged aquatic vegetation (Badgley *et al.*, 2011);
- Beach sand, soil and sediments (Obiri-Danso & Jones, 2000; Yamahara *et al.*, 2009; Halliday & Gast, 2011); and
- Forage crops (Muller *et al.*, 2001; Ott *et al.*, 2001).

It is worth noting that whilst persistence of enterococci has been observed in these reservoirs, the evidence of growth in these environments is more sporadic. The evidence for growth stems from the high bacterial densities observed in tropical soils and sediments, typically moist, beach sand and sediments as well as in vegetation where high bacterial densities have been attributed to growth. As with *E.coli*, the effect of these reservoirs, and potentially naturalised populations, on the efficacy of enterococci as an FIO remains largely unknown.

1.2.4 Methods to enumerate faecal indicator organisms

Current culture-based techniques

The BWD (2006/7/EC, CEU, 2006) uses culture-based techniques, either membrane filtration or most probable number (MPN), to enumerate *E.coli* and enterococci in bathing waters. The membrane filtration methods (ISO 7899-2 (ISO, 2000), ISO 9308-1 (ISO 2014)) used by the Environment Agency require an incubation step of either 24 or 48 hours for *E.coli* and intestinal enterococci, respectively.

Culture-based approaches are associated with a number of limitations, particularly their slow speed due to the required incubation step (24-48 hours). This limits the ability of FIO monitoring to communicate short-term pollution events to the public in a timely manner (Ashbolt, Grabow and Snozzi, 2001). In addition, studies have shown temporal changes in FIO concentrations on times-scales of a day or less (Leecaster & Weisberg, 2001; Boehm *et al*., 2002), which could result in beaches remaining open, while contaminated, during laboratory processing, and the same beaches being closed after the contamination has passed, or being closed while contaminated. Culture-based methods also fail to identify the VBNC fraction of FIO, although, the understanding of the importance of the VBNC on regulatory monitoring and public health estimates is poor (Hassard *et al.*, 2017).

Rapid techniques

Rapid techniques overcome the major limitation of conventional culture-based methods, the incubation step, through the detection of cell properties. In a review of rapid methods, Noble and Weisberg (2005) note nucleic-acid-detection and enzyme/substrate methods to be the most common in water quality monitoring, particularly in the US. The US Environmental Protection Agency (USEPA) currently approves an enzyme/substrate test to quantify *E.coli* (Fricker *et al.,* 1997), although this is still culture-based and requires an incubation step, and has recently approved a nucleic-acid-based method, quantitative PCR (qPCR), to quantify enterococci.

A lawsuit against the USEPA to stimulate the use of a qPCR method has led to global debate regarding the use of rapid techniques in a regulatory context (Gooch-Moore *et al.,* 2011; Oliver *et al.,* 2016). However, it remains uncertain as to whether qPCR will be adopted as a rapid method for European regulatory use due to:

- A lack of robust epidemiological evidence linking genetic targets to human health risk;
- limited studies addressing the wider costs and benefits of adopting qPCR; and
- a lack of case-studies into the use of qPCR techniques, evidence base compared to culture-based techniques (Oliver *et al.,* 2016).

The use of rapid techniques in England is largely negated by the centralised nature of the Environment Agency's (EAs) laboratory service; additionally, the World Health Organisation (WHO) have not recommended their inclusion in the Bathing Water Directive at this time due to a lack of epidemiological evidence in Europe linking more rapid, molecular techniques to the risk to human health from bathing, and concerns over reported poor correlation with culture-based techniques (WHO, 2018). Nevertheless, there is a desire among regulators to gain more insight into the use of rapid methods and DNA based techniques, particularly for environmental and ecological monitoring (Walsh and Rhodes, 2016).

The polymerase chain reaction (PCR), now over 30 years old (Saiki *et al*., 1985), is an enzymatic process for the amplification of DNA. Purified DNA (template) from the microbial community of an environmental sample is introduced into a reaction mixture containing; DNA polymerase, nucleotides, magnesium chloride, and primers specific to the DNA fragment to be amplified, and is cycled through a program of temperatures which make up the steps of DNA denaturation, primer annealing, and DNA extension (Kralik & Ricchi, 2017). Each temperature cycle will, in theory, double the number of DNA molecules until the constituents become limited, and after successive cycles produce a DNA fragment of a specified known length, which can be visualised using fluorescent stains. The fragment is usually a gene or part of a gene that is a marker specific for the organism(s) requiring detection.

Quantitative-PCR (qPCR) takes advantage of the theoretical exponential increase in DNA between each cycle. A fluorophore is added to the PCR-reaction and the intensity of fluorescence, representing DNA concentration, is measured after each cycle. The absolute gene abundance in a test sample can be calculated by comparing it to a standard curve using known concentrations of the DNA target.

PCR techniques do not require an incubation step, which is both useful and limiting. Without an incubation step PCR assays are: rapid (~3 h); not limited to easily cultured organisms; and are able to quantify multiple targets simultaneously in a single reaction. However, the removal of an incubation step reduces the detection limit of the assay. Environmental waters often contain low and varying levels of FIOs and substances, which act as inhibitors, reducing the efficiency of a PCR reaction. This is a two-sided dilemma. To remove problems of inhibition, samples are often diluted, however, this can reduce the already low target concentrations below the limit of detection (LOD). Conversely, increasing the concentration of bacterial numbers from samples compounds the problem of inhibition. A further limitation to their regulatory use is the need to use a standard curve to calculate absolute abundance. How this standard curve is generated can greatly affect the estimated abundance. Hou *et al.* (2010) found a 3-4 log overestimation when using un-linearized plasmid preparation as standards compared to linearized, or PCR products. However, this limitation could be overcome by having a single set of standards made by a single laboratory distributed to all other laboratories undertaking this analysis, or through the use of digital PCR which does not require a standard curve (Cao *et al.*, 2015). In addition, qPCR does not differentiate between viable and non-viable cells, which can lead to an overestimation of FIO numbers with overestimations of around 0.8 $\log_{10}$ being reported (Raith *et al.*, 2014). This overestimation could also be due to the choice of DNA target as some targets have multiple copies within the bacterial genome

(See Sequencing below). Nevertheless, these overestimations must be explored if qPCR is to be incorporated into regulatory methods.

There are few studies (Hassard *et al*., 2017) which have explored the potential of qPCR for monitoring regulatory organisms in the UK, and few evaluating the impact of these techniques on management decisions (Kinzelman and McLellan, 2009; Walker *et al.*, 2015; Goodwin *et al.*, 2016). For beach management decisions Raith *et al.* (2014) compared the use of qPCR and culture-based techniques noting that 87% of the samples resulted in the same beach management decision, although, in 12 % of samples, qPCR would result in management action, such as posting signs, when culture-based techniques would not.

1.2.5 Classification of bathing waters in the UK

Bathing waters in Europe are classified as Excellent, Good, Sufficient, or Poor, based on the concentrations of *E.coli* and enterococci in weekly samples across a 4-year rolling data set (Table 1.1). The collection of microbiological data over long periods allows assessments of the general 'state of the environment', and efficacy of management practices and policies in achieving their environmental outcomes (Oliver *et al*., 2014). The classifications (Table 1.1) are easier for the public to interpret than those in the previous BWD (76/160/EEC), "Mandatory" and "Guideline", to better inform the public of water quality. Information is disseminated to the public through signage (Figure 1.2) at each designated bathing water showing the classification and any additional information, such as the susceptibility of a bathing water to short-term pollution events.

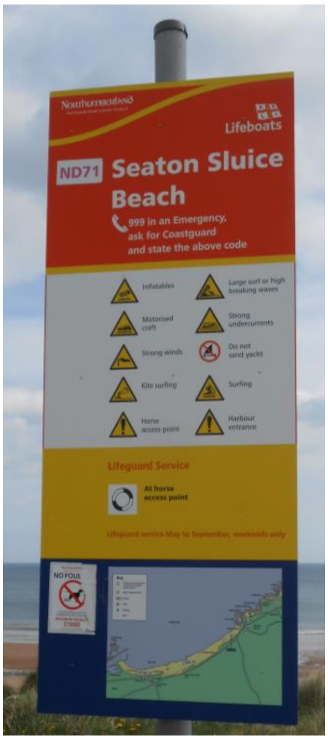


Figure 1.2.Example of a bathing water classification sign at Seaton Sluice beach (Chapter 6)

Table 1.1 Bathing water classifications used in the previous (76/160/ EEC) and revised (2006/7/EC) bathing water directives for coastal and transitional waters

Classification	**Requirements**
Excellent	Intestinal enterococci 100 CFU/100 mL* *Escherichia coli* 250 CFU/100 mL*
Good	Intestinal enterococci 200 CFU/100 mL* *Escherichia coli* 500 CFU/100 mL*
Sufficient	Intestinal enterococci 185 CFU/100 mL** *Escherichia coli* 500 CFU/100 mL**
Poor	If the percentile values for the last assessment period are worse than 'Sufficient values'.

CFU – Colony forming unit
** Based upon 95-percentile of samples taken through the bathing water season.*
*** Based upon 90-percentile of samples taken through the bathing water season.*

The prediction, management and communication of short-term pollution events to the public is particularly important to the classification of bathing waters. A maximum of 1 sample each year, or 15% of the total samples with high numbers of FIOs may be removed from the four-year rolling data set on the conditions that: 1) warning signs are present when the sample is taken (and the sampling team has seen the signage) and 2) attempts have been made to monitor or mitigate sources of short-term pollution. This can improve the classification of a bathing water for a long time due to the 4-year data set used. As a result Northumbrian Water has invested in automatic signage for willing local authorities to warn water users when water quality may be impaired due to short-term pollution events.

While the classification of bathing waters is maintained, the efficacy of this signage in reducing public health risk in the UK is unclear. A recent survey noted that many people overestimated the quality of water, 40% of those surveyed did not know or incorrectly stated the notified bathing water quality and 70% of respondents said they had seen the bathing water signage when there was no signage (Phillips *et al*., 2018). So whilst signs may maintain the classification of a bathing water, they may do little to protect public health, suggesting that policy should be directed at reducing short-term pollution events and increasing the resilience of bathing waters rather than limiting use during pollution events.

1.2.6 Faecal indicator organisms and health risk

The density of faecal indicator organisms has been shown to have some relationship to the health risk to swimmers (Kay et al., 1994a). This relationship allows regulators and beach managers to govern health risk by monitoring water quality and taking action such as closing beaches when FIO concentrations rise above a risk level.

The European bathing water classification system (Table 1.1) is based on the relationship between FIO density and health risk according to the World Health Organization's (WHO) Guidelines for Safe Recreational Water Environment (WHO, 2003). The WHO guidelines were based on epidemiological studies conducted in the UK in the 1990's (Fleisher *et al*., 1996; Kay *et al*., 1994). Good and excellent water quality approximately correspond to a 10% and 3.9% risk of gastrointestinal illness (WHO, 2003), respectively.

Unfortunately, the relationship between FIO concentrations and health risk is questionable. Evidence for this relationship is typically derived from epidemiology studies which rarely show a definitive relationship between FIO and health risk (Fewtrell & Kay, 2015). Two reviews summarise pre-2003 (Wade *et al*., 2003b) and post-2003 (King *et al*., 2015) epidemiological studies. Wade *et al.* (2003) noted that *E.coli* was a more consistent predictor of health risk whereas in marine waters enterococci showed a better relationship with health risk. The post-2003 evidence suggests a significant relationship between FIO in freshwater, but not in marine waters, although the review only considered 16 studies. Both reviews found significant heterogeneity between study protocols and severe methodological limitations in many papers. In addition, King *et al.* (2015) note that few studies were conducted in "Poor" quality water and none in "Sufficient" quality waters. Clearly, more epidemiological evidence is needed to support or refute the findings of the original studies on which the classifications (Table 1.1) are based.

The uncertainty in the FIO-health risk relationship could be due to different sources of pollution exhibiting a different level of health risk. Fewtrell and Kay (2015) noted that the poor FIO-health risk relationship was especially true where non-point sources of pollution were prevalent (See 4. Threats to water quality and public health). The studies at the foundation of the WHO guidelines (2003) were conducted in marine waters impacted by sewage (Kay *et al.,* 1994; Fleisher *et al.,* 1996), since then, only a handful of studies have considered the source of pollution. A summary of the epidemiological evidence linking non-human faecal pollution and health risk to bathers concluded that none of the studies used provided conclusive evidence for a relationship between non-human faecal contamination and gastrointestinal illness (Dufour *et al.,* 2012). The authors note, however, that other studies have shown a logical link between human infections and zoonotic pathogens, although, links between bathing in contaminated water and specific non-human sources are missing. More recently, studies using quantitative microbial risk assessment (QMRA) techniques have been used to support epidemiological evidence (Fewtrell & Kay, 2015).

Quantitative Microbial Risk Assessment (QMRA) takes a modelling approach to assess the risk of adverse outcomes from microbial agents using information about the spread, exposure to and dose-response model of specific microbial agents (Hass *et al.*, 1999). Soller *et al.* (2015) used a QMRA approach, taking published pathogen and FIO concentrations in faeces, to compare the risks of gastrointestinal illness (GI) from exposure to non-human and human sources of pollution. Risks were compared by normalizing the concentration of faecal matter to a known concentration of FIO. This analysis suggested that exposure to faecal contamination from gull, chicken and pig faeces presented a substantially lower risk of GI to bathers. Cattle faeces, however, presented a similar risk to human sources. While only six pathogens were used in this analysis, it highlights the importance of understanding the different sources of pollution that may affect a water body. The analysis also does not take into account the persistence of pathogens compared to FIO. We might therefore expect the relative risk from sewage to increase over time since FIO generally die-off faster than viruses and protozoa, the main purveyors of risk in sewage (Soller *et al.,* 2015). Nevertheless, an understanding of the potential health-risks associated with bathing waters and how best to mitigate these risks requires an understanding of the sources of pollution. It is worth noting that currently, management decisions are based on FIO concentrations, which are unlikely to reflect health risk when a mixture of pollutant sources are involved since identical FIO concentrations resulting from contamination with different sources present a different risk to health (Soller, *et al.*, 2010).

1.3 Threats to water quality and public health

Despite the range of policy and economic drivers, improvements to water quality are often difficult to achieve due to the difficulty in identifying, apportioning and mitigating threats to water quality. The origin of pollution can be described as point or diffuse source. Point sources have a definitive point-of-entry to a watercourse, e.g. the discharge point of a wastewater treatment plant. Diffuse pollution sources are often regarded only as agricultural sources of pollution (Oliver *et al*., 2014), however, urban diffuse sources exist. Pollution is considered diffuse when there is no single point of discharge (European Environment Agency, 2018) and urban when it originates directly from anthropogenic activities, such as, driving cars, or sewage discharging into streams (Lundy & Wade, 2013). Urban diffuse pollution is often overlooked and/or poorly understood; this may be due to the difficulty in applying modelling approaches to spatially and temporally sporadic events, the difficulty in remediating urban diffuse pollution, or the perception that agricultural pollution is that most problematic to water bodies.

1.3.1 Urban pollution

The realisation that water plays a major role in the transmission of certain diseases revolutionised our understanding of epidemiology and changed our sanitation practices. Before the link between the spread of cholera and drinking sewage-contaminated water, posited by Drs John Snow and William Budd, really gained acceptance (Cooper, 2001), the Report on the Sanitary Conditions of the Labouring Population of Great Britain (Chadwick and Flinn, 1842) was published. The resulting 1848 Public Health Act paved the way for local authorities to develop the combined sewer systems (Figure 1.3), many of which are still in use today across much of the UK (Cooper, 2001).

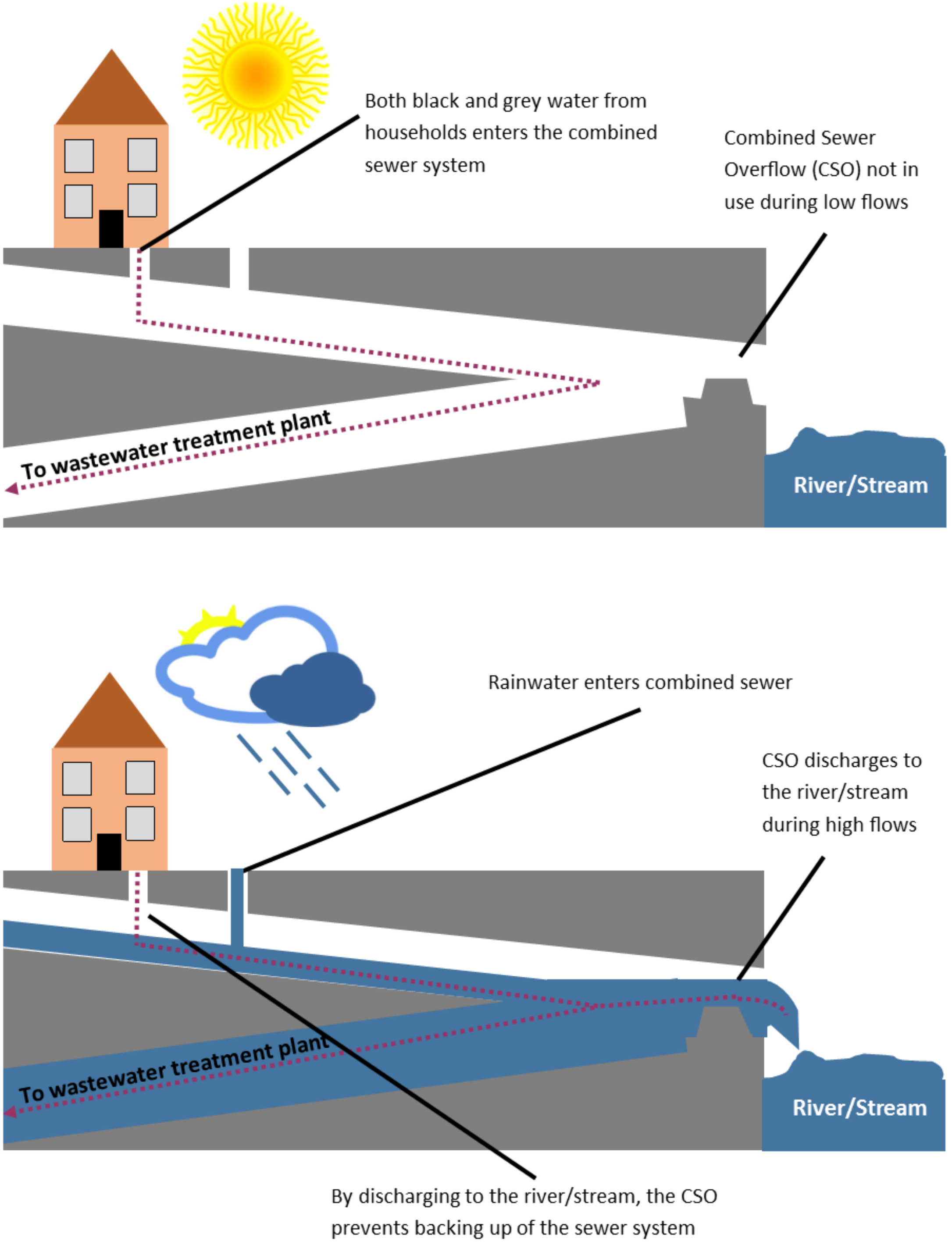

Both black and grey water from households enters the combined sewer system
Combined Sewer Overflow (CSO) not in use during low flows
To wastewater treatment plant
River/Stream
Rainwater enters combined sewer
CSO discharges to the river/stream during high flows
To wastewater treatment plant
River/Stream
By discharging to the river/stream, the CSO prevents backing up of the sewer system

Figure 1.3 Diagram of combined sewer overflow operation in dry weather (top) and during rainfall (bottom), adapted from (USEPA, 2004)

The UK's sanitation requirements are today served by a variety of sewer systems, which include combined sewers, separate sewers or private treatment systems such as a septic tank. Combined sewers (Figure 1.3) collect greywater from showers and sinks, blackwater from toilets, and surface run-off from rainfall in a single pipe where it is taken, ideally, for treatment before discharge to a water body. In separate systems, surface run-off enters a drain, running off directly to a watercourse, and is separated from black and greywater, which enters sewer systems.

Sewer systems present risks to water quality through both point and diffuse sources of pollution. Combined Sewer Overflows (CSOs, Figure 1.3) and Sewage Pumping Stations (SPSs) are infrastructure assets designed to relieve pressure on sewage systems by overflowing sewers filling beyond their design capacity in the event of blockages or during heavy rainfall events. This prevents sewers from backing up and sewerage flooding homes and streets. While discharges from CSOs are usually permitted, discharges often occur too frequently and these point sources of pollution present risks to public and environmental health (WWF, 2017). To tackle this, telemetry has been installed on all CSOs discharging to a bathing water. CSOs which discharge frequently, and to watercourses with a high amenity value, and which CSOs to monitor is determined by the EA, although, individual water companies are likely to install additional monitoring on CSOs which they believe will impact bathing water quality. Northumbrian Water currently monitor 1152 CSOs across the North East, recording the time and duration that a spill occurs (Snape, 2019). This telemetry is currently used to good effect to reduce public health risk via the Safer Seas app (SAS, 2018). When CSOs have overflowed for 30 minutes, the telemetry monitoring CSOs alerts the public via an app, highlighting areas of higher risk. In this sense, telemetry has reduced the risk to public health which CSOs pose, and allows maintenance to be carried out immediately if a CSO is overflowing too frequently, or unexpectedly.

Misconnections in separate systems, where the foul sewer pipe carrying black and grey water is connected to the surface water drain are common, difficult and expensive to detect and rectify, and have impacts on water quality which are often difficult to determine (Ellis & Butler, 2015). A government-commissioned review (Royal Haskoning, 2007) estimated that up to 1.25 million properties across the UK had misconnections, although Revitt and Ellis (2016) note reported rates to vary considerably with an average misconnection rate of from 3% up to 30% in some hotspot areas, the identification of which should be a priority (Ellis & Butler, 2015). Identifying misconnections is tedious and expensive. Misconnections have been estimated to cost the

UK water industry around £235 million/year in terms of asset management and maintenance (Royal Haskoning, 2007). While the estimated cost of identifying, repairing and rectifying misconnections varies considerably, for 500,000 homes this may cost between £393 million and £1.3 billion (Ellis & Butler, 2015). The exceptional costs of identifying misconnections, particularly those presenting a threat to public health, and the current lack of data (Ellis & Butler, 2015) makes methods to prioritise search areas paramount to reducing these issues in an economically efficient manner.

Leaking sewers should be classified as diffuse sources of pollution and are increasingly common with ageing sewer infrastructure that has exceeded original life expectancy across much of the UK. Sewers are likely to get older still with only 1% of sewer assets replaced between 2000 and 2008 (DEFRA, 2011), due to the high cost and disruption associated with sewer replacement. Failing sewer assets are therefore likely to become increasingly problematic. The timely detection and location of points of failure will be critical to cost-effective improvements and maintenance of sewers.

Septic tanks serve a large number of rural areas in the UK where connections to centralised sewer systems are not available (May *et al*., 2015a). Septic tanks typically consist of a two chambered tank where solids are removed through settlement and clarified effluent soaks away into the surrounding soil which is thought to be treated as it percolates through the soil (Wood *et al*., 2005). However, many older septic tanks are in use and may be undersized or receive rainwater causing them to overflow regularly. In addition, many older septic tanks discharge directly to water courses. Older, poorly functioning, or leaking septic tanks are major sources of nutrients, particularly phosphorus, in freshwater systems (May *et al.*, 2015b). For example, (Aitken *et al.*, 2001) found the 82% of septic tanks in a Scottish catchment discharged directly to water courses. These problematic septic tanks are likely sources of urban diffuse pollution in rural catchments.

1.3.2 Agricultural pollution

Agriculture covers around three-quarters of land use in England and Wales (DEFRA, 2018a). Both arable and livestock farming continue to impact the aquatic environment through routine agricultural processes; the use of pesticides, such as metaldehyde, and medicines, such as antibiotics and endocrine disrupting compounds, can release micropollutants, while slurry and fertilizer spreading and animal defection can lead to the leaching of nutrients, as well as potentially zoonotic microorganisms, into watercourses (DEFRA, 2018a). The well-known phenomenon, eutrophication, arises from excess nitrogen (N), phosphate (P) and potassium (K) and leads to the deterioration in the ecological quality of water. An estimated 50% of the phosphorus entering surface waters can be attributed to livestock and fertilizer (Morse *et al.*, 1993). In the UK poultry, sheep, cattle and pigs comprise the majority of livestock. Sheep and cattle make-up the majority of grazing stock (Figure 1.4), although by head of population, poultry has the highest number (~180,000,000 in 2017, (DEFRA, 2018b)).

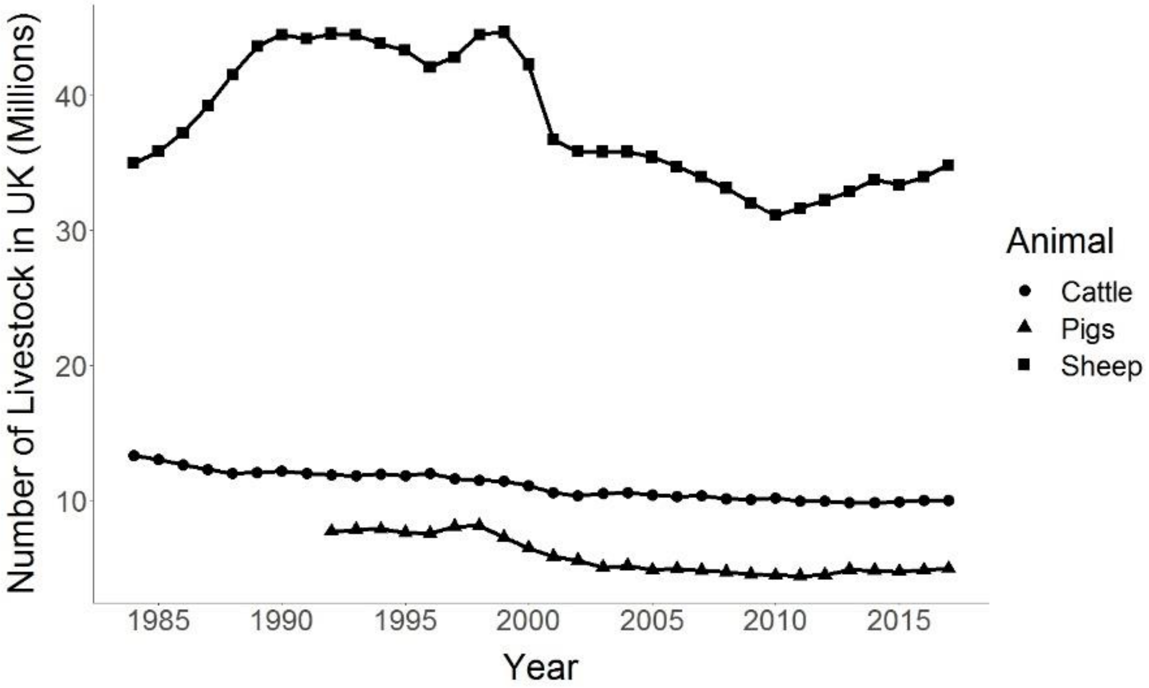


Figure 1.4. Trends in grazing stock numbers in the UK in 2017 (DEFRA, 2018b)

Algal blooms and the release of, potentially zoonotic, microorganisms into the environment are of concern for human health. Almost two-thirds of human pathogens and 60% of emerging pathogens have animal origins (Penakalapati *et al*., 2017).

1.3.3 Other threats

Wildlife such as birds, deer and rabbits can carry a number of human pathogens such as *Campylobacter spp.* (Lévesque *et al*., 2000) and therefore may present a threat to public health. While the risks to human health from animal sources are thought to be small, the presence of faeces from other sources may confound analysis of water quality by providing high numbers of FIO.

Pets are also potential reservoirs of FIOs and human pathogens. Many beaches (both globally and locally) restrict access to dogs and/or horses during bathing water season to reduce the risk to bathers and beach users. Justifying the banning of pets from beaches is rarely supported by with evidence. However, some studies have attempted try to determine potential loading of FIO by assessing the shedding of *E.coli* from the faeces of birds and dogs (Meerburg *et al*., 2011; Wright *et al*., 2009), or of faecal matter from bather themselves (Elmir *et al*., 2007). It is difficult to determine the accuracy of methods using the shedding rate of FIOs in faeces, although they provide a quick method of assessing the likely risks.

1.4 Microbial source tracking (MST)

Understanding the sources of pollution is vital to effective bathing and surface water management, however, the ubiquity of *E.coli* and enterococci in the mammalian gut limits their use in determining the sources of pollution (Hagedorm *et al.*, 2011). Despite the discussions regarding the poor efficacy of *E.coli* and enterococci as FIO, a recent World Health Organisation (WHO) review recommends their continued use within the Bathing Water Directive (WHO, 2018). This review (WHO, 2018) also highlights the potential for emerging techniques such as microbial source tracking (MST), which has developed as a useful tool for decision-makers, particularly in the US, and has influenced regulatory decisions (Nguyen *et al*., 2018).

It is important to note that a range of chemical source tracking techniques have proven useful in identifying sewage contamination. A range of chemicals, typically pharmaceuticals and personal care products, have been used as markers of sewage contamination. Chemicals such as caffeine, acetaminophen, and acesulfame have been noted to be useful as sewage markers and correlate well with FIO in sewage or environmental samples, while others such as carbamazepine have been less useful (Cantwell *et al*., 2016; Sauvé *et al*., 2012; Nödler *et al*., 2016). However, some of these chemicals may be less useful for detecting sewage from smaller scale decentralised works, due to the lower likelihood of members of the community, those contributing to the sewage, using medication such as acetaminophen and carbamazepine or using artificial sweeteners such as acesulfame. Differences in the environmental persistence of chemicals and FIOs may reduce their usefulness for microbial risk assessments, however, they are likely to represent the potential risks from some micro-pollutants better than microbiological agents. Advances in the use of chemical markers for the detection and monitoring of sewage contamination has been recently reviewed (Lim, *et al*., 2017).

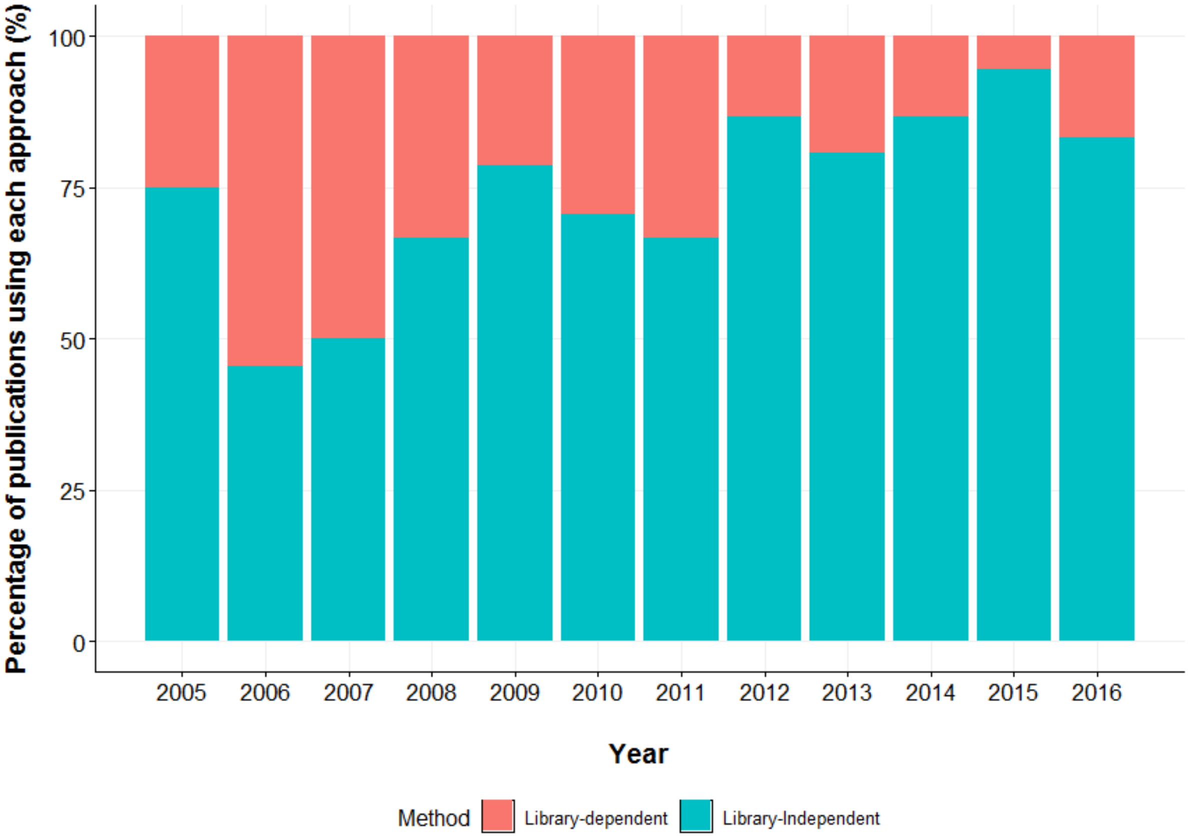


Figure 1.5. The percentage of published papers in Environmental Science and Technology (n=28), Water Research (n=84), and Applied and Environmental Microbiology (n=48), using either a library-dependent or independent approaches. Data was obtained through Web of Science and Scopus searches using the key term 'Microbial Source Tracking' on 6 Feb. 2017

MST describes a range of techniques that use microbes and their communities to identify and sometimes apportion the sources of faecal contamination in a receiving environment. There are two general approaches to microbial source tracking (MST). Library-dependent approaches compare the phenotypic or genotypic traits of a particular group of organisms isolated from impacted sites with pre-constructed libraries consisting of a large number of these organisms from likely sources of pollution (Simpson *et al.*, 2002). In contrast, library-independent approaches use previously identified genetic (often host-specific) targets, using the concentration of these as a proxy for the level of faecal contamination from a particular source (Harwood *et al*., 2014).

The majority of MST research involves the use of library-independent methods (Figure 1.5), which may be due to the more rapid nature of library-independent methods compared to library-dependent ones (Griffith *et al.,* 2003), the additional labour requirements of library-dependent methods, and the consistency in the performance of these approaches in different geographical areas (Ebdon & Taylor, 2006). Whether this trend will continue is unclear and may depend on how MST investigations are conducted, the purpose of the investigation and the expense, suitability and availability of techniques in the future. With new technologies, library-dependent approaches are becoming less labour intensive and may be able to distinguish between similar sources such as the faecal and non-faecal components of sewage (Newton *et al.,* 2013).

Although a wide range of methods exist, most recent research uses a narrow range of molecular techniques (Figure 1.5). The general dominance of PCR-based techniques is evident, and the transition from end-point PCR to quantitative PCR (qPCR) highlights the desire for MST to be more quantitative and rapid in nature. The increase in the popularity of sequencing (Figure 1.5) is due to both use in MST studies (Newton *et al*., 2013; Neave *et al*., 2014; Samarajeewa *et al*., 2015) and in biomarker discovery (Gomi *et al*., 2014; McLellan & Eren, 2014).

1.4.1 Library-dependent methods

The first attempts to link bacteria to sources involved building libraries of organisms from a single species, using techniques (Simpson *et al.,* 2002) such as; antibiotic resistance assays (ARA), ribotyping, and repetitive-sequence PCR. These techniques were quickly replaced by Pulsed Field Gel Electrophoresis (PFGE) and terminal restriction fragment length polymorphism (TRFLP) (figure 1.5). These methods generally target *E.coli* and enterococci which are easily isolated from environmental samples.

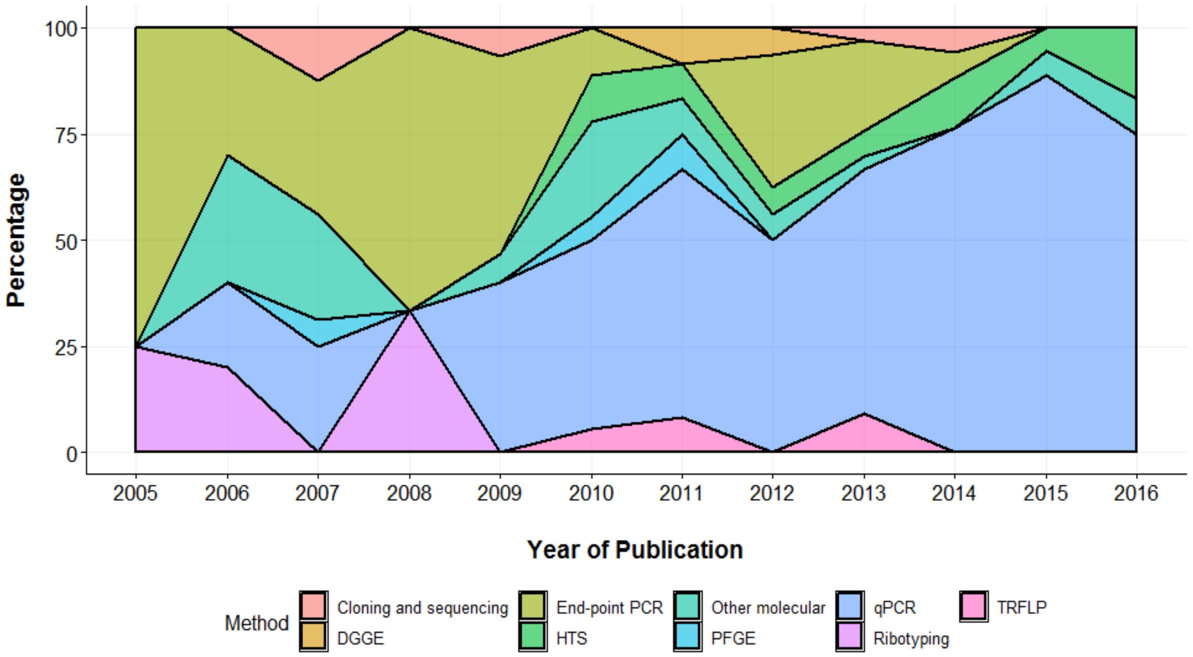


Figure 1.6. The proportion of molecular techniques used in studies published in each year between 2005 and 2016 in Water Research (n= 84), Environmental Science and Technology (n=28), and Applied and Environmental Microbiology (n=48) which use molecular techniques. Data was obtained through Web of Science and Scopus searches using the key term 'Microbial Source Tracking' on 6 Feb. 2017

Antibiotic resistance assays (ARA) compare the resistance profiles of *E.coli* (Parveen *et al*., 1997) or enterococci (Wiggins *et al*., 2003) isolated from different sources. However, when classifying bacteria by source, ARA can have variable and potentially low average rates of correct classification (57% - 94% (Wiggins *et al*., 2003; Scott *et al*., 2002)), with higher rates of correct classification generally occurring when library sizes are small and not representative of all sources. Wiggins *et al.* (2003) considered a representative library to consist of 6,587 isolates. In addition, factors such as gain or loss of a plasmid containing a resistance gene can complicate analysis (Scott *et al*., 2002) and resistance patterns are unlikely to be stable over time. In comparing seven different protocols to classify *E.coli* by source, Stoeckel *et al.* (2004) noted that ARA, carbon source utilization (CSU) and a ribotyping assay (RT-HindIII) classified less than 25% of blinded isolates that were already in the library, while CSU and RT-HindIII did not perform better than random at assigning a further 150 isolates to their known sources. In practice, false positive results would inhibit most protocols and larger libraries would be required to improve accuracy, reproducibility and geographical stability.

Pulsed-field gel electrophoresis (PFGE) is a DNA fingerprinting method used to infer relatedness between the genomes of organisms in epidemiology studies. While PFGE has been found useful in some studies, Parveen *et al.* (2001) found no association between the PFGE profile and source of an isolate. Repetitive-sequence PCR uses primers which target interspaced repetitive sequences in the bacterial genome to differentiate between similar strains of a single species, especially using the BOX primers (Versalovic *et al*., 1994). The BOX primers have been reported to correctly identified the source of 78-90% of isolates to sources (Dombek *et al*., 2000), and (Araújo *et al*., 2014) was able to correctly classify 78% of isolates between human and gull faeces, although this required a library of 592 isolates between the two sources. Unfortunately, the reproducibility of this technique and the fact that variability increases with library size means these techniques are generally unreliable (Meays *et al*., 2004; Dombek *et al*., 2000).

Between 2009 and 2014, TRFLP was a common method for community profiling (Figure 1.6). TRFLP involves the enzyme digestion of a single gene, amplified from a community and comparing the length and relative intensity of digested fragments. TRFLP targets a single gene possibly due to the volume of available literature and the inexpensive nature of the analysis (Cao *et al.,* 2013). However, comparisons against emerging next generation sequencing techniques allow a greater phylogenetic resolution and can identify a greater proportion of the microbial diversity than TRFLP methods (Cao *et al.*, 2013; Samarajeewa *et al.,* 2015).

1.4.2 Sequencing for water quality monitoring and MST

The progress in high throughput sequencing (HTS) technology, which has historically outpaced Moore’s law (Muers, 2011), means that tens or hundreds of samples can be processed simultaneously (multiplexed) and rapidly (< 24 hours). This progress has led to initial explorations in the use of HTS technology for water quality monitoring and MST.

By targeting the 16S rRNA gene, HTS allows for the identification of multiple bacteria to generate community-fingerprints. The 16S rRNA gene encodes the 16S ribosomal RNA, a structural component of the ribosome. Since ribosomes are a critical component in the production of proteins, vital to all life, the parts of the 16S gene are highly-conserved (Woese *et al*., 1975) across living organisms. The 16S gene is ideal for bacterial species identification since it contains nine hypervariable regions, separated by conserved regions. Differences in the hypervariable regions allow differentiation between bacteria while the conserved regions allow for the use of universal primers (Chakravorty *et al*., 2007). It is important to note, however, that there is currently a trade-off between the rate of errors introduced by sequencing and the length of DNA it is possible to sequence (read-length). Therefore most studies choose one or two hypervariable regions only. However, this is often insufficient identify and differentiate between all bacteria to species level, since each region varies in sequence diversity between different bacteria (Chakravorty *et al*., 2007). Importantly, Kumar *et al.* (2011) noted variations in the reported relative abundance of bacteria when different hypervariable regions were targeted in 16S rRNA sequencing, finding that the apparent dominant species changed when different hypervariable regions were targeted. Interestingly, averaging the results of the V1-V3 and V7-V9 regions gave communities similar to the result of Sanger sequencing of the whole 16S rRNA gene. Differences in bacterial communities was also observed when using different HTS technologies (Samarajeewa *et al*., 2015). Exploration of MST using different technologies may be beneficial in understanding these differences, and assessing how much differences in the perceived microbial community influences MST conclusions.

Identifying faecal associated bacteria and bacterial communities

Detection and comparison of bacterial communities has informed library-independent approaches in the development of new markers (Gomi *et al*., 2014; McLellan & Eren, 2014) and allows direct identification of source-associated bacteria and their communities. Iceton (2018) used a database and BLAST searches to identify host-associated bacteria in the communities of potential sources and sinks. This approach, however, was limited by the low sensitivity of the assay coupled with the low abundance of these bacteria and the low number of samples used in the study. A similar approach is occasionally taken in the simultaneous detection of putative pathogens (Tan *et al*., 2015; Batista *et al*., 2018). The efficacy of this approach is seriously limited by the sequencing methods' ability to reliably classify and identify taxa. Sequencing technologies are currently limited by the read length, the length of DNA it is possible to sequence with a high degree of certainty, which limits the taxonomic-level at which bacteria can be classified. Currently, it is accepted that genus-level classification is possible while species-level may be possible for some species depending on which gene and which region of a gene is sequenced. However, Tan *et al.* (2015) note that even if species-level classification is possible using the 16S rRNA gene, it would be insufficient to assess risk from pathogens with strain-specific virulence.

A more successful strategy for MST has been the identification of bacterial signatures, associated with faeces. Newton *et al.* (2013) identified faecal and non-faecal bacterial signatures within sewage by identifying three genera (*Acinetobacter*, *Arcobacter* and *Trichococcus*) and five families (*Porphyromonadaceae*, *Clostridiaceae*, *Lachnospiracea* and *Ruminococcaceae*) to represent sewer and faecal contamination respectively. These community signatures were successfully used to track the proportions of faecal and sewage communities during a combined sewer outfall event, sewage blending and the following four days of dry weather. While this approach was useful in determining sewer as opposed to faecal communities, the sharing of operational taxonomic units (OTUs) in the faecal signature between human and animal faeces limited this approach to defining a human signature. To differentiate human and non-human sources the SourceTracker software (Knights *et al*., 2011) was used.

Computation methods for source identification and source apportionment

Computational methods used to predict the relative contribution of sources of bacterial communities are popular. However, predicting the contribution of these sources to the overall bacterial community is difficult; while some bacterial taxa are host-associated, a number of taxa are shared between hosts. Random forests algorithms and the SourceTracker (Knights *et al*., 2011) are the most common methods to identify distinct sources and predict their relative contributions. The SourceTracker software takes a Bayesian approach to identify the sources and their relative contribution to ‘sink’ samples. This approach is discussed by Knights *et al.* (2011). Briefly, sink samples are considered as *n* sequences assigned to any one of the source samples or an unknown source. All possible assignments of sequences to each source are considered through the use of Gibbs sampling to integrate over the *posterior* distributions of taxa in the source environment and sources in the test samples, which are both Dirichlet distributions. Gibbs sampling works by firstly randomly assigning each sequence to a source, and estimating the current proportion of each source in the sink. A single sequence is reassigned to a new source with the probability of observing the sequence in the source. Each iteration of this procedure gives a representation of a single sample from the distribution of all possible sequence-source assignments. Large numbers of these iterations allows the variability of this distribution and mixing proportions to be estimated (Knights *et al*., 2011). Knights *et al.* (2011) evaluated SourceTracker against naïve Bayes and random forests approaches by mixing two simulated source communities and concluding that SourceTracker was superior to both, particularly when the communities of two source samples were very similar. Neave *et al.* (2014) used 454-pyrosequencing with SourceTracker to identify human pollution in receiving waters. SourceTracker results generally agreed with and were more sensitive than other source tracking methods (DGGE of enterococci and PCR for faecal markers), despite the predicted proportion of faecal contamination being extremely low. False positive results are a concern when using SourceTracker, particularly for low proportions of bacteria. A separate evaluation of SourceTracker recommended running the algorithm five-times with default settings to identify false positive results with relative standard deviations > 100% (Henry *et al*., 2016), following an analysis of simulated microbial communities. However, this analysis did not include potentially similar sources, such as cow and sheep, which are important for MST in the UK (See 3.2 Agricultural pollution).

The ability of SourceTracker to identify and differentiate sources of pollution depends on the similarity of microbiomes from the same host environment and dissimilarity of those from different source environments. Staley *et al.* (2018) found that source assignments were only accurate when the library was composed of samples considered local to the test samples. However, how local these samples are required to be is unclear since the tested samples were from Australia and the USA (Staley *et al*. 2018). It would be useful to know, for example, whether a single library would be representative of the whole of the North East of England. In addition, it has been suggested that SourceTracker may artificially conflate the background community with the faecal source community overestimating the relative contribution of a source (Staley *et al*. 2017). However, this appears to have little effect on the conclusions drawn from the results, though this may become an increasingly important consideration with decreasing levels of contaminations.

However, when the background community is omitted as a source, considerably greater contamination from inlet and outfall samples was found. Thus, when SourceTracker is evaluated without controlling for the environmental context of faecal pollution, in this case open ocean microbiota, the algorithm may artificially conflate the environmental signature with the faecal source and overestimate the burden of pollution from the source.

While large library-sizes were a requirement and a limitation of previous library-dependent methods, the number of samples required to be representative of source microbiomes is unclear. A number of studies have successfully used libraries composed of a single sample, and while Brown *et al.* (2017) suggested the potential of false negatives in their study when using less than 13 samples, the same authors later suggest that <10 samples are likely to be sufficient (Staley *et al*. 2018). A better understanding of how library size affects the accuracy of predictions is required, since accuracy may decrease with library size, especially when comparing similar sources (Staley *et al*. 2018).

Staley *et al.* (2018) note that several questions remain regarding the temporal stability of libraries, understanding of the decay of faecal communities and an understanding of how SourceTracker predictions relate to regulatory FIO concentrations. In addition, validation of more recent techniques for the analysis of sequence data and generation of bacterial community fingerprints are required since the composition of these fingerprints is affected by changes in: the sequencing platform used (Samarajeewa *et al*., 2015); the quality controls used in processing the sequencing data; and the bioinformatics approaches used to define an OTU. Currently, no study has described the effect of these upstream choices on the SourceTracker output, which may be important in long-term adoption of these techniques for water quality or catchment monitoring. A further consideration for MST is that SourceTracker may underestimate the contributions of highly diverse sources, such as soils due to the lack of overlap between the source community and the sources which may contain rare taxa (Flores *et al*., 2011).

Using the entire bacterial community, like in SourceTracker, may make analysis susceptible to changes in the composition of microbial communities introduced through DNA extraction, PCR and sequencing protocols (Roguet *et al.*, 2018). In this light, a random forests classifier using a narrow taxonomic focus, for the orders *Bacteroidales* and *Clostridiales*, was recently tested. Random forests models consist of numerous decision trees generated using a random subset of the training data, in this case the source communities of *Bacteroidales* and *Clostridiales*. When classifying samples into sources, the classifier reports averages of classifications from each of the decision trees. In evaluating this approach, using pet (cat and dog) and ruminant (sheep and cow) groupings gave lower error rates than using individual sources, possibly due to similarities in their *Bacteroidales* and *Clostridiales* communities. The classifiers appears to be largely specific, although the Cat, Dog and Pet classifiers had the lowest prediction accuracy giving some false positive results and were only identified in samples with higher levels of contamination (>10% of source sequences).

The random forests approach may allow for more rapid identification of pollution sources, but fails to identify the magnitude of contamination easily. Once random forest classifiers are generated, they can be used later to rapidly analyse pollution sources. In comparison, SourceTracker often requires the entire library to be analysed simultaneously (Roguet *et al.,* 2018), although, new bioinformatics approaches which use closed-reference techniques are overcoming this limitation. Roguet *et al.* (2018) overcame the inability of random forests to predict the magnitude of contamination by using the relationship between known proportions of contamination from test samples and the relative proportion of sequences matching all classifiers as a proxy. However, this approach was not tested at low levels of contamination expected from environmental waters (Neave *et al*., 2014), nor compared to the SourceTracker algorithm.

A number of MST investigations have successfully used HTS to generate community signatures from pollution sources and identified these in receiving waters (Newton *et al*., 2013; Neave *et al*., 2014). Newton *et al.* (2013) used the V6V4 and V6 region to produce community signatures, with a signature being composed of the relative abundance and the taxa distribution.

A further limitation of using SourceTracker in MST studies may be in the stability of faecal communities. Sassoubre *et al.,* (2015) note that the HTS approach to MST relies on the temporal stability of the microbial communities contributing to the pollution and it is therefore important to understand the factors affecting these communities. Indeed, in their study Sassoubre *et al.,* (2015) showed that the microbiota of sewage changed significantly with as little as 3% of the OTUs being identified by SourceTracker as originating as sewage after 48h. This is a limitation in the use of HTS in MST, although a community signature approach, such as that used by Newton *et al*. (2013) may overcome this. A signature is a number of OTUs specific to a source and the relative abundance of these OTUs; if the most persistent OTUs with similar persistence patterns are used the reliability of HTS as an MST method may be improved. A number of studies have identified microbial communities from humans and/or sewage (Newton *et al*., 2013, Koskey *et al*., 2014, Sassoubre *et al*., 2015), few have explored the degradation of these communities (Sassoubre *et al*., 2015) and no studies have explored the degradation of communities of faecal material from other animals which may also contribute to bathing water pollution such as dog, cow, horse and sheep etc.

1.4.3 Library-independent methods

Library-independent methods are based on the observation that some organisms (or genetic markers within the DNA of those organisms) exhibit a preference for a particular environment or host. Currently, the perfect host-specific organisms i.e. one that occurs in just a single host species but in every individual of that species, has not been identified (McLellan & Eren, 2014). However, a range of genetic targets which appear to be more prevalent in a particular host species, and are present in enough individuals of that species to be useful, have been identified.

Library-independent methods still dominate MST research and investigations. Compared to library-dependent techniques, library- independent techniques are generally less labour intensive, more cost-efficient, and more rapid in their implementation and analysis. One issue which reduces this rapidity, and adds to costs, is that the performance of library-independent markers varies depending on the geographical location, potential sources of pollution, and environmental conditions (Wuertz *et al*., 2011). Validation of marker performance is therefore required in each new location they are used, reducing the rapidity with which these techniques can be deployed (Wuertz *et al*., 2011; Harwood *et al.*, 2014).

Validation of marker performance

Currently, there are a wide range of MST markers, although, the selection of markers is difficult as few have been thoroughly validated. Selection of markers is further complicated as marker performance indicators can hold a range of values, changing with; the environment, geographical location (Gawler *et al.,* 2007; Wuertz *et al.,* 2011), assay used, and combination of markers selected (Caldwell *et al.,* 2007; Gomi *et al.,* 2014). Wuertz, *et al.,* (2011) suggests preliminary studies are necessary for marker validation even though this detracts from the rapidity of library-independent approaches.

The ideal validation of markers should include an evaluation of their: host-specificity, distribution in host population, temporal and geographical stability, environmental persistence, their correlation to public health risks, and the limits of their detection and quantification sensitivity of the methods of detection (Hagedorm *et al.,* 2011; Harwood *et al.*, 2014). However, these are not always possible due to time, cost and sampling limitations.

The host-specificity of a marker describes the extent the marker may be found in a single host and not found in faecal matter from non-target hosts. Specificity is evaluated by examining faecal matter from non-target hosts and calculating one minus the false positive rate (Stoeckel and Harwood 2007). The sensitivity of a marker is determined by evaluating the proportion of faecal samples from target-hosts that are confirmed as positive (Stoeckel and Harwood 2007).

In evaluating the specificity and sensitivity the USEPA (2005) suggest that ten samples per host type are used. It is worth noting that reported values often use different numbers and/or types of non-target hosts, depending on the likely sources of faecal pollution. The temporal and geographical stability of the specificity and sensitivity of a marker, as well as its environmental persistence and correlation to pathogens are less often tested prior to MST studies. The stability of markers refers to whether the specificity, sensitivity and presence of the marker are consistent over time and geographic location. Many persistence studies have been undertaken, however, it is important that new markers are evaluated for persistence particularly since this could allow the development of catchment models, which may move MST techniques from risk identification to risk prediction.

Currently, comparing marker performance is difficult, due to the vast range of markers and validation methods available, as well as the use of different units in performance measures, make it difficult to choose between markers (Harwood *et al.*, 2014). If MST is to be deployed in the water industry, a cost-effective approach may be to validate markers across the catchment served by a particular water company, although there is little knowledge on the variability of marker performance on local scales.

Current markers for MST investigations

As mentioned, the 16S rRNA gene is the genetic target for the organisms most commonly used in MST studies including: a number of organisms in the order *Bacteroidales*; *Bifidobacteria spp.*; *Firmicutes*; and archaea such as *Methanobrevibacter smithii*

(McLellan & Eren, 2014). Although the 16S rRNA gene is most commonly used, others genetic markers such as: The α-1,6-mannanase gene of *B. thetaiotaomicron*; the *E.faecium* surface protein gene (ESP) and mitochondrial DNA have been used (Yampara-Iquise *et al*., 2008; Harwood *et al*., 2014; McLellan & Eren, 2014).

Bacteroidales

Bacterioidales are the most common target for library-independent MST. *Bacteroidales* are obligate anaerobes which are abundant in the mammalian intestinal tract (Harwood *et al*., 2014) and whilst difficult to culture, they are not environmentally persistent; their presence therefore indicates recent faecal pollution (Bernhard & Field, 2000).

Bernhard and Field (2000) developed the human-specific marker, HF183 which, having received many field tests, has a reported specificity of 60 to 100% and sensitivity of 58.3 to 68% for human faeces and 100% for sewage. It is also worth noting that the concentration of HF183 is generally an order of magnitude higher than that of faecal coliforms (Bernhard & Field, 2000). Many more *Bacteroidales*-based human-faecal markers have been and continue to be developed. Harwood *et al.* (2014) gives a summary of 10 further *Bacteroidales* markers using the 16S rRNA gene target and 4 using other gene targets. However, only five 16S rRNA markers and three other markers have been field tested and only two 16S rRNA markers (HF183 and BacHUM-UCD) have been correlated with pathogen presence (Harwood *et al*., 2014). A number of *Bacteroidales* markers for animals other than humans have been identified and assays developed for their detection, including markers for; canine (BacCAN), bovine (BacCOW) and ruminant (BacR) hosts (Kildare *et al*., 2007; Reischer *et al*., 2006).

Bifidobacterium

Bifidobacterium is a genus of enteric anaerobes abundant in humans (Bonjoch *et al.,* 2004). *B. adolescentis* have been suggested to be specific to the human intestinal tract and therefore potentially useful to track human pollution (Ballesté & Blanch, 2011). A Taqman qPCR assay targeting *B.adolescentis* was developed and although the specificity of this assay was slightly less than the H183 assay (95%), with cross-reactivity occurring in agricultural wastes, this organism may be a useful MST target (Gourmelon *et al*., 2010). Unfortunately, few studies have currently further evaluated this assay so little is known about its geographical or temporal stability.

Methanobrevibacter smithii

A range of qPCR assays for human markers have also been developed to target the archaeon *Methanobrevibacter smithii* (Johnston *et al*., 2010). The nifH methanogen-specific gene, is unique among this group and it does not code for a functional nitrogenase group (Ohkuma *et al.,* 1999). The nifH gene is reported to be present in around 30% of individuals and in 93% of sewage and absent from non-human sources (Specificity = 100%) (Ufnar *et al*., 2006). Subsequent studies have reported some limitations with *M. smithii.* In a comparison of markers, the nifH gene appears to have low abundance in sewage and appears insensitive, particularly when compared to other markers (Ahmed *et al.*, 2012; McQuaig *et al.*, 2012). This led to the conclusion that *M. smithii* may have limited use as a sole marker, although the high specificity may give it some use as part of a multi-marker study (Ahmed *et al.*, 2012; Harwood *et al.*, 2014). However, subsequent studies have also observed *M. smithii* in porcine guts (Federici *et al.*, 2015) and as part of the core methanogen community in bovine samples (Cersosimo *et al*., 2016). In addition, a recent survey of 16S rRNA gene sequences in faecal samples revealed sequences highly similar to *M. smithii* in bovine, ovine and equine samples (Iceton, 2018). While this may suggest the unsuitability of *M. smithii* as a marker (Iceton, 2018), studies reporting cross-reactivity did not target the nifH gene which may prove to represent a human-specific strain. Nevertheless, caution would be advised before using this marker.

Lachnospiraceae

Bacteria of the family *Lachnospiraceae* are anaerobes. These have been of recent interest in MST. A phylotype of *Lachnospiraceace*, Lachno2, looks to be a promising marker, concentrations correlated well with enterococci (r = 0.86) and adenovirus (r = 0.91) suggesting that *Lachnospiraceae* may be as environmentally persistent as enterococci and represent some pathogens well (Newton *et al*., 2011). While the specificity of Lachno 2 was not evaluated, a previous study found no sequences of the genus *Blautia* (a member of the *Lachnospiraceae*) in cattle faeces. However, a comparison of five human-associated markers found Lachno 2 to have the lowest specificity, cross-reacting with 52% of animal faecal samples (Mayer *et al*., 2018). The Lachno 2 assay did, however, have the highest mean concentration in human faecal samples (6.0 Log_{10}(gene copies/100 mL)), supporting the development of other *Lachnospiraceae* based markers. Validation of a Lachno 3 and Lachno 12 are underway which appear to be more specific, although still showed low-levels of cross-reactivity (Feng *et al.,* 2018).

Lachnospiraceae were also recently highlighted as potential faecal markers using oligotyping, a computational method which can discriminate between similar strains (Eren *et al.*, 2013). Eren *et al.* (2014) analysed the V6 region of the 16S rRNA genes of bacteria of genus *Blautia*, to identify high-resolution OTUs termed oligotypes. Whilst most (86%) oligotypes identified were not present in all host faecal matter, 13 oligotypes specific to humans, swine, chicken, deer and cattle were identified. Although these are yet to be tested in MST investigations *Blautia* oligotypes are potential indicators and oligotyping a promising technique to identify further indicators.

Mitochondrial DNA

A number of laboratory and field studies (Martellini *et al.,* 2005; Schill and Mathes, 2008; Baker-Austin *et al.*, 2010; He *et al.*, 2015, 2016; Stea *et al.*, 2015; Villemur *et al.*, 2015) have reported the potential of eukaryotic mitochondrial DNA (mtDNA) as an MST marker. Mitochondrial-DNA can be isolated from faeces of animals, likely a result of the shedding of colonic epithelial cells (Iyengar *et al*., 1991). An advantage of mtDNA is that host DNA is detected directly, rather than using a microbial proxy (Caldwell *et al.,* 2007). This may be particularly useful where no microbial markers exist, e.g. pigeons (Waso *et al.,* 2018). In addition, the utility of mtDNA has been compared to 16S rRNA genes having regions of both highly-conserved and variable sequences, with multiple copies per cell thereby allowing differentiation between sources (Martellini *et al.,* 2005).

In laboratory and field trials mtDNA has achieved mixed results. Martellini *et al.,* (2005) designed PCR assays to differentiate human, bovine, ovine and porcine faecal pollution, no cross-reactivity was found suggesting high specificity of markers. Baker-Austin *et al.* (2010) reported sensitivity to sewage of 85%, lower than that of bacterial markers. This was similar to the performance of mtDNA in other studies. Schill and Mathes (2008) blind tested 20 faecal samples using mtDNA markers from nine pollution sources. The average sensitivity and specificity was 85% and 99% respectively. Field trails generally agree, with human mtDNA marker concentrations being at least one order of magnitude lower than HF183 (Villemur *et al*., 2015; Stea *et al*., 2015).

Aside from low sensitivity, a number of other limitations exist for the use of mtDNA markers, in particular, the presence and concentration of mtDNA may not be indicative or proportional to health risk. Recreational water users may increase the human mtDNA concentration in pristine waters through the shedding of skin epithelial cells. In addition, the gene copies per cell of mtDNA can vary by three orders of magnitude and little is understood about the shedding rates of epithelial cells from the colon. Despite limitations mtDNA markers may be useful in their application alongside bacterial markers due to their high specificities to their hosts.

Viral indicators

The relationship between FIO and health risk is a topic of much debate, and has led to the call for reliable viral indicators, which may better represent health risk from viruses (Marion *et al*., 2014; Shah *et al*., 2011); to monitor the efficiency of wastewater treatment processes and indicate faecal contamination of food and water. Harwood *et al.*, (2013) suggest that non-pathogenic viruses may be useful MST indicators since many are host-specific (McQuaig *et al.,* 2012), have environmental decay rates (Walters *et al.*, 2009), and wastewater treatment removal rates more similar to viral pathogens than culturable FIOs (Symonds *et al*., 2018).

However, viruses are typically present in low concentrations (Kitajima *et al*., 2014; Harwood *et al*., 2013) which can make enumeration difficult, expensive (McQuaig *et al*. 2009), and lead to false negative results (Harwood *et al*., 2013). Nevertheless, the low concentrations may be overcome using concentration techniques (Ahmed, Harwood, *et al*., 2015), and metagenomics analyses (Aw *et al.,* 2014) may identify viruses with higher concentrations. Recently, the pepper mild mottle virus (PMMoV), with high mean abundance and low seasonal variation, has been proposed and tested as an indicator of faecal pollution and treatment process indicator (Rosario *et al*., 2009) and although less abundant than HF183, may better represent viral health risk in environmental waters (Hughes *et al*., 2017).

A large amount of research has been dedicated to the use of bacteriophages as indicators of faecal pollution and a recent review suggests F-specific RNA (FRNAPH) and somatic coliphages are better indicators of viral contamination than current FIOs (USEPA, 2015). Bacteriophage are viruses which use bacteria as hosts and can be measured at low cost using culture-based assays. FRNAPH RNA bacteriophages have been suggested as a model organism for enteric viruses, exhibiting a correlation with viral concentrations in freshwater (Havelaar *et al.,* 1993). Among four classifications of FRNAPH group II and group I appear to be the most promising groups as indicators of human and animal pollution respectively (Table 1.3).

Table 1.2 Examples of some viral indicators used for MST and their relationship to host environments

Viral indicator	Host	Concentration (gene copies/ 100mL)	Sensitivity (%)	Specificity (%)	References
FRNAPH (Group II)	Wastewater		94.3 57.9	89.6 Avian 69.3 Cow	(Gourmelon *et al.*, 2010) (Harwood *et al.*, 2013)
FRNAPH (Group III)	Wastewater		17.4	NA 67	(Gourmelon *et al.*, 2010) (Blanch *et al.*, 2006)
FRNAPH (Group I)	Animal faeces		52.1 – 100	94.1	(Gourmelon *et al.*, 2010)
Bacteriophage infecting *Bacteroides fragilis* GB124	Wastewater Faeces	2.5 x 10^3 – 5 x 10^4 1.0×10^1 to 1.0×10^2 (PFU/g)	100% 4%	100% 100%	(Payan et al., 2005; Ebdon et al., 2012) (Diston & Wicki, 2015)
Bacteriophage infecting *Bacteroides thetaiotaomicron* GA17	Wastewater	3.2 x 10^3 – 1.2 x $10^{5.5}$	100%		(Payan et al., 2005)
Adenoviruses	Wastewater	1.4 x 10^6	91.6		(Kitajima *et al.*, 2014) (Hughes *et al.*, 2017)
polyomaviruses	Wastewater	2.6 x 10^7	58-79		(Kitajima *et al.*, 2014)
Enteroviruses	Wastewater	1.8 x 10^5			(Hughes *et al.*, 2017)
Noroviruses	Wastewater	9.7 x 10^4	75		(Kitajima et al., 2014)
Pepper Mild Mottle Virus (PMMoV)	Wastewater Human faeces	3.7 - 4.4 x 10^5 5.7 x 10^6	98 11.3		(Kitajima *et al.*, 2014) (Hughes *et al.*, 2017)
CrAssphage	Wastewater	6.4 - 9 Log$_{10}$	100	59 92.7 98	(García-Aljaro *et al.*, 2017) (Ahmed *et al.*, 2018) (Stachler *et al.*, 2018)

Other phage targets, such as bacteriophage infecting *Bacteroides spp.* (Tartera *et al*., 1989; Payan *et al.*, 2005; Jofre *et al.*, 2014; Diston & Wicki, 2015) and *Enterococcus spp.* (Purnell *et al.,* 2011) are also promising MST targets. While different *Bacteroides* host strains may be required for MST in different regions, two host strains, *B. thetaiotaomicron* GA17 and B. *fragilis* GB124 (Table 1.2), have shown promise in the UK and Southern Europe (Payan *et al.*, 2005; Blanch *et al.*, 2006; Ebdon *et al*., 2012). Blanch et al., (2006) found bacteriophage infecting *B. thetaiotaomicron GA17* to have greater specificity to wastewater than somatic coliphages, FRNAPH, and bacteriophages infecting Bacteroides fragilis RYC2056. *Enterococcus* hosts, susceptible to lysis from bacteriophage from specific hosts, have been identified, using a tiered approach (Purnell *et al.,* 2011). Interestingly, Purnell *et al.* (2011) noted a negative correlation (R = -0.480, p < 0.01) between the sensitivity and specificity of *Enterococcus* hosts. Strains 100% specific to cattle and pig faeces were only 33% and 20% sensitive, respectively, and an *E. faecium* isolate, MW47, had a specificity of 100% and sensitivity to raw and treated wastewater of 100% and 25% respectively. This suggests that a trade-off between these performance parameters may be necessary when selecting MST markers. Wangkahad *et al.* (2017) also isolated two strains of *E. faecalis*, which appear to be 100% specific and 90% sensitive to sewage. Further work to understand the environmental survival of MW47 suggests that caution is required when using phage-based MST due to differential die-off between different phage families (Purnell *et al*. 2018).

Recent metagenomic monitoring of wastewater suggests that there is an array of uncharacterized viruses which may provide better human and/or animal specific markers in the future (Aw *et al.,* 2014). While metagenomic detection and monitoring of viromes are in their infancy, metagenomics approaches have produced new viral indicators. Metagenomic assembly of virus genomes led to the recent discovery of crAssphage, a bacteriophage which appears to be 6-times more abundant in metagenomes than all other phages together (Dutilh *et al*., 2014). Stachler and Bibby, (2014) first suggested the utility of crAssphage as an MST marker, following an evaluation of 86 publically available metagenome data sets. CrAssphage appeared to be more abundant in sewage from the US and Europe, compared with that from Asia and Africa and had low cross-reactivity with other samples, with the exception of bat guano. The development of two new primer pairs targeting crAssphage appear to be better than previous markers, although still show cross-reactivity with dog and gull faeces through qPCR assays (Stachler *et al*., 2017). It is important to note that crAssphage was also observed in 41% of animal faecal samples tested which is likely to limit its efficacy as a human MST marker (García-Aljaro *et al*., 2017).

Relating MST outcomes to regulatory indicators

It is important that conclusions drawn by MST assays relate to regulatory FIO concentrations, since in bathing water catchments FIO compliance is the main management objective, although this rarely occurs. Difficulties in relating MST marker and FIO concentration are well appreciated. Whilst library-independent genetic markers give indications of FIO sources they only occasionally explain a high proportion of *E.coli* variation (Reischer *et al*., 2008; Heaney *et al*., 2015), and drawing FIO source conclusions from marker data remains difficult (Wang *et al*., 2013). These difficulties are to be expected since genetic markers and FIO differ in many respects: their initial faecal concentrations; transport and attenuation mechanisms (Johnston *et al*., 2010); their environmental decay rates both within faeces (Oladeinde *et al*., 2014) and in environmental waters (Brown & Boehm, 2015; Wanjugi *et al*., 2016; Korajkic *et al*., 2014); the assays used for their detection (Ahmed *et al.*, 2015); and the ubiquity of FIO and possible environmental persistence as opposed to the host-specificity of genetic markers.

Modelling efforts to relate FIO and MST marker concentrations, so far, are not applicable for a range of environmental conditions. Wang *et al.*, (2013) developed a ratio model to determine the proportion of FIO originating from sewage (F) using genetic marker concentrations,

$$F = \left[\frac{R_{aw}}{R_{sewage}}\right] * \ e^{(t * \Delta k)} \qquad \textit{Equation 1.1}$$

Where R_{aw} is the ratio of genetic-marker to FIO (e.g., enterococci) concentrations in ambient water, R_{sewage} is the same ratio in raw sewage, Δk is the difference in first order decay rate constants between the genetic-marker and FIO of choice, and t is the time spent in environmental waters. Determining the correct values of Δk and t is, however, challenging. The age of the faecal contamination is often impossible to determine (Mattioli *et al*., 2016) and estimates of decay rate constants show a large variation, only some of which is explained by different environmental conditions (Brooks & Field, 2016). Nevertheless, Mattioli *et al.* (2016), note that where ΔK = 0, the model simplifies to:

$$F=[Rw/Rsewage] \qquad \textit{Equation 1.2}$$

However, Mattioli *et al.*, (2016) note that when determining the fraction of culturable FIO (F) from sewage, ΔK ≈ 0 only occurs when light intensity is adequately low. The authors estimated that at a latitude of 37.50 degrees low light conditions only occurred at depths of >100 m or >1 m in mixed and unmixed water columns, respectively. While these depths may be less in temperate climates, Equation 2 is unlikely to be applicable in the majority of circumstances.

E.coli as a source indicator

A range of approaches and techniques has been used to attempt to apportion *E.coli* by source. Library-dependent methods were the first developed and field tested, although these methods have been noted to be overly complex, preventing source identification (Neave *et al*., 2014). This is possibly due to the large diversity in *E.coli* a result of large variations in genome size (Bergthorsson & Ochman, 1998).

The distribution of *E.coli* phylogenetic groups is not identical between different host species or environments. Whilst *E.coli* typically belong to phylogenetic groups A, B1, B2, D, E or F, studies suggest an affinity of B2 *E.coli* strains to the human intestinal tract (Bailey *et al*., 2010). Further analysis of *E.coli* phylogenetic group and subgroup distributions among animal faeces also found the $B2_3$ subgroup to be human-preferred; unfortunately this subgroup showed a low sensitivity with only 7% of human *E.coli* clones analysed belonging to this subgroup (Carlos *et al*., 2010). Whilst some host preference can be inferred from the phylogenetic grouping of *E.coli* strains, the sensitivity and specificity are too low to be useful for source tracking studies.

While library-dependent approaches rely on identifying genotypes occupying different hosts or environments, highly similar genotypes may occupy different hosts (Naziri *et al*., 2016) and environments (Byappanahalli *et al*., 2012). Subtle variations in the *E.coli* genome allow highly similar strains to exist in the differing conditions (such as nutrient concentrations, pH, temperature, predation, UV) and therefore hosts and environments. Recently, with improvements in molecular methods, a number of studies (Gomi *et al*.

2014; Warish *et al*. 2015; Deng *et al*. 2015; Zhi *et al*. 2015) have made use of these variations in the *E.coli* genome to identify specific genes or variants which allow *E. coli* to survive in a range of mammalian intestinal tracts. There have been two main approaches to identifying host-specific *E.coli*: i) using genes in the accessory genome (Gomi *et al*., 2014), and ii) using single nucleotide polymorphisms (Deng *et al*., 2015; Zhi *et al*., 2015).

The *E.coli* core genome may be composed of as little as 10% of the 15,741 to 16,373 gene families which make up the pan-genome (Rasko *et al*., 2008; Lukjancenko *et al.*, 2010). DNA acquisition is therefore a likely strategy for *E.coli* to occupy a host-specific niche. Two studies have identified host-associated genes in the accessory genome (Luo *et al*., 2011; Gomi *et al*., 2014) both using whole genome sequencing (WGS) of *E.coli* isolates through high throughput sequencing (HTS). Gomi *et al.* (2014) identified host-associated and preferred genes in human, bovine, porcine and chicken derived *E.coli* in Japan. Four human-associated *E.coli* markers H8, H12, H14 and H24 appeared useful for MST (Table 1.3). In a catchment study, 47.9% of *E. coli* isolates were allocated a source, although, 4.4% of isolates contained 2 marker genes from different sources. Whilst this is a promising approach, 52% of environmental *E. coli* isolates remained unclassified. This could be due to: the limited number of sources originally tested; the unclassified samples being of environmental origin or naturalised; or that a limited proportion of the *E.coli* diversity from each source was sampled (Gomi *et al*., 2014). To improve the number of isolates classified it may be necessary to sample a greater number of sources as well as a greater number of isolates to sample a greater diversity of each strain and explore the genomes of environmental *E.coli*. These human markers (H8, H12, H14, H24) were subsequently tested in Australia and only H8 and H12 had specificities >85%, although, the authors note the possibility of human contamination in cow runoff leading to reduced predicted specificities(Ahmed, et al., 2015).

The identification of single nucleotide polymorphisms (SNPs) may also prove to be a rapid and effective method of determining sources of host-specific *E.coli* host origins. Luo *et al.* (2011) identified 84 and 120 genes more prevalent in environmentally derived and enteric *E.coli* strains, respectively. Targeting the enteric specific glucosyltransferase gene (ycjM) Deng *et al*., (2014) identified local sequence changes within the *ycjM* gene, allowed the identification of a human preferred genotype H-*yjcM* (Deng *et al*. 2015). The H-*yjcM* genotype was found be present in ~50% of *E.coli* (Deng *et al*. 2015), although the single study using this marker suggests it is much less prevalent (Kataržytė *et al*., 2018).

Table 1.3. Summary of performance of current E.coli biomarkers from recent studies.

Host	Biomarker(s)	Sensitivity	Specificity	Identification/ validation method	Reference
Human	H8	50%	99%	WGS/ Multiplex PCR	Japan (Gomi *et al*., 2014) $^{¥}$Australia (Warish *et al*., 2015)
		45%$^{¥}$	94%$^{¥}$		
	H12	30%	100%		
		14%$^{¥}$	85%$^{¥}$		
	H14	30%	98.%		
			72%$^{¥}$		
	H24	37%	99%		
			57%$^{¥}$		
	At least one marker	66.7%	97%		
Human	human-*ycj*M marker	53%	99.7%	PCR and qPCR	(Deng *et al*. 2015)
Human	SNP pattern in *asnS-ompF*, *uspC-flhDC*, *csgBAC-csgDEFG* intergenic regions.	60%	98%	PCR and Sequencing	(Zhi, *et al*., 2016b)
Human	*ydeR-yedS*	56%	99%	Predicted using *E.coli* Database	(Zhi, *et al*., 2016b)

Sewage	IS30	100% - Chlorine treated wastewater (59% of chlorine treated isolates) 94% - secondary treated wastewater 75% - UV-treated wastewater	97% - Surface water 95% Groundwater	PCR	(Zhi, *et al.*, 2016a)
Cow	Co2 Co3 Both Markers	20% 23.3% 30%	100% 100% 100%	Multiplex PCR	(Gomi *et al.*, 2014)
Cow	SNP pattern in *asnS-ompF*, *uspC-flhDC*, *csgBAC-csgDEFG* intergenic regions.	40%	100%	Predicted from *E.coli* Database	(Zhi *et al.*, 2015)
Cow	*cutC-torYZ*	30%	98%	Predicted from *E.coli* Database	(Zhi, *et al.*, 2016b)
Cow	*ydeR-yedS*	92%	98%	Predicted from *E.coli* Database	(Zhi, *et al.*, 2016b)
Chicken	Ch7 Ch9	76.7% 70%	100% 98.9%	Multiplex PCR	(Gomi *et al.*, 2014)

	Ch12	66.7%	96.7%		
	Ch13	56.7%	95.6%		
	All markers	80%	73.3%		
Chicken	SNP pattern in *asnS-ompF*, *uspC-flhDC*, *csgBAC-csgDEFG* intergenic regions.	54%	99%	Predicted from *E.coli* Database	(Zhi *et al.*, 2015)
Pig	P1	16.7%	100%	Multiplex PCR	(Gomi *et al.*, 2014)
	P3	13.3%	100%		
	P4	13.3%	100%		
	All Markers	30%	100%		
Pig	SNP pattern in *asnS-ompF*, *uspC-flhDC*, *csgBAC-csgDEF*G intergenic regions	79%	97%	Predicted from *E.coli* Database	(Zhi *et al.*, 2015)
Pig	*uspC-flhDC*	74%	97%	Predicted from *E.coli* Database	(Zhi *et al.*, 2015)
Dog	*uspC-flhDC*	60%	95%	Predicted from *E.coli* Database	(Zhi *et al.*, 2015)
Dog	SNP pattern in *asnS-ompF*, *uspC-flhDC*, *csgBAC-csgDEFG* intergenic regions	63%	93%	Predicted from *E.coli* Database	(Zhi *et al.*, 2016b)

Horse	SNP pattern in *asnS-ompF*, *uspC-flhDC*, *csgBAC-csgDEFG* intergenic regions	36%	99%	Predicted from *E.coli* Database	(Zhi *et al.*, 2016b)
Sheep	SNP pattern in *asnS-ompF*, *uspC-flhDC*, *csgBAC-csgDEFG* intergenic regions.	46%	99%	Predicted from *E.coli* Database	(Zhi *et al.* 2016b)
Gull	SNP pattern in *asnS-ompF*, *uspC-flhDC*, *csgBAC-csgDEFG* intergenic regions	5%	99%	Predicted from *E.coli* Database	(Zhi *et al.*, 2016b)

Zhi *et al.* (2015) hypothesised that evolution of the regulatory transcriptome is a likely mechanism for host-speciation; since regulation of core genes allows phenotypic adaptation to adverse environments (Ziebuhr *et al.*, 1999; Prüß *et al.*, 2006), and gene regulation can be altered through mutations in promoter sequences (Ando *et al.*, 2011). Host-associated SNPs patterns were identified using logic regression analysis in the sequenced and concatenated intergenic regions (ITGRs) between the *uspC* and *flhDC, csgBAC* and *csgDEFG,* and *asnS* and *ompF* genes (Zhi *et al*., 2015) and later between the *cutC-torYZ, metQ-rcsF* and *araH-otsB* (Zhi, *et al.*, 2016) (Table 1.3). This logistic regression approach was subsequently applied to a database compiled from publically available *E.coli* genomes, allowing the rapid assessment of ninety ITGRs and identification of ITGRs harbouring host-specific information (Zhi, *et al.*, 2016). The logic regression method was noted as highlighting more robust associations with *E.coli* isolates than phylogeny based approaches did. This approach identified a human-specific SNP pattern present in 53% of *E.coli* isolates contained. The sensitivities for SNP patterns for other host animals ranged from 31 to 94%, although, all patterns showed a very high specificity (>96%). This variation in sensitivity may be explained by host specialist strains co-existing alongside generalist strains, and hosts, which share the same ecological niche, increasing the potential to share generalist strains. There is also the possibility that other genetic regions may hold more host-specific variations (Zhi *et al.,* 2015). With increasing access to complete and draft genomes on public databases, the use of genomic databases may be a key approach to future biomarker discovery.

While a logical regression approach highlights host-specific patterns, the results are not applicable for the current suite of more rapid assays used in MST. The requirement to isolate and sequence multiple intergenic regions from every *E.coli* would be highly labour intensive and costly. The logistic-regression derived markers may therefore be more useful in a clinical setting, however, other biomarkers such as H8-H24 may be useful in MST, although Gomi *et al*. (2014) reported difficulties in the use of qPCR to identify markers from environmental waters.

1.4.4 Limitations of library-independent MST

There are a number of limitations to library-independent MST. As yet, no truly host-specific genetic marker has been identified (McLellan & Eren, 2014). There must always, therefore, be a disclaimer associated with conclusions drawn from genetic markers.

Correlation with FIO and pathogen presence may be low (Kildare *et al*., 2007), specificity of markers can vary (Gawler *et al.*, 2007; Wuertz *et al.*, 2011) and evidence of extra-intestinal growth of *Bacteroidales* in poultry litter has been observed (Weidhaas *et al*., 2015). Therefore combinations of other markers, or MST approaches may be required.

1.5 MST method selection and application

After almost three decades of research and field testing, there is no consensus on the most appropriate approach or microbial target for MST applications (Hagedorm *et al.,* 2011). This could be due to: the range of potential applications and MST approaches available; the use of individual techniques to try and solve a problem; the limited field testing of MST methods; different performance criteria used in separate studies (Hagedorn *et al*., 2011); or, the lack of data on the costs and benefits associated with MST. A further limitation is the differences in the protocols and data-analysis techniques, such as how to deal with samples below the limit of quantification, led to different outcomes between laboratories using the same MST approach (Stewart et al., 2013; Cao et al., 2018).

This lack of consensus may go some way to explain why the use of MST has been limited in the UK, particularly for public health and water quality monitoring and management. Indeed, a review of techniques to identify, monitor and control urban diffuse pollution in the UK fails to even mention MST as a potential option (Lundy & Wade, 2013). For the UK water industry in particular it is, therefore, paramount that methods are robustly tested using similar performance criteria, tested in ‘real-world’ studies and evaluated for their economic and business related outputs.

MST investigations typically involve either a tiered or toolbox approach, or a combination of both. A toolbox approach involves the use of a range of techniques on the same set of samples, such as a mixture of single-markers and/or whole community-based approaches. A number of studies have used a tiered approach, which has minimised costs through the use of inexpensive assays to determine sites where more expensive assays could then be applied (Griffith *et al*. 2013; Ahmed *et al*. 2015). A toolbox approach may be necessary where cross-reactivity with other sources or low levels of pollution are expected as marker concentrations may be below the limit of detection in environmental samples. However, a balance is required since uncertainty in metrics may lead to a

Box 1. Tiered approach to microbial source tracking (Griffith *et al*., 2013).

1) Develop a list of potential faecal contamination sources through: Maps, interviewing relevant local experts, and visual inspections.

2) Analyse available FIO monitoring data for spatial and temporal trends to help identify conditions that result in elevated FIO levels and determine linkages to the greatest potential sources of faecal contamination.

3) Where leakage from a sanitary system is a potential source, investigate it using traditional tools such as smoke testing, dye testing, or camera inspection.

4) Where human sources are a potential contributor, test ambient waters for human source specific genetic markers (even if traditional tools have not identified a leaking sanitary system). Place high priority on either detecting or confirming a human faecal source, as this source may pose the greatest relative health risk.

5) Where human sources have been accounted for and the relative human loadings are better understood, and/or a likely animal faecal pollution source (e.g., runoff from a horse farm) has been identified, test ambient waters using non-human (animal) source-specific genetic markers.

6) Where source-specific genetic markers have yet to be developed for the suspected source(s), consider testing ambient waters using genetic community analysis methods.

toolbox approach, which over-compensates by using unnecessary methods and becomes inefficient and costly. A tiered approach can be less costly since methods are employed in stages, starting with the lowest cost methods. Enough information to draw robust conclusions may, therefore, be gleaned from low cost methods and negate the need for further analysis (Cao *et al.*, 2013; Walker *et al.*, 2015).

An example of a tiered approach is suggested by Griffith *et al.* (2013) (Box 1). While this exact framework is unlikely to work for all situations in the UK, an evaluation of forty-one MST assays highlighted the importance of establishing such frameworks for MST methods and risk assessments to establish consistent methods and methodologies (Boehm *et al*., 2013).

1.6 Aims and objectives

The overall aim of this research is to evaluate the performance of two emerging MST techniques, *E. coli* biomarkers and community analysis, and assess the feasibility of their incorporation into workflows for Northumbrian Water to carry out MST investigations, with a particular focus on diffuse and low levels of human pollution, through case studies.

Objective 1. Compare the performance of *E.coli* biomarkers and community analysis using high-throughput sequencing to identify sources of human pollution.

The performance of the *E.coli* biomarkers with the most potential for MST (H8, H12, H14 and H24) have not been evaluated outside of the Indo-Australasian region, nor have their results been compared to other source tracking methods. Northumbrian Water are particularly interested in linking the high throughput nature of new community-analysis techniques with regulatory methods.

Research questions:

1a. Are the H8, H12, H14 and H24 *E.coli* biomarkers suitable for the detection of human pollution in the North of England? (Chapter 3 and 4)

1b. Can we use current regulatory methods that assess water quality to detect *E.coli* biomarkers for MST? (Chapters 3 and 4)

1c. Does community-based MST sequencing support the results and conclusions drawn by *E.coli* analysis? (This objective will be answered through case studies in chapter 3 and 6)

Objective 2: Assess the performance of *E.coli* biomarkers in the North East of England.

The performance of the *E.coli* biomarkers developed by Gomi *et al.*, (2014) during the first catchment study (Chapter 3) was markedly different to previous studies (Gomi *et al*., 2014; Warish *et al*., 2015). A database was curated and interrogated to elucidate the expected performance of these biomarkers and develop new *E.coli* biomarkers in a cost-effective manner. This tried to answer the following questions, which arose as a result of the initial catchment study.

Research questions:

2a. Do better markers exist which are specific to *E.coli* to track sewage pollution in the North East of England? (Chapter 4)

2b. Does the concentration of markers in sewage vary between different communities? Is marker concentration representative of total *E.coli* concentration? I.e. is the sensitivity stable between communities? (Chapter 4)

Objective 3: Evaluate the performance of HTS community analysis to discriminate between common sources of pollution in UK catchments and develop a robust method for Northumbrian Water to carry out MST using HTS community analysis in the North East of England.

As yet no studies have assessed the ability of HTS to distinguish between similar sources, which affect UK water quality such as that from ruminants like cows, horses and sheep. This is particularly important since currently no suitable markers exist for some of these individual sources (Boehm *et al*., 2013). There also remain questions as to the appropriate size of the library and whether it is possible to use a generalised library of the whole of a particular region, like North East England.

Research questions:

3a. Can a single database of faecal sources for a particular region be built? (Chapter 6)

3b. Can current methods distinguish between sources that share similar taxa? (Chapter 5)

Objective 4: Identify the most cost-effective way to integrate MST methods into Northumbrian Water operations.

Northumbrian Water have expressed an interest in undertaking MST studies as a result of the case studies undertaken during this project. Identifying the socio-economic and environmental opportunities and benefits of MST studies can help to build and evaluate the business case to establish MST as a routine technique used by the wider UK water industry.

Research questions:

5a. What is the most cost-effective way to implement MST methods into Northumbrian Water’s current operations? (Chapter 7 using case studies in chapter 3 and 6)

5b. What economic and environmental benefits can MST bring? (Chapter 6)

5c. What are the areas where MST will have the largest benefit in the future? (Chapter 6)

Chapter 2 Methods

2.1 Sample collection

2.1.1 Sample preservation and transport

Following the collection of all environmental (river and sea water) and faecal samples, samples were kept on ice in a cool box during transportation and always arrived at the lab in less than 3 hours but usually in less than 1 hour. This was deemed acceptable since the storage duration (1 to 14 days) and temperature (between 20 ^{o}C and -80 ^{o}C) was found to have little effect on the microbial composition analysed by 16S rRNA gene sequencing (Lauber *et al*., 2010; Tedjo *et al*., 2015). Once at the laboratory, samples were kept at 4 ^{o}C until processing, which occurred in under 3 hours.

2.1.2 Faecal samples

Individual faecal samples were collected using a sterile spatula in a 250 mL sterile container. Care was taken to take the sample from the part of the faeces not in contact with the soil, although this was particularly difficult with some types of faeces; cow faeces often have a low solids content and are therefore difficult to collect without disturbing the ground, horse and pig faeces (when indoors) are difficult to separate from hay used as bedding in the stables or paddocks. Samples were transported back to the lab and stored at 4 ^{o}C and DNA was extracted within 24 hours of sample collection.

Faecal swabs, for the culture of *E.coli*, were taken from the centre of each faeces to minimize the likelihood of contamination. It should be noted that gull faeces were particularly difficult to collect without contact of the surface from which it was collected.

2.1.3 Raw sewage

All sewage samples were collected post-screen from municipal wastewater treatment plants (WWTPs) and septic tanks using a 1 L sampling bucket. The bucket was rinsed three times prior to sampling and 2 x 2.5 L samples were taken 15 minutes apart and mixed to make a composite sample.

2.1.4 Environmental water samples

All river, sea, CSO and land surface drain (LSD) samples were collected in 1 L pre-sterilised polyethylene containers which were acid-washed with a 1% HCl solution and autoclaved prior to sample collection according to the Bathing Water Directive Annex V (BWD, 2006/7/EC). A single 750 mL river water sample was collected without disturbing the sediment using a telescopic sampling pole. When possible, samples were taken from ~30 cm below the surface, in the centre of the river, by inserting the sampling pot inverted into the water.

Seawater samples collected during the bathing water season were collected by the EA's sampling team, at the same time as their own regulatory sample, according to Annex V of the BWD (2006/7/EC). Briefly, one litre of water was collected from a depth of ~30 cm below the surface of the water at a point that is at least one metre deep, using pre-sterilized, 1 L disposable polyethylene sample bottles. For samples taken outside of the bathing water season, I took seawater samples using an identical process.

Bathing water samples collected by the EA were transported and stored at 4 ^{o}C at the Newcastle EA site (Tyneside House, Newcastle, UK). These were collected within 2 h and transported to Newcastle University (< 15 minutes) and usually processed within 30 minutes.

2.2 Enumeration of faecal indicator organisms

E.coli were enumerated through membrane filtration according to the bathing water directive (2006/7/EC, CEU, 2006) and international standard ISO 7899-2 (ISO, 2000). After discussion with the EA these protocols were updated to be identical to EA protocols which use tryptone bile x-glucuronide (TBX) chromogenic agar (Oxoid, UK) as the culture media for *E.coli*.

Four volumes of river and seawater sample, 0.1 mL, 1 mL, 10 mL and 100 mL were filtered with 10 mL of PBS solution for the 0.1 and 1 mL volumes, at least in duplicate, through hydrophilic mixed cellulose ester membrane filters with a 0.45 µm pore size (Pall Laboratories, UK). For enumeration of *E.coli*, following filtration, membrane filters were incubated on TBX agar (Oxoid, UK) at 37 ^{o}C for 4-6 h and 44 ^{o}C for 18-20 h. Plates with between 5 and 100 colony forming unit (CFU) were counted and recorded.

2.3 DNA extraction, storage and quality control

2.3.1 E.coli isolates suspended in lysogeny broth

Individual *E.coli* were isolated from faeces by spreading faecal swabs onto tryptone bile x-glucuronide (TBX) agar (Oxoid, UK) and incubating at 37 ^{o}C overnight. A single, green colony was picked from the incubated plate using a sterile needle and spread again onto TBX agar and incubated at 37 ^{o}C. This was repeated 2 more times before a sterile needle was used to inoculate 8 mL of lysogeny broth (Thermo Fisher, UK). The inoculated broth was incubated at 37 ^{o}C. DNA was extracted from 0.2 mL of an overnight (16-18 h) cell culture using the Wizard Genomic DNA Purification Kit (Promega, UK) according to the manufacturer's instructions for Gram-negative bacteria.

For storage, 1 mL of overnight culture was added to 1 mL of autoclaved, 30% glycerol (v/v) solution, vortexed and stored at -80 ^{o}C.

2.3.2 Direct extraction from E.coli isolated from plate counts

E.coli colonies were picked and placed in 50 µL of DNA/RNA free water (Thermo Fisher, UK) and incubated at 95 ^{o}C for 15 minutes. Solutions were then centrifuged at 14,000 × g for 5 minutes at 21 ^{o}C, and the supernatant recovered to remove any cell debris. The recovered DNA was stored at -80 ^{o}C until further use.

2.3.3 Extraction of DNA from faecal samples

DNA was extracted directly from 150 – 300 mg of fresh faecal samples using the FastSpin kit for faeces (MP Biomedicals, USA), with a modification to include 4 cycles at 60 m s^{-1} for 40 seconds, suggested to increase the DNA yield (Albertsen *et al*., 2015). This kit was used as this was the current method used in the laboratory; additionally, it was used by a previous study (Iceton, 2018) that may provide a useful comparison to some analysis in this thesis. It was therefore beneficial to keep the methods as consistent as possible.

DNA was extracted from between 10 mL and 100 mL of post-screen sewage, depending on the concentration, and filtered through hydrophilic mixed cellulose esters membrane filters with a 0.22 µLm pore size (Pall Laboratories, UK). Filters were folded 5-times and stored in 2 mL Eppendorf tubes at -80 ^{o}C until use. DNA was extracted directly from torn membrane filters using the DNA FastSpin kit for soil (MP Biomedicals, USA) as per the manufacturer's instructions with the following modifications. Torn filters were placed into the E-lysis tube with 898 µL and 110 µL of sodium phosphate buffer and MT buffer and subject to 4 cycles at 60 m s^{-1} for 40 seconds. To avoid contamination of the final DNA, an additional ethanol wash-step was included as recommended by the manufacture for samples with a high organic content.

2.3.4 DNA extraction from environmental waters

The volumes of river and seawater samples processed were typically 250 mL for river and 800 mL for seawater, however, it was necessary to reduce this for very turbid samples to 100 mL and 500 mL, respectively. These volumes are larger than those typically taken (Mattioli *et al.*, 2016; Iceton, 2018), however, larger volume sizes may reduce uncertainty in downstream analysis used in community analysis (Mattioli *et al*., 2016).

Sea and river water samples were filtered through hydrophilic mixed cellulose esters membrane filters with a 0.22 µL pore size (Pall Laboratories, UK) and folded in half five times and stored in a 2 mL Eppendorf tube at -80 ^{o}C until use. DNA was extracted using the DNA FastSpin kit for soil as per the manufacturer's instructions with the following modifications. Filters were torn and placed into the E-lysis tube with 898 µL and 110 µL of sodium phosphate buffer and MT buffer added to the E-lysis tube. For environmental samples, 5 µL of 10 µg µL^{-1} Salmon Sperm DNA was added to the lysis buffer to act as an internal control for DNA extraction and qPCR (Haugland *et al*., 2005). The lysis steps were modified to include four cycles at 60 m s^{-1} for 40 seconds, suggested to increase the DNA yield (Albertsen *et al*., 2015). The recovered DNA was stored at -80 ^{o}C until used.

2.3.5 DNA quality control

DNA quality was assessed using a Nanodrop 1000 (Thermo Fisher, UK) according to the manufacturer's instructions. DNA quantity was also assessed using the Qubit high sensitivity, double-stranded DNA kit (Thermo Fisher, UK) without modification to the manufacturer's instructions using a Qubit 2.0 fluorimeter (Life Technologies, Carlsbad USA).

2.4 PCR and qPCR assays

2.4.1 Polymerase Chain Reaction (PCR)

Amplification of DNA fragments was carried out using the Fast Start High Fidelity PCR System (Sigma-Aldrich, UK) according to the manufacturer's protocol. Briefly, each 25 µL reaction contained: 1.8 mM $MgCl_2$, 0.2 – 0.4 µM of each primer, 200 µM of each dNTP, 2.5 U of high fidelity enzyme blend, and 5-15 ng of DNA. PCR reactions were carried out in a PCR Max AC1 thermo-cycler (PCR Max, UK) using the protocol in table 2.1.

Table 2.1. PCR cycling conditions used throughout this study.

Step	Number of cycles	Time	Temperature
Initial Denaturation	1	2 minutes	95 °C
Denaturation Annealing Elongation	35	30 seconds 30 seconds 1 minute	95 °C Variable 53-60 °C 72 °C
Final elongation	1	7 minutes	72 °C
Cooling	1	No limit	4 °C

2.4.2 Identification of E.coli biomarkers using multiplex PCR

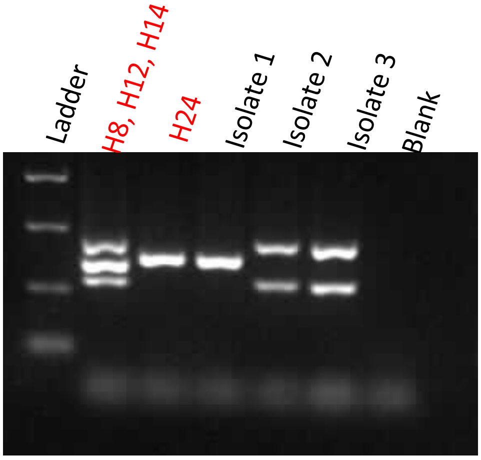


Figure 2.1. End-point PCR and visualization by agarose electrophoresis

A multiplex PCR of human-specific *E. coli* markers H8, H12, H14 and H24 using the primers designed for them (Gomi *et al.*, 2014) was optimised by altering the annealing temperature and $MgCl_2$ concentration. The reaction mixture for the FastStart, High Fidelity PCR kit (Sigma-Aldrich) was modified from that above to contain 2.8 mM $MgCl_2$ and 0.2 µM of each primer. DNA was replaced with water in the negative control and standards kindly supplied by Gomi *et al.* (2014) in the positive controls (H8, H12, H14 in one control and H24 in another) (Figure

2.1). Reactions were performed as described above (Table 2.1) with an annealing temperature of 60 ^{o}C. DNA in the reaction was added at amounts of less than 30 ng, as amounts greater than 30 ng resulted in unwanted amplification, which was difficult to differentiate from the H8 and H14 markers. PCR products were run on a 1.5% agarose gel, stained with Nancy-520 DNA stain (SigmaAldrich, Bumbleby, Ukraine) and visualized using a Dual-Intensity Transilluminator (UVP, USA).

PCR reactions which were positive for a human marker were assumed to be inhibition free. All isolates negative for human markers were subject to a further PCR targeting the *RodA* gene (Primers in table 2.4), as described above (1.1 Polymerase Chain Reaction (PCR)), present in most *E.coli,* as described previously (Chern, *et al*., 2011) to ensure that a negative result was not due to problems with DNA extraction or inhibition of the PCR reaction.

2.4.3 Quantitative PCR (qPCR)

Quantitative PCR reactions were carried out using the SsoAdvanced universal SYBR green supermix (Bio-Rad, UK) chemistry according to the manufacturer's instructions. Each 10 µL reaction contained 300-500 nM of forward and reverse primers, 1 Unit of SsoAdvanced universal SYBR Green supermix. All reactions were carried out in clear-welled, 96-well plates (Bio-Rad, UK) on a CFX96 Real-Time PCR Detection System (Bio-Rad, UK) and cycled through the following conditions: 95 ^{o}C for 30 seconds followed by 37 cycles of 95 ^{o}C for 10 seconds, and 30 seconds at 60 ^{o}C (all markers used for qPCR had the same annealing temperature). A melt-curve analysis was undertaken after each qPCR run by increasing the temperature from 65 ^{o}C to 95 ^{o}C in 0.5 ^{o}C increments for 5 seconds per step.

Primer concentrations were optimized by performing qPCR reactions with varied mixtures of forward and reverse primer concentrations of: 300 nM, 500 nM, 900 nM. Standard curves ranging from 2 x 10^{0} – 2 x 10^{6} gene copies per reaction, and a negative control comprising DNA-free water, were run in triplicate for each qPCR run. Standards were made from stock solutions prior to each qPCR run. Standards were constructed by extracting DNA from an overnight culture of *E.coli* in LB broth (1.3.1 *E.coli* isolates suspended in LB broth), determining DNA concentration (1.3.5 DNA quality control) and using equation 1 to calculate the number of gene copies (gc). For *E.coli* associated genes, the total weight of DNA was assumed to be 5.51 x 10^{-15} g, the average length of an *E.coli* genome, 5.4 x 10^{6} base pairs multiplied by the average weight of a base 1.02 x 10^{-21} g/molecule.

Equation 2.1

$$gene\ copies = \frac{Concentration\ of\ extracted\ genomic\ DNA\ (\frac{g}{\mu l})}{Total\ weight\ of\ genomic\ DNA\ (\frac{g}{molecule})}$$

The estimates of the gene copies in the stock solution (Equation 2.1) were checked by enumeration of the *E.coli* culture prior to DNA extraction (1.1 Enumeration of faecal indicator bacteria). Stock solutions of standards were made by diluting the DNA to 1x10^{8} gc *RodA* µL^{-1}. Multiple stock concentrations were stored at -80°C and one working stock solution was stored at -20 °C.

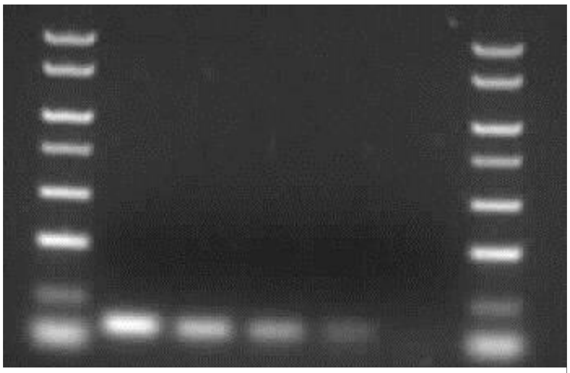


Figure 2.2. Sketa test with salmon sperm approximate concentrations (from left to right) of 50 ng, 5 ng, 0.5 ng, 0.05 ng

As a processing control 50 ng of salmon DNA (Thermo Fisher, UK) was added to the lysis buffer prior to DNA extraction of environmental samples (1.3.4). 50 ng was added as this was easily detectable through PCR based assays (Figure 2.2) and any reduction would likely be from inefficiencies in DNA extraction methods. A salmon DNA blank sample was prepared using DNA- free water and processed using the same method as the environmental samples. QPCR reactions using the sketa primers (Table 2.2) targeting the salmon DNA were used as an internal control (Haugland *et al.*, 2005).

Table 2.2. Sketa primers for the detection of Salmon DNA used as an internal control in qPCR

Primer	**Sequence**	**GenBank reference**	**Reference**
Sketa forward	GGTTTCCGCAGCTGGG	AF170538 (23–38)	(Haugland *et al.*, 2005)
Sketa reverse	AGTCGCAGGCGGCCACCGT	AF170538 (41–59)	

A qPCR result where the cycle threshold value was three units greater than that of the salmon DNA blank sample indicated a problem with either the DNA extraction procedure or the presence of inhibitory substances (Chern *et al.*, 2011; Haugland *et al.*, 2005, 2010).

Where a problem was indicated, samples were checked for inhibition by dilution as described above. Samples where inhibition was not resolved through dilution, are normally removed from further analysis, although this did not occur in this study.

Due to previously reported inconsistencies in how quality controls are used in the interpretation of qPCR data, generally, the most stringent quality controls were taken from literature, all of which exceeded recommendations in the minimum required information for publication of quantitative real-time PCR experiment's (MIQEs) guidelines (Bustin *et al*., 2009). A number of quality control measures were implemented to ensure consistency in the analysis of qPCR data, these are outlined in Table 2.3.

Table 2.3. Quality control measures used in the analysis of qPCR data.

Control measure	Control value	Action	Reference
Standard Curve	$R^2 > 0.99$	If standard curve values remain below control value after removing outliers, repeat qPCR run.	(Broeders *et al.*, 2014)
PCR amplification efficiency, estimated from the standard curve using - Amplification efficiency = [10(-1/slope)] - 1	Between 90% - 110%	Repeat run if efficiency is outside control values when outliers of the standard curve are removed.	(Broeders *et al.*, 2014)
Relative standard deviation	< 25%	Remove sample from analysis.	(Broeders *et al.*, 2014)
Limit of detection (LOD)	At least 2 out of 3 positive reactions, or > calculated LOD	If less than control value, report as "Below limit of detection". Do not use value in further analysis.	(Symonds *et al.*, 2016; Hughes *et al.*, 2017; Forootan *et al.*, 2017)
Limit of quantification	At least 2 out of 3 positive reactions with Cq values within ±0.5 of each other, or estimated for each assay.	Report as "below limit of quantification". Use in further analysis, but note limitations.	(Symonds *et al.*, 2016; Hughes *et al.*, 2017; Forootan *et al.*, 2017)
Test for inhibition	a) Change of Ct value < 1 with a 1:10 dilution of DNA. b) Difference of < 3 cycles between Sketa assay for blank DNA extraction and sample	a) Test with further dilution. If inhibition still a problem remove samples from further analysis. b) If difference in Ct is > 3 after further dilution, remove sample from further analysis.	(Chern *et al.*, 2011; Haugland *et al.*, 2005, 2010)
Melt curve analysis	Peak melting temperature with 1 °C of expected melting temperature.	Remove sample from analysis.	

The limit of detection (LOD) is defined as the lowest number of gene copies which can be detected with a definite probability (Forootan *et al*., 2017). No standard method for determining the LOD exists. The LOD has previously been identified by: Simply choosing a threshold cycle (Ct) value as a cut-off (Odagiri *et al*., 2015; Schriewer *et al*., 2013); the lowest gene copy number where 2 out of 3 reactions are positive (Symonds *et al*., 2016; Hughes *et al*., 2017); or simply assumed to be a single molecular target (Hassard *et al*., 2017). Forootan *et al.* (2017) suggest that LOD is assessed by taking replicates and defining a confidence level, for example a LOD at 80% confidence is the genetic target is positive in 80% of samples. While qPCR practitioners in the medical fields often chose a 95% confidence interval for their limit of detection (Forootan *et al*., 2017), genetic targets within environmental samples are often at much lower concentrations than in medical samples; this may explain why there are a number of ways used to define the LOD of an assay in environmental studies, and may justify the lower level of confidence. Here, where a single run was used (Chapter 4, Finding Host Specific Biomarkers), the LOD was defined as 2/3 positive reactions. However, estimating the LOD using 2/3 positive reactions was noted to often result in a very low LOD (< 1 gene copy per reaction) which may be difficult to differentiate from machine noise (Forootan *et al*., 2017). For the catchment study then, the probability of detection of the *RodA*, Hu100 and HF183 assays was estimated by combining the standard curves for 9 (*RodA*) or 10 (Hu100 and HF183) qPCR runs. For all assays, 2 gene copies per reaction was used as the LOD, this was deemed reasonable since this gave a probability of detection of >68% for all assays, but is more stringent than previous studies (Hassard *et al*., 2017; Symonds *et al*., 2016).

The limit of quantification (LOQ) of an assay defines the lowest gene copy number that can be determined with stated and acceptable precision and accuracy (Forootan *et al*., 2017). Similarly to the LOD, in environmental studies the LOQ is either overlooked, taken to be the sample as the LOD (Hassard *et al*., 2017), estimated from a single run (Symonds *et al*., 2016), or determined from a standard curve made from multiple runs (Forootan *et al*., 2017). For assays only using a limited number of runs (e.g., Chapter 4, Finding Host Specific *E.coli* Biomarkers), the approach taken by Symonds *et al.*, (2016), using the gene copy number having 2/3 positive reactions with quantification cycle (Cq) values within ± 0.5 was used to define the LOQ. However, this often resulted in the LOQ and LOD being the same or very similar values. To approximate the LOQ for assays used in the Seaton Sluice catchment study (Chapter 7), replicate reactions from 9 or 10 qPCR runs were aggregated to form a single standard curve with the range 2 – 2 x 10^5 gc/ reaction. Outliers were identified using Grubbs' test (Grubbs, 1969) in the outliers package version 0.14 (Komsta, 2011) and the standard curve was used to calculate the coefficient of variation (CV) for each concentration. For log-normally distributed data, Forootan *et al.*, (2017) describe the CV value as:

Equation 2.2

$$CV = \sqrt{(1+E)^{((SD(Cq))^{2} * ln(1+E)} - 1}$$

Where: E is the qPCR efficiency of all replicates plotted together and SD(Cq) is the standard deviation of replicate Cq values across all runs (Table 2.4).

While no guidelines on what CV values are appropriate (Forootan *et al*., 2017), The TATAA Biocentre (TATAA Biocentre, 2018) use a CV value of 35% for medical samples (Forootan *et al*., 2017). Here that approximately corresponds to 20 gc/sample (Table 2.4). The CV values in Table 2.4 may be higher than expected since they are an amalgamation of 9 or 10 runs completed over one year whereas studies often use a single run, some run-to-run variation is therefore expected. A LOQ of 5 gc was used for all assays, which means accepting a slightly higher CV value than the 35% recommended in medical settings (Forootan *et al*., 2017). This seems appropriate since the aim of the study was MST where low levels of pollution are expected, and the quality control measure of an RSD of <25% (Table 2.3) for sample replicates is likely to reduce variation among study samples since samples which are a high variance will be removed. Moreover, this appears to be more stringent than previous definitions of the LOD in environmental studies.

Table 2.4. Characteristics of the standard curves composed of 9 (RodA) and 10 (Hu100 and HF183) qPCR runs.

Marker	Starting quantity (gc)	Number of replicates	Proportion positive	Mean	SD	CV (%)	Amplification efficiency of standard curve	R^2 of standard curve
RodA	200000	27	1.00	17.01	0.2667	18.47	0.9871	0.9952
RodA	20000	27	1.00	20.47	0.2525	17.47	0.9871	
RodA	2000	27	1.00	23.93	0.3221	22.39	0.9871	
RodA	200	27	1.00	27.34	0.4300	30.19	0.9871	
RodA	20	27	1.00	30.67	0.4936	34.89	0.9871	
RodA	2	28	0.68	33.53	0.7332	53.72	0.9871	
Hu100	200000	30	1.00	16.40	0.3642	25.24	0.9783	0.9952
Hu100	20000	30	1.00	19.69	0.4293	29.93	0.9783	
Hu100	2000	33	1.00	23.11	0.4703	32.93	0.9783	
Hu100	200	33	1.00	26.37	0.3611	25.01	0.9783	
Hu100	20	30	1.00	29.66	0.5216	36.75	0.9783	
Hu100	2	25	0.68	32.77	0.5255	37.04	0.9783	
HF183	200000	30	1.00	12.92	0.3382	22.92	0.9521	0.9942
HF183	20000	30	0.97	16.29	0.3231	21.87	0.9521	
HF183	2000	33	0.94	19.76	0.3927	26.73	0.9521	
HF183	200	33	1.00	23.09	0.4155	28.34	0.9521	
HF183	20	33	1.00	26.65	0.5031	34.63	0.9521	
HF183	2	32	0.97	29.94	0.6576	46.20	0.9521	

2.5 Enumeration of *E.coli* and related biomarkers.

To enumerate *E.coli* from environmental samples and sewage, the *RodA* gene was used. The *RodA* gene was chosen over the *UidA* gene used by Gomi *et al.* (2014) or 23S rRNA gene (Warish *et al.*, 2015). Due to concerns over the specificity of the *UidA* gene (Sabat *et al.*, 2000) and the better reliability of *RodA* as a single copy gene (Chern *et al.*, 2011), the *RodA* gene is a better proxy for *E.coli* cell counts than the 16S or 23S rRNA genes, which typically have multiple copies.

Table 2.5.Primers and fragment details for qPCR analysis of E.coli and associated biomarkers

Assay	Primer sequences		Fragment size (bp)	Annealing temperature (ºC)	melting temperature (ºC)	Reference
RodA984 (*RodA)*	Forward	GCAAACCACCTTTGGTCG	158	60	85.0	(Chern *et al.*, 2011)
	Reverse	CTGTGGGTGTGGATTGACAT				
H8	Forward	ACAGTCAGCGAGATTCTTC	117	60	93.0	(Gomi *et al.*, 2014)
	Reverse	GAACGTCAGCACCACCAA				
H12	Forward	GTAAAAGGACTGCCGGGAAA	213	60	87.0	(Gomi *et al.*, 2014)
	Reverse	TCAGATCGTCCTTTACCAG				
H14	Forward	CAGCCTGAGCGTCTTTTAC	271	60	86.0	(Gomi *et al.*, 2014)
	Reverse	CGGTGGGAAAAGAAGTTGAA				
H24	Forward	CTGGTCTGGCTTTATAACAC	229	60	82.0	(Gomi *et al.*, 2014)
	Reverse	ATCATTTCCACTTGTCGGG				
Hu100	Forward	ACGGTTATCAGCTCACGTCG	98	60	82.0	Chapter 4
	Reverse	TCGCCCCTCGAAAAGCATTA				
Hu9	Forward	AAGCCAATGATGATGTGGGC	163	60	80.5	Chapter 4
	Reverse	TAGGCCAACTTTCTACCGCA				

Standards for the H8, H12, H14 and H24 biomarkers were kindly supplied by Gomi *et al.* (2014) and shipped from Kyoto University, using the FedEX priority service (2 days). Standards for the *RodA* gene and human markers identified in Chapter 5 were made as described above (2.4.3 Quantitative PCR (qPCR)). Primer sequences and primer concentrations of commonly used targets are given in Table 2.5.

2.5.1 Enumeration of HF183

The most commonly used human marker for MST targets the HF183 16S rRNA gene cluster within *Bacterodes spp.* (Bernhard & Field, 2000; Green *et al*., 2014; Harwood *et al*., 2014). While a range of primer sets have been tested, the HF183/BacR287 primer set (Table 2.6) has been shown to be superior to the HF183/BFDrev primer set, with no spurious nonspecific amplification, greater precision and a lower limit of detection (Green *et al*., 2014; Haugland *et al.*, 2010). Stock, linearized standards were kindly received from Bunce *et al.,* (*In prep*) and stored at -20°C.

Table 2.6. Primers used for the detection of the HF183 marker gene

Primer	Sequence	Genbank Accession	Melting Temperature (°C)	Reference
HF183	ATCATGAGTTCACATGTCCG	AB242142 (179 to 346 bases)	78.5	(Green *et al*., 2014)
BacR287	CTTCCTCTCAGAACCCCTATCC			

2.6 DNA Sequencing and data analysis

2.6.1 Whole genome sequencing and analysis of E.coli isolates

To prevent the sequencing of identical genotypes of *E.coli*, repetitive element PCR was performed, using the BOX-A1R primers, on selected isolates to aid selection for sequencing. Repetitive element PCR targets highly conserved, repetitive elements that occur at different intervals within individual bacterial genomes. The BOX repetitive elements are comprised of three subunit sequences, boxA, boxB and boxC, located in intergenic sequences (Koeuth *et al.*, 1995). Here, the 59 base-pair boxA sequence was targeted using a single primer; PCR of this repetitive element produces fragments of different lengths, allowing individual bacteria to be distinguished to sub-species level. BOX-PCR was performed as previously described (Mohapatra *et al.*, 2007), using the Fast Start, High Fidelity PCR kit (Sigma-Aldrich, UK). The gel was run at 50 V for 10 minutes and then at 80 V (4 V cm^{-1}) at 4 °C until the marker reached the end of the gel (c.

4.5 h). Gels were visualized using a Dual-Intensity Transilluminator (UVP, USA). Analysis of the BOX-PCR images was conducted using Bionumerics Version 4 (ApplidMaths). The Pearson's product-moment coefficient was used to create a similarity matrix of *E.coli* isolates for individual hosts (Mohapatra & Mazumder, 2008).

DNA was extracted (2.3.1 *E.coli* isolates suspended in lysogenic broth) and the quality and quantity of DNA checked using a Nanodrop 1000 and qubit (3.5 DNA quality control) as described above. Library preparation and sequencing were performed using a TruSeq DNA HT library preparation kit (Illumina, US), on an Illumina MiSeq using a 250 bp paired-end metric, respectively, according to the manufacturer's instructions at the Earlham Institute, UK. The genomic data was quality trimmed with Trimmomatic (Bolger *et al.*, 2014) using a 4 bp sliding window with an average quality score of 20. Reads less than 150 bp in length were removed from further analysis. Read quality was assessed using FastQC (Andrews, 2010). Unpaired reads were concatenated into a single fasta file and both paired and unpaired reads were used to assemble contigs using SPAdes (Bankevich *et al*., 2012), all commands are given in Appendix A.1. *De novo* assembly was chosen over mapping to a reference to ensure all of the accessory genome was captured. The quality of each assembly was evaluated using QUAST (Gurevich *et al*., 2013).

2.6.2 16S rRNA amplicon sequencing by Ion Torrent

To prepare DNA for 16S rRNA gene sequencing, extracted DNA, stored at -80 ^{o}C, was defrosted on-ice. A PCR amplification targeting the V4 and V5 regions was performed using the defrosted DNA, as described above (4.1 Polymerase Chain Reaction (PCR)) with the following modifications. One of fifty unique, Golay error-correcting barcoded primers, incorporated onto fusion primers containing the forward primer (Table 2.6) and Ion Torrent sequencing adapters were placed in each PCR reaction, as recommended by the manufacturer (Life Technologies, UK). A reverse primer (Table 2.7) was added to each reaction, and PCR amplification was performed with an annealing temperature of 56 ^{o}C.

Table 2.7. Primers used in the preparation of Ion Torrent 16S rRNA gene libraries for Ion Torrent sequencing

Primer	Direction	Sequence	Reference

V4 (515f)	Forward	GTGNCAGCMGCCGCGGTAA	(Quince *et al.*, 2011)
V5 (926r)	Reverse	CTTCCTCTCAGAACCCCTATCC	(Quince *et al.*, 2011)

PCR products were visualized through gel electrophoresis on a 1.5% agarose gel stained with Nancy-520 DNA stain (SigmaAldrich, Bumbleby, Ukraine) and visualized using a Dual-Intensity Transilluminator (UVP, USA). PCR fragments were cleaned to remove unwanted products such as primer-dimer, where primer molecules hybridise to each other instead of their DNA target, using Ampure XP beads (Beckman Coulter, Brea USA), according to the manufacturers instruction. The cleaned PCR products were then quantified using the high sensitivity DNA kit for Qubit (Life Technologies, Carlsbad USA) as described above (3.5 DNA quality control). An equimolar pool with a concentration of 100 pM was created from all samples. The equimolar pool was enriched using emulsion PCR with an OneTouch V2 machine with 400 base-pair chemistry (Life Technologies, Carlsbad USA), according to the manufacturer's instructions. The enriched library was sequenced on an Ion Torrent Personal Genome Machine (Life Technologies, Carlsbad USA), according to the manufacturer's instructions.

2.6.3 Analysis of data from Ion Torrent sequencing

The Quantitative Insights Into Microbial Ecology (QIIME) package v1.9.1 (Caporaso *et al.*, 2010) was used to analyse all Ion Torrent data. The raw data file (BAM format) was converted to fastq format, required by QIIME, using the Torrent Suite software v4.4.2. A generalised QIIME script is given in Appendix A.2. Briefly, sequences were simultaneously demultiplexed and quality filtered to remove sequences less than 100 bp in length or with a mean quality score below 20. OTU were picked using an open reference method. OTUs are clustered at the 97% similarity level using the UCLUST package (Edgar, 2010) against the SILVA 138 database (Pruesse *et al.*, 2007); sequences that are not aligned to any in the SILVA database are clustered *de novo*. Representative sequences, defined as the first sequence in an OTU cluster, were aligned using PyNAST (Caporaso *et al.*, 2010) and filtered to remove chimeric sequences using the ChimeraSlayer package (Haas *et al.*, 2011). A phylogenetic tree was constructed from the remaining phylogenetic sequences using the FastTree package (Price *et al.*, 2010).

It should be noted that open reference OTU picking methods, while still used widely, are becoming less common due to their propensity to overinflate the number of OTUs (Callahan *et al*., 2016). These methods are being replaced by methods such as Divisive Amplicon Denoising Algorithm (DADA) 2 (Callahan *et al*., 2016) and Deblur (Amir *et al*., 2017), which attempt to correct for errors incurred during PCR and sequencing and allow for a higher resolution analysis of sequences (See below). However, I used these methods since they are standard methods in 16S rRNA gene sequencing analysis and previous work in MST has used these methods. In addition, the DADA2 and Deblur packages were developed to deal with Illumina data and are often unsuccessful with other data types, such as Ion Torrent.

2.6.4 16S rRNA amplicon sequencing by Illumina

During the project I decided to move from sequencing using the Ion Torrent PGM sequencing platform to the Illumina Miseq platform. This decision was made because:

- The environmental microbiology field has largely moved to Illumina platforms.
- The Illumina platforms are less prone to error and can deal better with homopolymers (Shokralla *et al*., 2014).
- The larger data producing capacity of the Illumina Miseq machine allowed 150 samples to be processed per run, achieving a similar sequencing depth as using 50 samples per run on the Ion Torrent PGM. This resulted in a lower cost per sample.

The quality and quantity of DNA was determined as described above, and Illumina sequencing was undertaken using the NU-OMICS sequencing service (NU-OMICS, 2018). Illumina sequencing was performed on an Illumina Miseq, using the 515F (GTGCCAGCMGCCGCGGTAA) and 806R (GGACTACHVGGGTWTCTAAT) primers (Caporaso *et al*. 2011) targeting the V4 region of the 16S rRNA gene, with the V2 500 chemistry.

2.6.5 Analysis of data from Illumina sequencing

Illumina data was received as separate fastq files for the forward and reverse reads which were imported into QIIME2 (Caporaso *et al*. 2010; Caporaso, 2018) and demultiplexed. The quality of the imported reads was assessed using the summarize feature of QIIME2 and the Qiime2View software (Caporaso, 2018). The demultiplexed reads were then error-corrected and filtered to remove chimeric sequences, sequences shorter than 100 bp, and phix reads, used as an internal control, using the DADA2 pipeline (Callahan *et al*., 2016). The DADA2 pipeline was chosen since it has been shown to be more accurate and less prone to including spurious sequences than other common methods such as UCLUST (Edgar, 2010) used in QIIME1, and average linkage used in Mothur (Schloss *et al*., 2009). The DADA is reference and cluster free and is therefore capable of classifying sequences to a higher resolution than the typical classification of an OTU of 97% similarity. Instead of the OTU table produced by QIIME1 (Caporaso *et al*., 2010), an amplicon sequence variants (ASV) table is produced. A multiple sequence alignment was then conducted using the mafft (Katoh *et al*., 2002) program to remove highly variable base positions from the ASVs before construction of a phylogenetic tree using the FastTree package (Price *et al.*, 2010).

2.6.6 Analysis with SourceTracker

The SourceTracker program (Knights *et al*., 2011) was used to estimate the contribution of potential faecal sources to sinks. OTU tables or ASVs, were converted from BIOM format to tab-separated files and mapping (QIIME1) or metadata (QIIME2) files were prepared for SourceTracker by adding “Env” and “sourcesink” columns to describe the sample environment and whether it is a source or a sink, respectively. SourceTracker was run using default settings as recommended by Henry *et al.* (2016). Henry *et al.* (2016) also recommend that, for microbial source tracking, SourceTracker is run five times and sources with a relative standard deviation >100% adjusted to zero contribution.

2.7 General statistical analysis

2.7.1 Summary statistics

All statistical analyses were performed in R (R Core Team, 2017) using the Rstudio graphical user interface (R Studio Team, 2015). All figures were made using ggplot2 (Wickham, 2016), supported by the gridExtra package, version 2.3 (Auguie, 2017).

To allow comparison of bacteria and gene copy concentrations between sample sites the geometric mean and geometric standard deviation were calculated using equations 2 and 3, respectively. The geometric mean was chosen as it accounts for the potentially large variation and zero skew in bacteria data better than the arithmetic mean.

Equation 2.3

$$Geometric\ mean = e^{\frac{\Sigma log(x)}{n}}$$

Equation 2.4

$$Geometric\ standard\ deviation = e^{\frac{1}{n}\Sigma\ (logx - log(geometric\ mean))^2}$$

Linear regression was performed using the lm function in R (R Core Team, 2017). The measure of influence of data points on linear regression models was assessed using Cook’s distance and DFFITS analyses conducted using the “olsrr” package, version 0.5.1 (Hebbali, 2018). To determine if the slope of a linear regression model was significantly different to 1 an analysis of variance (ANOVA) test was used.

The normality of count data was evaluated statistically using the Shapiro-Wilk normality test and visually through histograms, quartile-quartile (Q-Q) and residual plots. The homoscedasticity of the data was evaluated using the Bartlett and the Fligner-Killeen tests.

While the vast majority of studies use a log-transformation to normalise bacterial and gene-copy count data, log-transformations have been shown to perform poorly, except where mean counts are large and dispersion is small (O’Hara and Kotze, 2010). To normalise data, a box-cox transformation was conducted using the MASS package (Venables and Ripley, 2002) to evaluate lambda values between -6 and 6 at 0.1 intervals and values were transformed using Equation 2.5.

Equation 2.5

$$Transformed\ x = x^{\frac{\lambda - 1}{\lambda}}$$

To compare mean bacterial concentrations at each site, analysis of variance (ANOVA) was conducted using the ANOVA function in base R (R Core Team, 2017).

2.7.2 Sensitivity and specificity of faecal markers

The sensitivity of biomarkers, synonymous with the true positive rate, of *E.coli* biomarkers was by:

Equation 2.6

$$Sensitivity = 100 \times \frac{Number\ of\ true\ poisitives}{Total\ number\ of\ samples}$$

The specificity of biomarkers, or one minus the rate of false negatives, was calculated using Equation 2.7.

Equation 2.7

$$Specificity = 100 \times \left(1 - \frac{Number\ of\ false\ negatives}{Total\ number\ of\ samples}\right)$$

Chapter 3 Human-associated *E.coli* genetic markers and community analysis to identify pollution from decentralized systems

3.1 Introduction

Faecal pollution contributes greatly to reduced water quality through nutrient loading and can present significant public health risks through pathogen loading. In the UK, only 36% of rivers were classified as good or excellent water quality in 2015 (Priestley, 2015) according to the Water Framework Directive (2000/60/EC, European Commission, 2000). The economic burden of surface water pollution in England and Wales is estimated to be £1.3 billion per annum, largely to the issues of identifying and mitigating diffuse pollution sources (National Audit Office, 2010). Methods to establish the occurrence, location and sources of pollution are therefore invaluable in making informed investment decisions to ensure efficient improvements to water quality and quantity.

The microbiological quality of water is monitored and regulated using the faecal indicator organisms (FIOs) *E.coli* and enterococci. While FIO presence indicates recent faecal contamination their ubiquity in the intestinal tract of most warm-blooded mammals means current, culture-based enumeration methods fail to identify or differentiate the source(s) of FIO (Reischer *et al*., 2008). Modelling approaches are promising in estimating agricultural contributions to FIOs (Whitehead *et al*., 2016; Dymond *et al*., 2016), although, contributions from misconnections, leaking sewers and poorly positioning of malfunctioning septic systems can be difficult or impossible to model and delineate from agricultural sources.

Microbial source tracking (MST) describes techniques which attempt to identify and apportion sources of pollution. Whilst a plethora of MST approaches exist (Scott *et al*., 2002; Harwood *et al*., 2014), identifying relationships between MST results and regulatory FIOs is difficult (Marti *et al*., 2013; Ridley, *et al*., 2014) due, in part, to the differing behaviour of distinct bacteria in environmental waters. The suitability of *E.coli* and enterococci as indicator organisms is well debated (Wade *et al*., 2003b; Marion *et al*., 2010; Lamparelli *et al*., 2015). Nevertheless, their current regulatory role increases the desire to link MST conclusions to FIO concentrations. Recently, four human-associated *E.coli* genes (H8, H12, H14, H24) have shown promise as library-independent biomarkers in Japan (Gomi *et al*., 2014) and Australia (Warish *et al*., 2015) to link human pollution to *E.coli* concentrations. Lower sensitivities and specificities, however, were noted in Australia (Warish *et al*., 2015) compared to Japan, highlighting the necessity to assess marker performance prior to use in new locations using likely sources of pollution (Stoeckel & Harwood, 2007). The applicability of these markers outside of the Indo-Australian region has never been evaluated previously.

Decentralized wastewater treatment systems, which are common in rural catchments throughout the UK, present particular difficulties for library-independent MST approaches. The problem that MST approaches are not 100% sensitive to human faeces is often not important since sewage often contains faeces from a large population. However, in decentralized systems, this will not always be the case. The use of biomarkers where small decentralized treatment is prevalent may result in false negative results. The increasing speed and accuracy of high throughput sequencing (HTS), coupled with reducing costs which have historically outpaced Moore's Law (Muers, 2011), and the ability to characterize bacterial communities from environmental samples using the 16S rRNA gene, may lend this technology to water quality monitoring (Vierheilig *et al*., 2015; Schang *et al*., 2016) and MST (Unno *et al*., 2018). The SourceTracker software (Knights *et al*., 2011) widens the applications of HTS to MST, allowing the contribution of microbial communities from possible sources to environmental samples to be estimated. A number of MST investigations (Newton *et al.*, 2013; Neave *et al.*, 2014; Ahmed, *et al.*, 2015) have used both marker and community-based approaches, generally finding a consensus between conclusions drawn from host-associated markers and SourceTracker outputs, albeit with poor correlation between the two (Ahmed, *et al.*, 2015). No studies, however, have investigated relatively small catchments consisting of decentralized treatment systems nor compared community analysis and human-associated *E. coli* marker approaches.

The aims of this study were two-fold. Firstly, to evaluate the performance of the human-derived *E.coli* markers, H8, H12, H14 and H14, discovered in Japan, for their ability to detect human-associated *E.coli* in the UK. Secondly, to identify and assess human pollution in a Northern England catchment potentially impacted by agriculture and small decentralized wastewater treatment systems using the human-associated marker assay in conjunction with community analysis.

3.2 Methods

We first assessed marker sensitivity and specificity using a multiplex-PCR assay to identify markers in target (sewage) and non-target (non-human) hosts (chicken, cow, horse, dog, pig and sheep) collected from Northern England. This ensured the suitability of markers for use in the catchment study. Following communications with the Environment Agency and Northumbrian Water Ltd., we attempted to streamline the detection of markers from regulatory plate-counts, since picking individual *E.coli* from non-regulatory plates to test marker presence/absence would limit their regulatory and industrial application. A qPCR method was therefore tested and used to estimate the proportion of colonies on a plate which contained a human-associated biomarker.

3.2.1 Assessing marker performance

3.2.1.1 Sample collection

As recommended by the U.S. Environmental Protection Agency (USEPA, 2005), ten faecal samples from each non-human source: chicken, cow, horse, dog, pig and sheep were collected as previously described (2.1.2 Faecal samples). Nine sewage samples were collected from five sewage treatment works (< 2,000 PE) and directly from a septic tank (Appendix B.1) and transported to the lab as previously described (2.1.3 Raw sewage)

Some of the sources used to assess biomarker performance in this study were located outside of the catchment, for three reasons: access to septic tanks within the catchment was limited; samples from a single household may not be representative of all septic tanks in the area, and to assess biomarker distribution across the North of England.

3.2.1.2 Faecal DNA extraction

DNA was extracted (2.3.3 Extraction of DNA from faecal samples) and quality checked as previously described (2.3.5 DNA quality control).

3.2.1.3 E.coli culturing, isolation and DNA extraction

Thirty *E.coli* were cultured from sewage samples and twenty from each non-human source (Appendix B.1), using different faeces from those above (3.2.1.2 Faecal DNA extraction) as previously described (2.3.1 *E.coli* isolates suspended in lysogenic broth). To reduce the likelihood of re-sampling identical *E.coli*, only 1-3 colonies were selected from each initial plate and samples were collected from five different small (< 2,000 PE), decentralized WWTPs (Appendix B.1).

3.2.1.4 Multiplex-PCR

A multiplex PCR method (2.4.2 Identification of *E.coli* biomarkers using multiplex PCR) was developed and used for the detection of H8, H12, H14 and H24 (Gomi *et al.*, 2014).

3.2.1.5 Data analysis

The sensitivity and specificity of each human marker was evaluated and reported as the percentage of *E.coli* from sewage possessing a marker gene (Warish *et al.*, 2015; Gomi *et al.*, 2014), as described above (2.7.2 Sensitivity and specificity of faecal markers). The specificity was determined using *E.coli* isolates (n = 120) (Gomi *et al.*, 2014) and faecal sources (n = 60) (Warish *et al.*, 2015) from non-human sources using Equation 2.7.

3.2.2 Catchment study

Catchment area and sampling strategy

A full description of the catchment can be found in Appendix B.2. Briefly, the Newby catchment is a sub-catchment of a larger catchment management program (Eden Demonstration Test Catchment). The Newby catchment is comprised of agricultural land interspersed with small settlements, typically farms and villages which are served either individually or communally by septic tanks.

River water samples (n=36) were collected on 6 days over two months (May to July) targeting different flow conditions in the catchment. Sampling locations were targeted above and below farm settlements in two sub-catchments and at the catchment outlet (Figure 3.1).

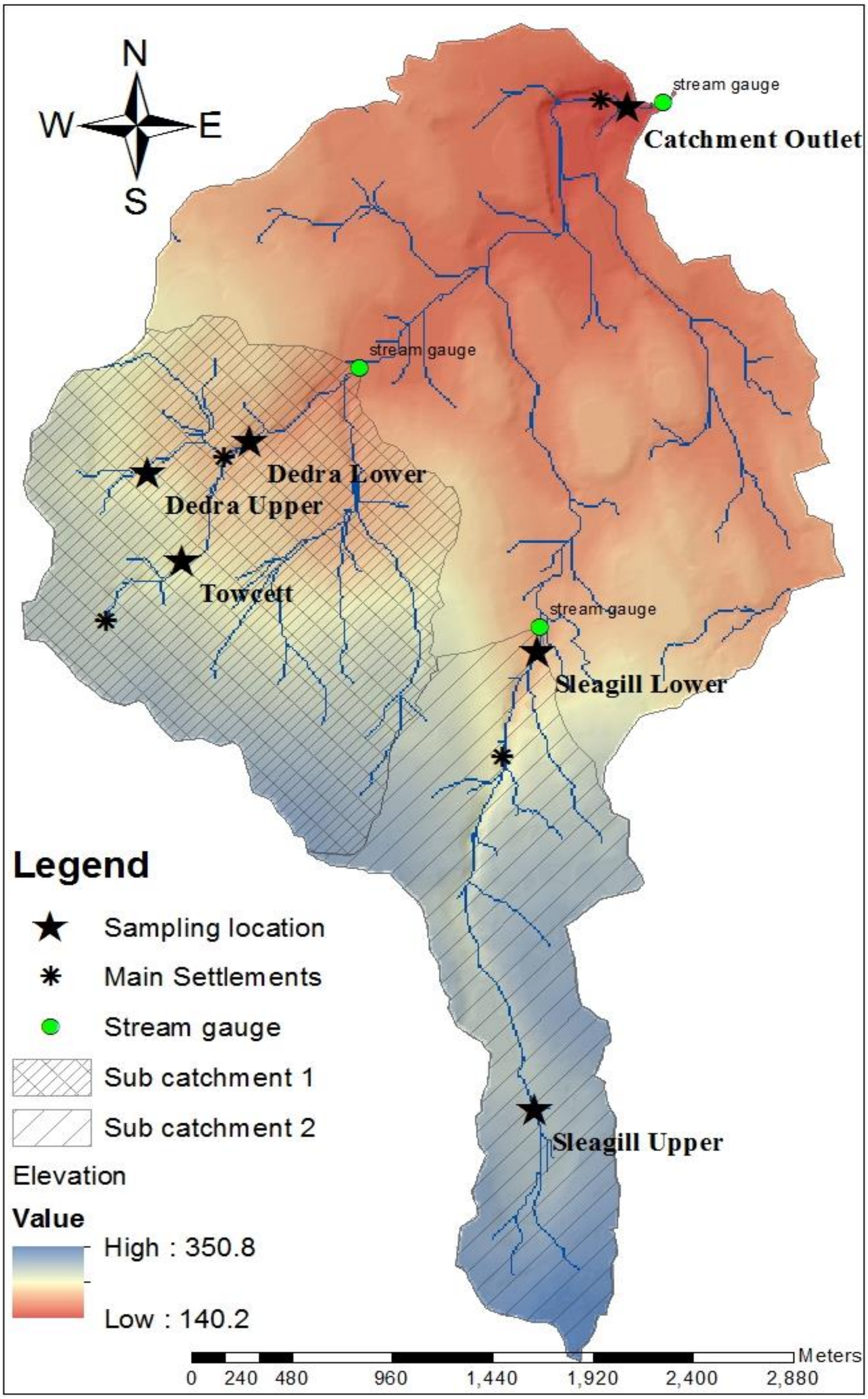


Figure 3.1. Map of the Newby catchment in the Eden Valley, Northern England. Created using open source data (Ordinance Survey, 2017).

Environmental sample processing

River water samples were collected and transported as described above (2.1.4 Environmental water samples). *E.coli* were enumerated (2.2 Enumeration of faecal indicator bacteria) and the plates were frozen at -20 ^{o}C until further use (c. eight weeks).

All colonies were picked from a single plate and combined for DNA extraction (2.3.2 Direct extraction from *E.coli* isolated from plate counts) and identification of marker genes (Multiplex-PCR and inhibition control) was performed on the DNA. In addition, colonies from eleven sets of duplicate plates were picked individually and processed in the same way. These eleven plates were chosen as they gave the largest range of proportions where both duplicate plates had > 5 colonies on each plate. QPCR was carried out on samples that were PCR-positive for human marker genes.

QPCR

QPCR was carried out in triplicate for each human marker and the *RodA* gene as described above (2.4.3 Quantitative PCR (qPCR)). The *RodA* gene was chosen over the *UidA* gene used by (Gomi *et al*., 2014) or 23S rRNA gene (Warish *et al*., 2015) due to its reliability as a single-copy gene (Chern *et al*., 2011), providing a closer relationship with cell counts compared with multi-copy genes. While no inhibition was noted a 1:10 dilution was used to keep sample values within the range of the standard curve (10^2 – 10^8 gene copies).

Data analysis and reporting

The proportion of *E.coli* containing marker genes was estimated using the ratio of all marker gene copies to *RodA* gene copies. Whilst it is possible that a single isolate may contain more than one marker gene, it was assumed that this overestimate would impact results less than the variability in sensitivities of individual markers between different septic systems. The absolute abundance of *E.coli* containing a marker was calculated by multiplying the mean *E.coli* concentration with the proportion of *E.coli* containing a marker gene at each site. All statistical analysis was conducted in R (R Core Team, 2017), DFFITS analyses were performed with the 'olsrr' package version 0.5.2 (Hebbali, 2018). ANOVA and determination of Pearson's correlation coefficient was performed using the

‘Hmisc’ package version 4.1-1 (Harrell & Dupont, 2018) to assess the correlation between the PCR and qPCR assays. For the catchment study, the geometric mean was used to summarize the data since the data set was relatively small (n=36), ranged across 5 orders of magnitude and was skewed towards 0. To determine if there was a significant difference between human *E.coli* concentrations at each site, a Box-Cox transformation (Box & Cox, 1964) was performed (2.7.1 Summary statistics).

16S rRNA gene sequencing and SourceTracker analysis

16S rRNA gene sequencing was performed (2.6.2 16S rRNA amplicon sequencing by Ion Torrent) and analysed as described above (Analysis of data from Ion Torrent sequencing).

SourceTracker (Knights *et al*. 2011) was run (2.6.6 Statistical analysis) using a faecal taxon library (FTL) consisting of three human (one septic tank and two raw influent), one ovine, two bovine, two equine and two avian (chicken) samples. Human samples were split into septic tank and sewage samples to run the SourceTracker analysis, and the contribution of each source, estimated by SourceTracker, was added together to give a human contribution. All commands and parameters are given in Appendix A.1, SourceTracker outputs are given in Appendix B.3.

3.3 Results

3.3.1 Human marker performance in rural catchments

All markers had a 100% sensitivity to sewage (n =9). The H24 marker had the highest sensitivity when using isolates from sewage (50%, Figure 3.2), slightly higher than in Japan (37%, Figure 3.2; Gomi *et al*., 2014), although all other markers had lower sensitivities than in Japan, namely 17%, 10% and 3% for H14, H8 and H12, respectively. The sensitivities for H14 and H24 were not tested in Australia due to their poor specificities (Warish *et al*., 2015). The aggregated sensitivity of all the markers was 69%, which is marginally higher than that in Japan (67%).

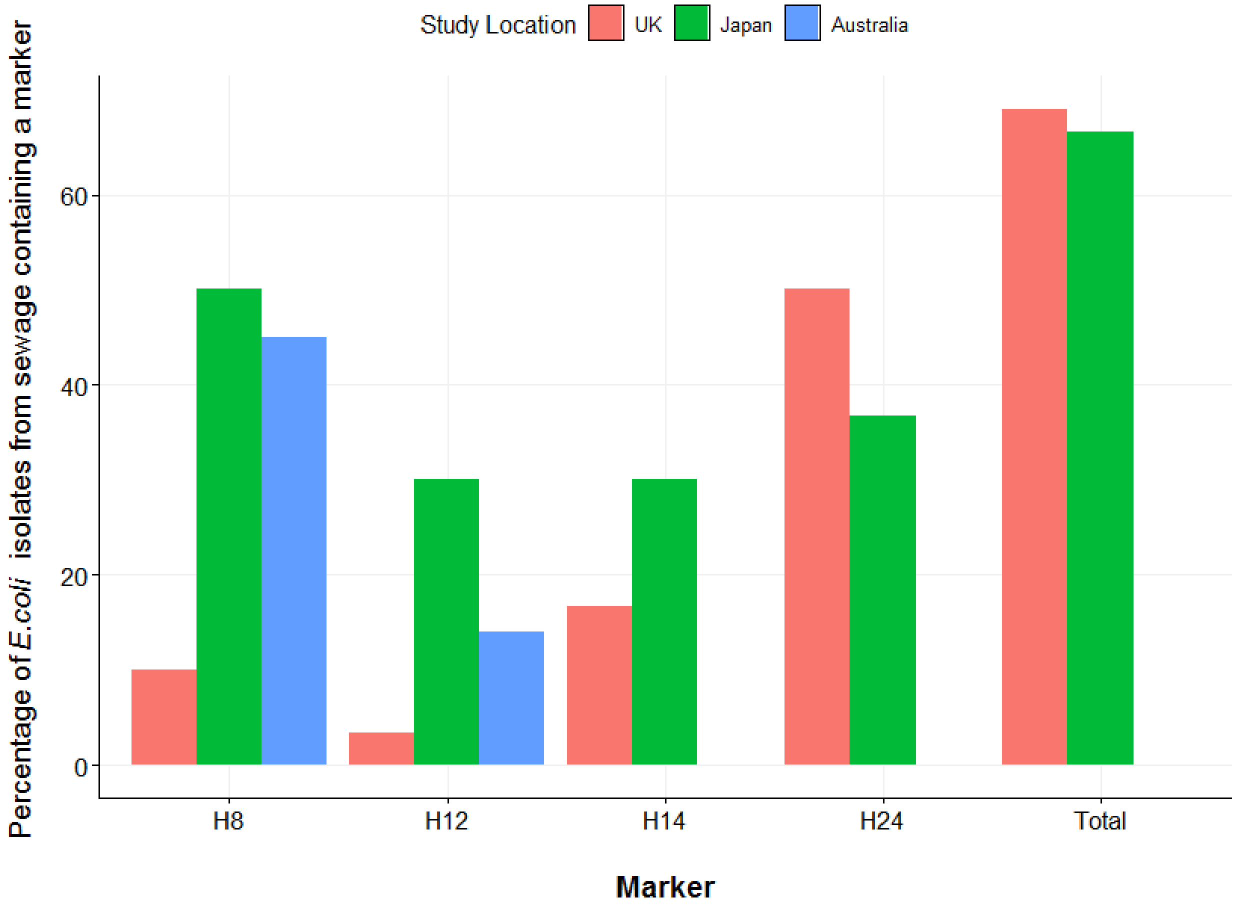


Figure 3.2. The sensitivity of the H8, H12, H14 and H24 markers determined by laboratory tests in the UK (this study), Japan (Gomi et al., 2014) and Australia (Warish et al., 2015). It should be noted that the sensitivity of H14 and H24 was not tested in Australia due to the low specificity of these markers there (Warish et al., 2015)

The specificity, determined by testing faecal samples, was 100%, 100%, 93% and 93% for H8, H12, H14 and H24, respectively (Table 3.1). Interestingly, while the specificities when testing *E. coli* colonies isolated from faeces largely agree, 99%, 100%, 93% and 96%, respectively, some cross-reactivity with Sheep (H8) horse (H14), pig (H14 and H24) and dog (H24 and H14) was noted that was not identified through faecal sampling alone. The high specificities of markers noted via both methods justified their use in the field trial.

Table 3.1. The number of each faecal source positive for each marker detected through PCR in studies using these markers so far. Only host sources tested in this study are shown, Warish et al. (2015) also tested faeces from emu, deer, cat, kangaroo and possum

	Number of samples/isolates PCR positive for each marker															
Study Location	**UK (This Study)**				**UK (This Study)**				**Japan** (Gomi *et al*., 2014)				**Australia** (Warish *et al*., 2015)			
Marker	H8	H12	H14	H24	H8	H12	H14	H24	H8	H12	H14	H24	H8	H12	H14	H24
Cow	0	0	0	1	0	0	1	1	2	0	2	0	5	5	7	8
Chicken	0	0	2	3	0	0	0	1	0	0	0	0	ND	ND	ND	ND
Pig	0	0	0	0	0	1	4	1	1	0	0	1	ND	ND	ND	ND
Sheep	0	0	2	0	1	0	1	2	ND	ND	ND	ND	ND	ND	ND	ND
Dog	0	0	0	0	0	1	1	0	ND	ND	ND	ND	0	0	4	2
Horse	0	0	0	0	0	0	1	0	ND	ND	ND	ND	0	0	4	0
Sample size	60				120				90				90			
Specificity (%)	100	100	93.3	93.3	99	98.3	93.3	95.8	97	100	97.8	98.9	94	94.4	83.3	88.9
Sample type	Faecal				Isolates				Isolates				Faecal			

**The total number of samples and total specificity observed in the study are shown here, even though for other studies only the sources which were the same as those tested in this study are shown above.*

3.3.2 Estimating the proportion of human E.coli using qPCR

The proportion of *E.coli* on a given plate containing human marker genes was estimated from the ratio of gene copy numbers of markers to *RodA* gene copies determined by qPCR. To attempt to verify this method, 270 individual colonies were picked from duplicate plates of eleven samples covering a range of predicted percentage contributions; human markers were identified in individual colonies using end-point PCR. The percentage of *E.coli* containing human markers as estimated by qPCR was compared to that identified by end-point PCR. One set of duplicate plates was removed from further analysis since one plate contained 26% of *E.coli* with a human marker and the duplicate plate contained none. This could be due to the duplicate plate only containing 13 isolates. Three plates which had < 20 isolates were unavoidably used in the comparison since plates obtained from lower or greater dilutions had too few (<10) or too many (>100) isolates.

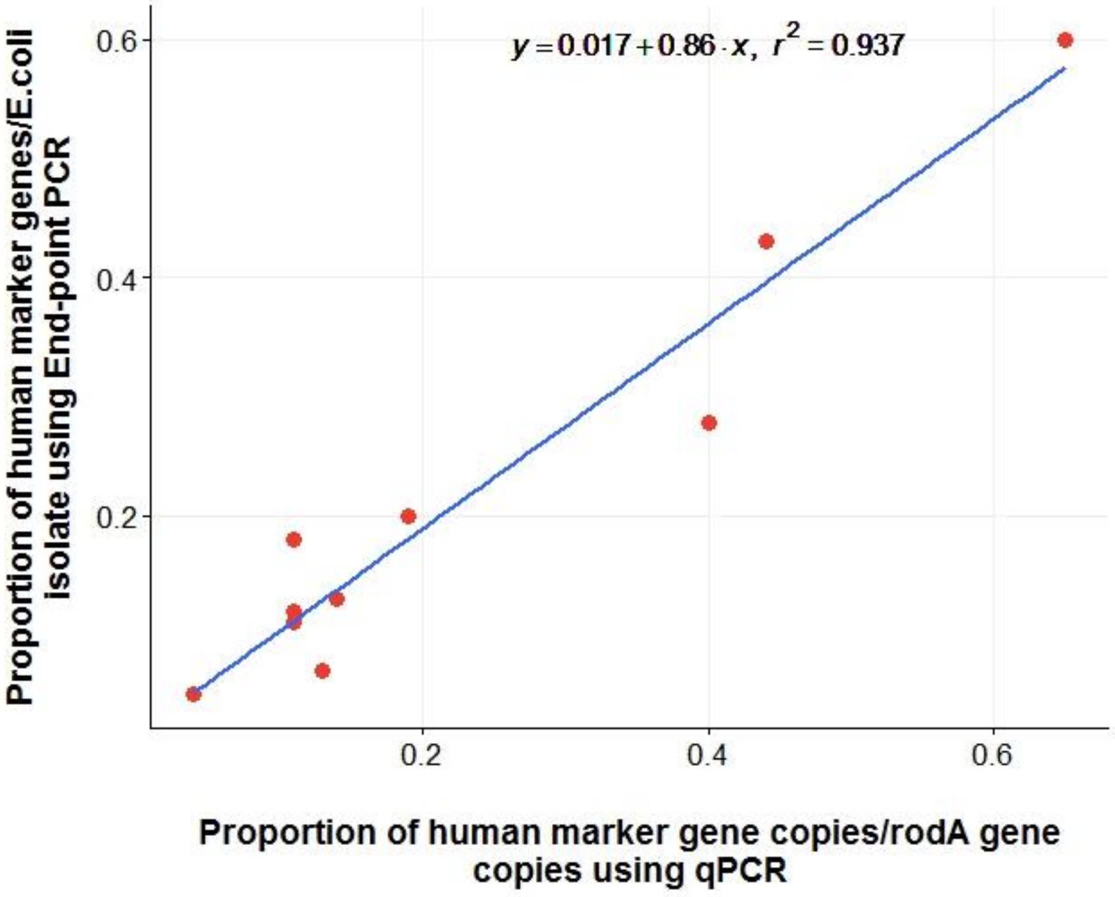


Figure 3.3. Comparison of qPCR and multiplex end-point PCR methods to estimate proportion of human-derived E.coli in environmental samples

The qPCR assay gave a good agreement of the proportion of marker genes among cultured *E.coli* identified through end-point PCR (r^2 = 0.937, p = 4.434e-06, Figure 3.3). Two points were noted as highly influential to the fit and intercept of the graph (0.64, 0.60 and 0.4, 0.29, Figure 3.3) using Cooks distance and DFFITS analysis. When these points were removed, the fit of the regression line changed little (r^2 = 0.914) and the intercept and slope were not significantly different to 1 and 0, respectively (y = 0.011 + 0.94x).

Although the between-plate variability was expected to be high, when a low number of isolates possess markers, the relationship observed in figure 3.3 was strong. For example here, a difference of a single marker between duplicate plates with 20 colonies would account for a 5% difference between plates. The maximum difference between the two assays was 8% when comparing the PCR and qPCR assays for the two most abundant markers (Appendix B.4). The low error and strong relationship (Figure 3.3 and Appendix B.4) justified the use of the qPCR data in the field trial to estimate the proportion of *E.coli* containing a human-associated marker, hereon in termed the proportion of human *E. coli*, with which it is synonymous.

3.3.3 Catchment study

3.3.3.1Human pollution

The H24 and H14 were most commonly detected, in 28 and 14 of the 36 samples, respectively, while H8 and H12 were only present in 3 samples (Table 3.2).

Table 3.2. Number of samples at each sampling location PCR positive for each E.coli marker in the Morland catchment

Sample Location	H8	H12	H14	H24	At least one marker
Outlet	0	1	3	4	5
Dedra lower	1	0	5	6	6
Dedra upper	0	0	2	3	4
Towcett	0	0	0	5	5
Sleagill lower	1	1	2	6	6
Sleagill upper	1	1	2	4	5
Total	3	3	14	28	31 (/36)
Percentage	8.3	8.3	38.9	77.8	86.1

E.coli concentrations in the catchment ranged between 20 and 8300 CFU/100 mL, with the exception of a single sample with 21,000 CFU/100 mL. Up to 65% of the *E.coli*, although typically between 10 and 20%, contained a human marker. Human *E.coli* marker data are summarized in figure 3.4 and broken down by sample day and location (Appendix B.2).

The abundance of *E.coli* containing human markers increased following each of the farm settlements at Dedra ($p = 0.0183$, Figure 3.4 (Top)) and Sleagill ($p = 0.0565$, Figure 3.4 (Top)). The increase in abundance of *E.coli* containing human markers at Sleagill was not statistically significant (marginally), which was attributable to the large ranges in *E.coli* concentrations. However, the human-associated *E.coli* concentration increased in all except a single sample (Sleagill upper – Sleagill lower Day 4, Appendix B.2).

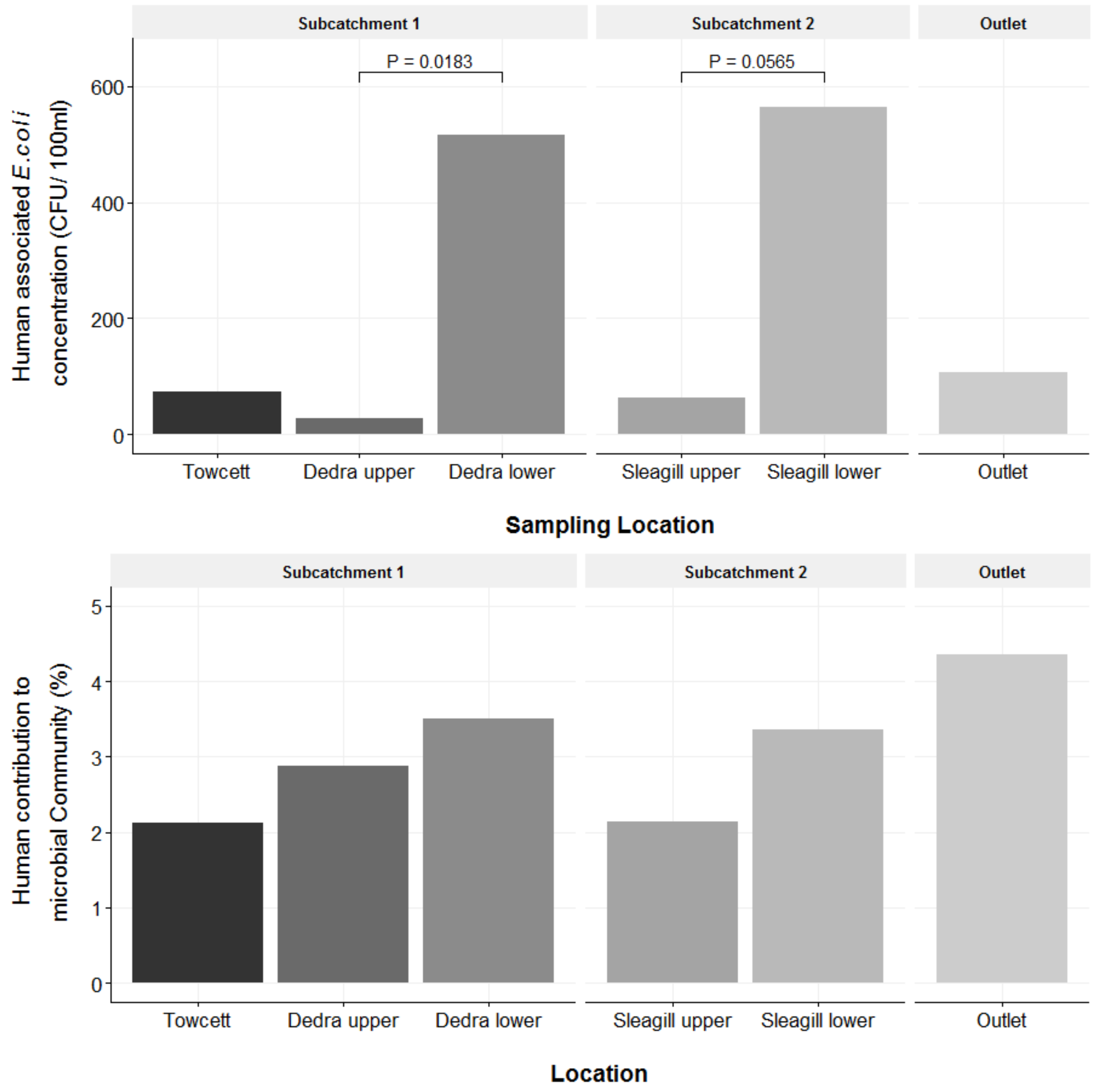


Figure 3.4. Top – Back-transformed means of Box-Cox transformed data (lambda value of 0.2) of human-associated E.coli. Significance values determined using one-way ANOVA. Bottom – Geometric mean values of the human proportion of the microbial community at each location predicted by SourceTracker

The ubiquity of human pollution was evident from both MST methods. Human sources were identified in all samples by community analysis and 86% (31/36) of these samples also contained human *E.coli* markers (Table 3.2). Community analysis also showed an increase in the human proportion of pollution at Dedra and Sleagill (Figure 3.4). While this increased the confidence in conclusions drawn from the human *E.coli* marker studies, there was a weak, non-significant (marginally) correlation between the human *E.coli* contribution and the reported contribution of the sewage microbial community (ρ = 0.32, p = 0.0577, Appendix B.5).

At the catchment outlet, community analysis showed an increasing human contribution, while the human proportion of *E.coli* appeared to decrease (Figure 3.4). The human-associated *E.coli* loading rate was calculated using flow rates from within the catchment. The *E.coli* loading rates from each of the two streams (Figure 3.5) suggest that the Sleagill sub-catchment contributes a greater human-associated *E.coli* load to the catchment outlet.

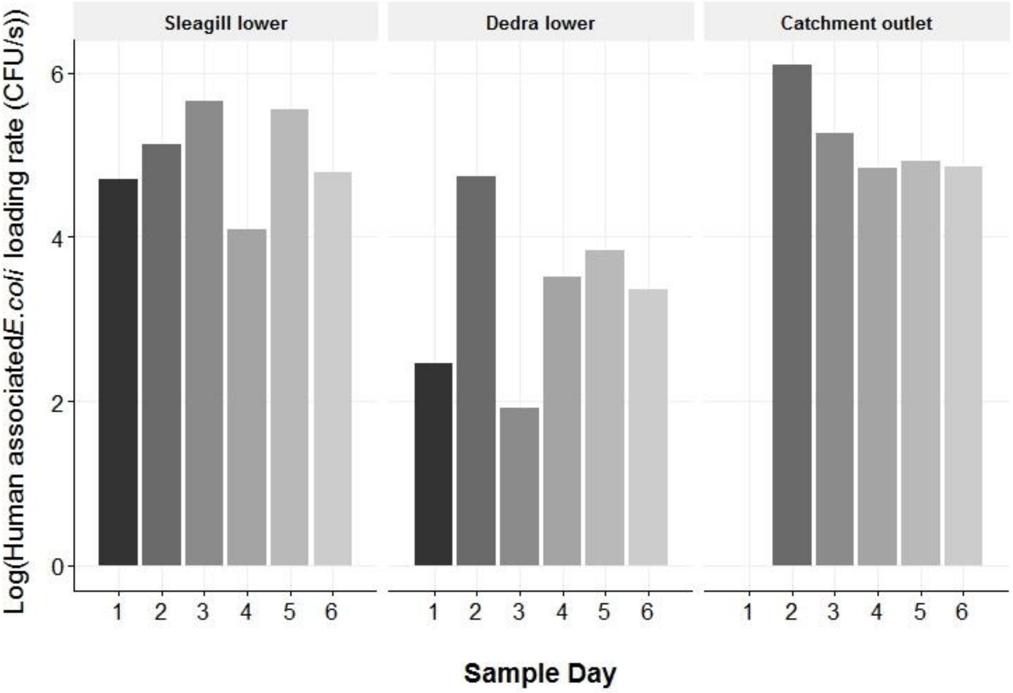


Figure 3.5. Loading rates (Log-transformed) of human-associated E.coli on each sample day in the two streams and at the catchment outlet of the Morland catchment

3.4 Discussion

3.4.1 E.coli biomarkers

An ideal marker for MST should be present only in the target host, abundant in the target host and present in every individual of the target host species. Aligning the results of microbial source tracking and regulatory microbial monitoring techniques is also desirable since MST investigations are often undertaken as a result of high counts of faecal indicator organisms and typically form the basis for catchment management and investment decisions. Unfortunately, no marker that is entirely specific to a host has been previously identified (McLellan *et al.*, 2013). Moreover, relating MST markers to *E.coli* and enterococci concentrations remains challenging. Recently identified host-specific *E.coli* markers (Gomi *et al*., 2014) may, therefore, provide an invaluable link between MST approaches and regulatory parameters. While these markers have been tested in Japan and Australia, little was known about their suitability as MST markers in the UK. It would appear that the human-associated *E. coli* markers investigated here are not ideal sole candidates when used individually to apportion faecal contamination to humans in the UK. While they showed high specificity (>93%), their sensitivities varied between 3% and 50% in *E.coli* isolated from sewage samples. This was similar to sensitivities of 14 to 50% reported elsewhere (Gomi *et al*., 2014; Warish *et al*., 2015) and when aggregated, the sensitivity of all markers increased to 69%, very similar to the 67% previously reported (Gomi *et al*., 2014).

The specificity of *E.coli* biomarkers varies between studies. Here, the specificities of H8, H12, H14 and H24 (99%, 98%, 93% and 96%, respectively) were more similar to those observed in Japan (97%, 100 %, 98%, and 99%, respectively) based on isolates, rather than those based on faecal samples (100%, 100%, 93%, and 93%, respectively in this study) compared to those in Australia (94%, 85%, 57%, and 72%, respectively). This could be due to the range of potential pollution sources tested, since some pollution sources important to Australia (Warish *et al*., 2015), such as emu faeces, are not a concern in the UK. It may also be due to the manner of sample collection. Warish *et al.*, (2015) observed a relatively high level of cross-reactivity with cow faeces when using composite faecal samples, compared to those obtained from individual faecal samples used in this study (Table 3.1). Here, we evaluated specificity through two methods, either using 10 faecal samples or 20 *E.coli* isolates from 6 non-human faecal sources, which are likely to be prevalent in the UK.

Interestingly, a greater number of non-human targets were positive for the four human markers when using *E.coli* isolated from faeces compared with DNA directly extracted from faeces. This may be due to the fact that different samples, from different locations, were used for DNA extraction and *E.coli* culturing; an uneven distribution of marker genes throughout non-human populations would lead to animals in some areas possessing a greater proportion of marker genes. The higher number of non-human hosts positive for a marker gene when using isolates may also be due to human-associated isolates being present in <10% of non-human hosts. For example, only a single horse faeces, from the twenty tested, was positive for the H14 marker. The difference observed between these methods may also be due to inhibition in the PCR reaction, or the low abundance of markers in non-human faeces. However, all faeces were PCR positive for the *RodA* gene, a gene highly associated with *E.coli spp.* (Chern *et al.*, 2011), suggesting that no significant inhibition took place. *E.coli* markers may be below the limit of detection for faeces, and only present when identified through the culture-based approach. It seems more likely that the disagreement between the two methods was due to the low prevalence of these markers across non-human populations, and the general agreement of high specificities for all markers by both methods suggest they are suitable markers for MST in the UK.

The proportions of *E.coli* containing the H8 and H12 markers (10% and 3%) were markedly different to those observed in Japan (50% and 30%, Gomi *et al.*, (2014)) and Australia (45% and 15%, Warish *et al.*, (2015)), but were similar to the 16.3% of *E.coli* isolates containing H8 reported in Bangladesh (Harada *et al.*, 2018). Variation in sensitivity between studies could be due to different sampling methodologies or geographical distributions of markers between different communities. Such variation could reduce the efficacy of these markers in decision making, particularly if this variation occurs on a localized scale. It is convenient to report the proportion of *E.coli* containing a marker for catchment studies (Gomi *et al.*, 2014; Warish *et al.*, 2015; Harada *et al.*, 2018), however, if the proportion of *E.coli* containing markers differs widely between local sources of human pollution, comparing the amount of human pollution between sampling points becomes difficult using this metric. A number of examples of where this variation may confound MST results are available. (Warish *et al.*, 2015) observed a much greater proportion of *E.coli* containing the H12 marker (>20%) compared to the H8 marker (<~8%) in four out of six sites, even though the proportion of *E.coli* containing the H8 and H12 marker in sewage was 45% and 14%, respectively. Here, on one occasion, H24 accounted for ~60% of *E.coli* (estimated with both qPCR and end-point PCR methods, Dedra Lower Sample Day 1, Appendix B.2 Figure 2), greater than the expected sensitivity (50%), but on other occasions, the value was much lower. In both examples we cannot discount the possibility that these markers had non-human origins, however, the between-study variation in sensitivity warrants further investigation into the variation in sensitivity on a local scale, before these markers can be recommended for quantitative use in decision making.

The identification of *E.coli* biomarkers through culturing, picking individual cultures and end-point PCR is laborious. While enumeration directly from water samples with qPCR would be more rapid and less labour intensive, it does not currently fit in with EU regulations and previous difficulties have been reported (Gomi *et al*., 2014). As a compromise, the proportion of *E.coli* containing biomarkers was estimated using qPCR on all isolates simultaneously removed from a plate. This qPCR “compromise” gave a good estimate of the ratio of *E.coli* containing human-associated markers, although there are limitations. The precision of this method was limited by the number of isolates of *E.coli* on a plate - a plate with only 20 isolates yields a precision of only ±5%. Additionally, an error of up to 8% was noted and attributed to between-plate variability, which questions whether a single plate is representative of the environmental sample, especially where a low percentage of isolates are human-associated. Sampling additional plates may be useful, although this becomes increasingly laborious and, therefore, expensive. Future studies may be better served using qPCR to directly apportion faecal biomarkers without a culture step and accept discrepancies between regulatory culture-based and MST nucleic-acid based methods; although Gomi *et al.* (2014) noted difficulties in direct enumeration using qPCR due to the low abundance of markers. New molecular technologies such as digital PCR (dPCR) (Cao *et al.*, 2015) with a greater tolerance to inhibition and improving sensitivities are valuable avenues of research which may offer a solution.

3.4.2 Comparison of community analysis and E.coli biomarkers.

Community analysis complemented *E.coli*-based MST in the catchment study. The use of multiple markers and community analysis was vital to avoid false negative results. Whilst H24 was the most commonly detected marker, using H24 alone would have resulted in at least one false negative result. The conclusions drawn in the catchment study were based primarily on the H14 and H24 markers, those with the lowest specificities (~93%). However, it should be noted that without these markers, using only H8 and H12, the rate of false negative results would render *E.coli* based MST unreliable in this catchment. Community analysis identified human sources in all samples, whereas, *E.coli* biomarkers were only identified in 81% of samples. The greater sensitivity of community analysis compared to *E.coli* biomarkers is likely to be due to the limitation of the culture-based method compared to community analysis. The culture-based method is limited to the detection and analysis of *E.coli* containing human sources that can selectively grow on plates, whereas community analysis can detect a much greater proportion of sewage-related taxa in an environmental sample.

The use of both marker and community analysis based approaches may improve confidence in MST results and relevance to regulatory monitoring techniques. Community analysis is limited by the proportional nature of the data, and the inability to relate this to quantitative, regulatory measures. While marker-based approaches appeared to be necessary to draw imprecise quantitative conclusions, community analysis provided an important role in improving confidence in the conclusions drawn from *E.coli* markers, where the cross-reactivity of markers has the potential to confound MST conclusions. However, direct comparison of marker and community analysis data remains difficult. The poor correlation between human biomarkers and community analysis (Appendix B.5), reflects previous findings (Ahmed, *et al*., 2015) and highlights the difficulty in interpreting community analysis results for governance of FIO. The poor correlation between the estimated human contributions from each method is likely to be due to the differential die-off, transportation and sedimentation rates (Walters *et al.*, 2009) between culturable organisms and DNA since the human contribution to the microbial community consistently increases, in contrast to the fluctuating concentrations of *E.coli* biomarkers down the catchment (Figure 3.4, bottom). The poor correlation between the estimated human contributions could also be due to the variation in the contributions from other sources to the microbial community that can alter the predicted contribution from human sources by community analysis. For example, any increase in the proportion of the human bacterial community after a farm may be suppressed due to the overall increase in faecal sources which increases the density of microbes downstream of a farm (Appendix B.2, Figure 3.4).

The composition of the faecal library input into SourceTracker remains the subject of some debate. A number of studies successfully used single samples (Henry *et al*., 2016; Sun *et al*., 2017) and while one study suggested that larger sample sizes are required to avoid false negatives, Staley *et al.*, (2018) suggest that less than 10 individuals may suffice if geographically-associated sources are used. Here, three human samples were used, and no false negative values were noted. For the water industry or catchment managers, using a large number of samples for all potential hosts in each monitored catchment is likely to be unfeasible with high associated costs. Further work to understand the required library size, and how geographically representative a faecal library is, would be valuable to inform further MST studies.

3.5 Conclusions

Human *E.coli* markers (H8, H12, H14 and H24) were tested in the North of England. Markers with the highest sensitivities, and which were most useful in a catchment study, were H24 and H14, although, these also had the lowest specificities. Marker sensitivities differed to those reported in Japan (Gomi *et al*., 2014) and Australia (Warish *et al*., 2015). This variation in marker sensitivity may limit the quantitative application of *E.coli* biomarkers, for example in regulatory monitoring, or comparing two sampling sites, particularly if this variation occurs on a local scale.

A qPCR based assay to estimate the percentage of *E.coli* isolates containing markers from a cultured plate, fitting into regulatory testing of *E.coli*, was used successfully to reduce labour and its associated costs, although, enumeration of gene copies directly from environmental samples would reduce labour further.

A combination of community analysis and multiple human-associated *E.coli* biomarkers improved confidence in MST conclusions and may make community analysis and an MST tool more relevant to decision makers. The conclusions drawn from each method agreed, although no direct correlation was found between the percentages of human contribution predicted by each assay. These differences are most likely due to the disparate persistence of culturable *E.coli* and DNA in a river environment, although, these differences provide additional information which would likely be missed without this "toolbox" approach.

The field trial highlighted the importance of MST in rural catchments, where human impacts are often overlooked, to disentangle human and agricultural inputs for management decisions or improving catchment models.

3.6 Acknowledgements

The demonstration test catchments belong to a DEFRA funded project and data was kindly received from the Eden test catchment. DNA for the Human-associated *E.coli* biomarkers was kindly received from Ryota Gomi at the Department of Environmental Engineering, Kyoto University.

Chapter 4 Identifying human-Specific *E.coli* biomarkers using a database approach for tracking sewage pollution in the UK.

4.1 Introduction

There is an urgent need to improve the microbiological quality of water on national and international scales. The magnitude of the global challenges sanitation engineers face is exemplified by global statistics: there are 700,000 annual deaths attributable to diarrheal disease (Prüss-Ustün *et al*., 2014), 1.8 billion people access drinking water contaminated with faeces, and 80% of wastewater is discharged to the environment without treatment (UNDP, 2018). While many OECD countries have high wastewater treatment rates (>99.9%), only 25% of surface waters in England (Salvidge, 2016) are on track to achieve the government's aim to improve 60% of surface waters to their natural state by 2021 (Priestley, 2015).

Methods to support investment and management decisions are critical to ensure cost-effective improvements in water-quality, particularly where urban diffuse pollution is a contributing factor to poor water quality. Leaking sewers, faulty combined sewer overflows (CSOs) and misconnections are difficult to identify and their contribution to pollution often remains unknown or unaccounted for. Misconnections in an estimated 1.25 million UK properties cost the water industry around £235 million/year (Royal Haskoning, 2007). Methods to determine when and where investment in infrastructure is necessary, and indeed cost-beneficial, to achieve desired water quality improvements are becoming increasingly important in the UK and across the world.

Microbial source tracking (MST) methods, which attempt to identify and often apportion sources of faecal pollution, could inform investment decisions. Currently, the most popular MST methods are library independent (Harwood *et al*., 2014), where genetic markers, previously validated as highly associated with a given host, are used to identify contamination in environmental waters. The ideal MST marker is often described as being highly host-associated, abundant in all members of the target host, similar to FIO and/or pathogens in terms of their inactivation rates, and exhibit geographic and temporal stability in their sensitivity, specificity and abundance (Stoeckel and Harwood 2007). For industrial applications of MST it is also important that assays are cost-effective and relate to current regulatory methods.

Currently, MST is rarely used in the UK water industry, possibly due to the difficulty in relating the current suite of library independent MST markers to regulatory faecal indicator organisms (FIO's) (Reischer *et al*., 2008; Wang *et al*., 2013; Mattioli *et al*., 2016). *Escherichia coli* are used across Europe, and much of the world, as an FIO to regulate and monitor the quality of recreational water, both fresh (US and Europe) and coastal (Europe), and in drinking water according to the Bathing Water (2006/7/EC) and Drinking Water (98/83/EC) Directives, respectively.

Apportionment of *E.coli* by source is increasingly of interest to direct catchment investment decisions, inform epidemiology studies (Fewtrell and Kay, 2015) and apportion viable antimicrobial resistant *E.coli* (Leonard *et al*., 2018). Relating non-*E.coli* MST marker concentrations to FIO concentrations is complicated since genetic markers and FIO differ in several aspects, including: their initial faecal concentrations; transport and attenuation mechanisms (Johnston *et al*., 2010); their environmental decay rates both within faeces (Oladeinde *et al*., 2014) and in environmental waters (Brown and Boehm 2015; Wanjugi *et al*. 2016; Korajkic *et al*. 2014); the assays used for their detection (Ahmed *et al.*, 2015); and the ubiquity and possible environmental persistence of FIO as opposed to the host-specificity of genetic markers. The use of *E.coli* as an MST marker is, therefore, attractive for MST studies.

A number of studies have attempted to apportion *E.coli* to different source. Early studies used a variety of DNA fingerprinting methods (Parveen *et al*., 2001; Versalovic *et al*., 1994; Dombek *et al*., 2000; Araújo *et al*., 2014). Unfortunately, the poor reproducibility and variability that increased with library size, made these methods generally unreliable (Meays *et al*., 2004; Dombek *et al*., 2000). The rise in popularity, and reduction in costs, of sequencing has allowed the exploration and comparison of a large number of *E.coli* genomes. Gomi *et al.*, (2014) used whole genome sequencing of 22 *E.coli* genomes from chickens, cows, humans, and pigs. A comparison of genomes led to the development of a number of markers highly-associated with human (H8, H12, H14 and H24), cow, pig and chicken faeces. Other methods to detect host-specific markers have also successfully identified host specific *E.coli* associations. Deng *et al.* (2015) analysed polymorphisms in the *ycjM* gene finding a human-associated genotype. While high sensitivities (34%-86%

of *E.coli* from sewage contained the human-associated ycjM genotype) and specificities were reported locally (99%), Kataržytė *et al.*, (2018) found H8 concentrations were higher than *ycjM* gene concentrations in environmental samples. A similar approach, using logistical regression to identify host-associated patterns of polymorphisms across multiple intergenic regions, revealed a number of host-specific patterns (Zhi *et al*., 2015). This approach, however, requires the sequencing of multiple genomic regions to identify these patterns in individually isolated cultured environmental isolates, which could be extremely expensive for water quality monitoring where analysis of up to 100 isolates per plate may be required (2006/7/EC, CEU, 2006).

Across all markers, H8 has been favoured due to its apparent high specificity to sewage (Gomi *et al.,* 2014; Warish *et al.*, 2015). However, some variations in the sensitivity and specificity of markers has been noted which, may limit the usefulness of these markers particularly for decision-making. Recently, variation in the sensitivity of *E.coli* markers has been noted, while H8 was highly prevalent in Japan (Gomi *et al*., 2014) and Australia (Warish *et al*., 2015), with sensitivities of 50% and 45%, respectively, it was less prevalent in the UK (10%) and Bangladesh (16.3%, Harada *et al.*, 2018). Whilst this could be due to different sampling methodologies (single septic tank; single wastewater treatment plant (WWTP) and multiple small WWTPs), it may also be due to differences in the geographic distribution of these markers in the environment. There is also some evidence that this geographical variation may occur on a local scale. For example, Warish *et al*., (2015) found H12 to occur more often than H8 in environmental isolates, despite the predicted sensitivity of H8 being more than three times that of H12 in DNA extracted from sewage samples (45% compared to 14%). Such differences could be attributed to; the presence of these markers in non-human sources, differential die-off rates of *E.coli* containing these different markers, and variation in the sensitivities of these markers between geographically different host (human) communities.

Low, and variations in, the sensitivity of human-associated markers, reduces the efficacy of markers in identifying sources with low levels of pollution. For example, in a recent catchment study (Chapter 3), the detection of H8 from cultured *E. coli* isolates alone would have resulted in a large number of false negative results. Moreover, variation in marker sensitivity between geographically distinct host (human) populations may make it difficult to make confident investment and management decisions. It is therefore vital to understand whether variation in marker sensitivity occurs at a local level and whether there are better markers for use in some areas. Nevertheless, apportioning *E.coli* by source is desirable for MST since it would allow decision makers to easily determine where financial investment will have the largest impact. Using nucleic acid detection methods, as opposed to culture-based methods, may overcome the limitation of low concentration of *E.coli* biomarkers. While the proportion of *E.coli* containing H8 may be low, the concentration of this marker in sewage has been found to be similar to HF183 (Hughes *et al.*, 2017), the most commonly used human associated marker (Harwood *et al.*, 2014). However, the direct detection of H8 from environmental DNA has some limitations. H8 is not specific to *E.coli*, highly similar sequences occur in *Yersinia* and *Klebsiella* sp. which could confound MST results, particularly if their behaviour in the environment is different to that of *E.coli* (Gomi *et al.*, 2014).

The aims of this study were two-fold: i) to determine if human biomarkers in *E .coli,* other than those previously published (Gomi *et al.*, 2014), exist and ii) to assess variation in the sensitivity of markers across the North East of England. This was done by identifying likely genetic markers using a database approach, and assessing their abundance and sensitivity in small, decentralised treatment plants, which represent small communities where urban diffuse pollution is likely to occur.

4.2 Methods

In this study we assumed that gene recombination is a causative process of host specificity in *E.coli*, although other processes, such as local sequence changes and DNA rearrangement all likely influence host-specificity (Arber, 2000). Coding sequences in the accessory genome of 221 publicly available and 23 locally assembled *E.coli* genomes were compared using the Large Scale Blast Score Ratio (LS-BSR) software (Sahl *et al.*, 2014), and ranked according to their specificity and sensitivity to human hosts. The sensitivity and specificity of coding sequences were then determined (2.7.2 Sensitivity and specificity of faecal markers) using 12 screened sewage samples from wastewater treatment plants (WWTPs) and 60 non-human faecal samples. Raw sewage samples were collected during the winter from 11 small (< 2,000 population equivalent (PE)), decentralised wastewater treatment plants and one medium-sized wastewater treatment plant (~30,000 pe). WWTPs were between 2 and 100 miles of where the local human derived *E.coli* were isolated.

4.2.1 Database development

Since biomarker performance has been shown to vary with geographic location, a library of genomes was constructed from those available on the National Centre for Biotechnology Information (NCBI) database to give a broad indication of their likely global performance. An existing database of publically available *E.coli* genomes (Zhi, Li, *et al*., 2016) was adapted; genomes not definitively of faecal origin, e.g. an isolate found on chicken breast meat, were removed and recently available *E.coli* genomes were added along with 23 locally sourced and sequenced *E.coli* genomes. Details of the 263 genomes in the final database are available in Appendix C.1. Table 4.1 shows the number of *E.coli* from each source in the database.

Table 4.1. The number E.coli genomes used to construct the local and global databases.

	Human	**Bovine**	**Ovine**	**Porcine**	**Canine**	**Equine**	**Avian (Laridae)**	**Avian (Gallus)**	**Environmental**	**Other**	**Total (non-human)**	**Total**
Global	99	33	9	27	8	13	4	29	6	11	141	239
Local	4	3	3	3	3	3	2	3	0	0	20	24
Total	103	36	12	30	11	16	6	32	6	11	159	263

The database consisted of *E.coli* isolated from the faeces of likely non-human sources of pollution in the UK. Eight environmental strains, were also included in the database since these are emerging as a potential complication of environmental monitoring using *E.coli* (Luo *et al*., 2011); these environmental strains are considered non-faecal, naturalised strains of *E.coli* and although they are phenotypically and taxonomically indistinguishable from faecal strains, they were isolated from an environment with no apparent faecal contamination and have been assigned to one of five cryptic *Escherichia* lineages (Walk *et al*., 2009). Where raw reads were available, these were assembled as described below.

4.2.2 Isolation, sequencing assembly of local genomes

Local *E.coli* were selected from a previously constructed library (Chapter 1). Table 4.2 shows the sampling areas and location. All non-human *E.coli* were isolated from individual faeces.

Table 4.2. Dates and areas samples were collected from which E.coli isolated and processed for whole genome sequencing.

Host	Sample type	Area	Date of collection	Number of isolates
Human	WWTP 1	County Durham	08/2015	1
	WWTP 2	County Durham	08/2015	2
Chicken	Free range, individual faeces	County Durham	04/2015	1
	Free range individual faeces	N. Northumberland	08/2015	2
Cow	Beef cow Individual faeces	Newcastle	05/2016	2
	Beef cow Individual faeces	County Durham	12/2016	1
Horse	Individual faeces	N. Northumberland	04/2015	1
	Individual faeces	County Durham	05/2015	2
Pig	Individual faeces	N. Northumberland	04/2015	2
	Individual faeces	S. Northumberland	08/2015	1
Sheep	Individual faeces	S. Northumberland	04/2015	3
Dog	Individual faeces	S. Northumberland	04/2015	3
Gull	Individual faeces	N. Northumberland	04/2015	2

BOX-PCR (2.6.1 Whole genome sequencing and analysis of *E.coli* isolates) was used to ensure duplicate strains were not sequenced. Genomic DNA was purified from isolates as previously described (2.3.2 Direct extraction from *E.coli* isolated from plate counts). The quantity and quality of *E.coli* gDNA was assessed (2.3.5 DNA quality control) prior to sequencing (2.6.1 Whole genome sequencing and analysis of *E.coli* isolates). Sequencing data was processed as previously described (2.6.1 Whole genome sequencing and analysis of *E.coli* isolates).

4.2.3 Biomarker identification

The LS-BSR software (Sahl *et al*., 2014) was used to interrogate the *E.coli* database, comparing the presence of coding sequences (CDSs) from individual genomes (Query sequences) in all other genomes. A number of options exist for the use of LS -BSR, here we used Prodigal (Hyatt *et al*., 2010) to predict CDSs from all genomes and USEARCH (Edgar, 2010) at a pairwise identify of 0.9 (Sahl *et al*., 2014) to identify unique CDSs. Query CDSs were then aligned against each genome in the database using BLASTN to get a query score. The BLAST score ratio (BSR) is calculated by first, generating a query score by conducting a BLAST search of the query sequence with itself and then dividing this by the reference score, generated by conducting a BLAST search of the query sequence with all other sequences in the database. This results in a BSR value between 0.0 and 1.0 (Sahl *et al*., 2014), with 1.0 indicating an exact match. The BSR attempts to reduce both the bias introduced by short sections of highly similar sequences, which artificially deflate E-values, and the variation in the raw BLAST score with length, which limits its applicability for comparative analytics (Rasko *et al.,* 2005).

Two values of BSR score were used to identify the presence of CDSs in query genomes. CDSs with a BSR value >0.4 were assumed to be present in non-target organisms, for the purpose of specificity; CDSs with a BSR value >0.8 were assumed to be present for the purpose of sensitivity (Sahl *et al*., 2014) i.e., we tried to underestimate CDS presence in target hosts and overestimate in non-target hosts to reduce the likelihood of chance matches in the BLAST searches affecting the markers selected for lab trials. The output spreadsheet from LS-BSR was uploaded into R (R Core Team, 2017) and the BSR values change to 1 or 0 to reflect presence or absence of a CDS in each genome, respectively.

The sensitivity$_{\text{isolates}}$, the proportion of *E.coli* from humans containing a marker, and the specificity$_{\text{isolates}}$, one minus the proportion of *E.coli* from non-human hosts containing a marker, of each CDS were then calculated separately for the the locally sourced *E.coli* isolates only (local database) and then the entire (global) library.

4.2.4 Biomarker selection

Biomarkers were selected using the process outlined in Figure 4.1. A total of 7930 and 152 CDSs had a sensitivity$_{\text{isolates}}$ > 1% and >10%, respectively, and a specificity$_{\text{isolates}}$ > 95%. Sequences with either a global sensitivity >10% or a local sensitivity$_{\text{isolates}}$ > 25%

were subject to BLAST searches of the NCBI database. Where sensitivity$_{\text{isolates}}$ is defined as the proportion of *E.coli* from the target host containing a marker and specificity$_{\text{isolates}}$ is one minus the proportion of *E.coli* form a non-target host containing a marker. CDSs which were highly similar to sequences found in more than one organism other than *E.coli* were discarded. Sequences highly similar (97% similarity) to those in *E.coli* from non-target hosts were sense checked, the CDSs were retained if they fulfilled the following criteria: i) the number of *E.coli* from non-target hosts seemed reasonable, given the suggested specificity$_{\text{isolates}}$, i.e., less than one non-target host in 20 sequences for a 95% predicted specificity, and ii) they matched one of the seven non-target hosts suggested by the database, all others were discarded. In addition, sequences with no matches to the NCBI database were also discarded. From the remaining 81 sequences, seven CDS (Table 4.2) that were not tested in a previous study (Gomi *et al*., 2014) were selected for in vitro testing. The selection of the seven CDS was inevitably slightly subjective, based on the reasoning in Table 4.3; briefly, the seven sequences were selected due to their high sensitivity$_{\text{isolates}}$ in either the global (Hu100) or local (Hu9) database. Other markers (Hu113, Hu117, Hu112, and Hu42) were selected from a range of sensitivities to test whether the concept of a global database was effective at identifying useful markers for source tracking, i.e. to test whether the sensitivity$_{\text{isolates}}$ and specificities$_{\text{isolates}}$ in the database were reflected in *in vitro* sensitivities$_{\text{isolates}}$ and specificities$_{\text{faecal}}$.

Table 4.3. A summary of the reasons for selecting markers identified computationally, for in vitro testing.

CDS	Predicted sensitivity* in global database	Predicted sensitivity* in local database	Predicted specificity*	Reasoning
Hu100	27.9%	0%	97.6%	Had the highest sensitivity* in the global database.
Hu9	9.6%	75%	94.1%	Had the highest sensitivity* and a high specificity in the local database.
Hu112	9.9%	0%	99.3%	Selected as a marker with a global sensitivity* between those of Hu100/Hu117 (~27%), and Hu42/Hu56 (~6%) to give a range of sensitivities to test the database approach.
Hu113	15.8%	0%	97.5%	Selected as a marker with a global sensitivity* between those of Hu100/Hu117 (~27%), and Hu42/Hu56 (~6%) to give a range of sensitivities* to test database approach.
Hu117	26.7%	50%	94.6%	This had the highest sensitivity* in the global database of markers which were also present in the local database.
Hu56	5.9%	0%	100.0%	Had 100% specificity* to humans
Hu42	5.8%	75%	95.8%	Had the highest sensitivity and a high specificity* in the local database.

*Sensitivity and specificity refer to the isolate sensitivities and specificities.

4.2.5 Biomarker validation

For biomarkers selected through the process outlined in Figure 4.1, PCR Primers were designed (Table 4.4) and checked for specificity using primer BLAST (Ye *et al.*, 2012). Primers were optimised by varying the annealing temperature between 48 and 65 ^{o}C using DNA extracted, as previously described (2.3.3 Extraction of DNA from faecal samples), from post-screen raw sewage from a medium-sized (~30,000 pe) WWTP (Tudhoe Mill, UK) and choosing the annealing temperature resulting in the brightest band, with the expected fragment length and no unwanted amplification. PCR was carried out on a further 30 isolates from the same WWTP to determine the sensitivity and test the selected annealing temperature. A library of 12 sewage and 60 faecal samples from 6 non-target organisms, sheep, cow, pig, dog, horse and chicken were interrogated for biomarker presence through endpoint-PCR as previously described (2.4.1 Polymerase Chain Reaction (PCR)), using the optimised annealing temperature, with a primer concentration of 0.4 µM.

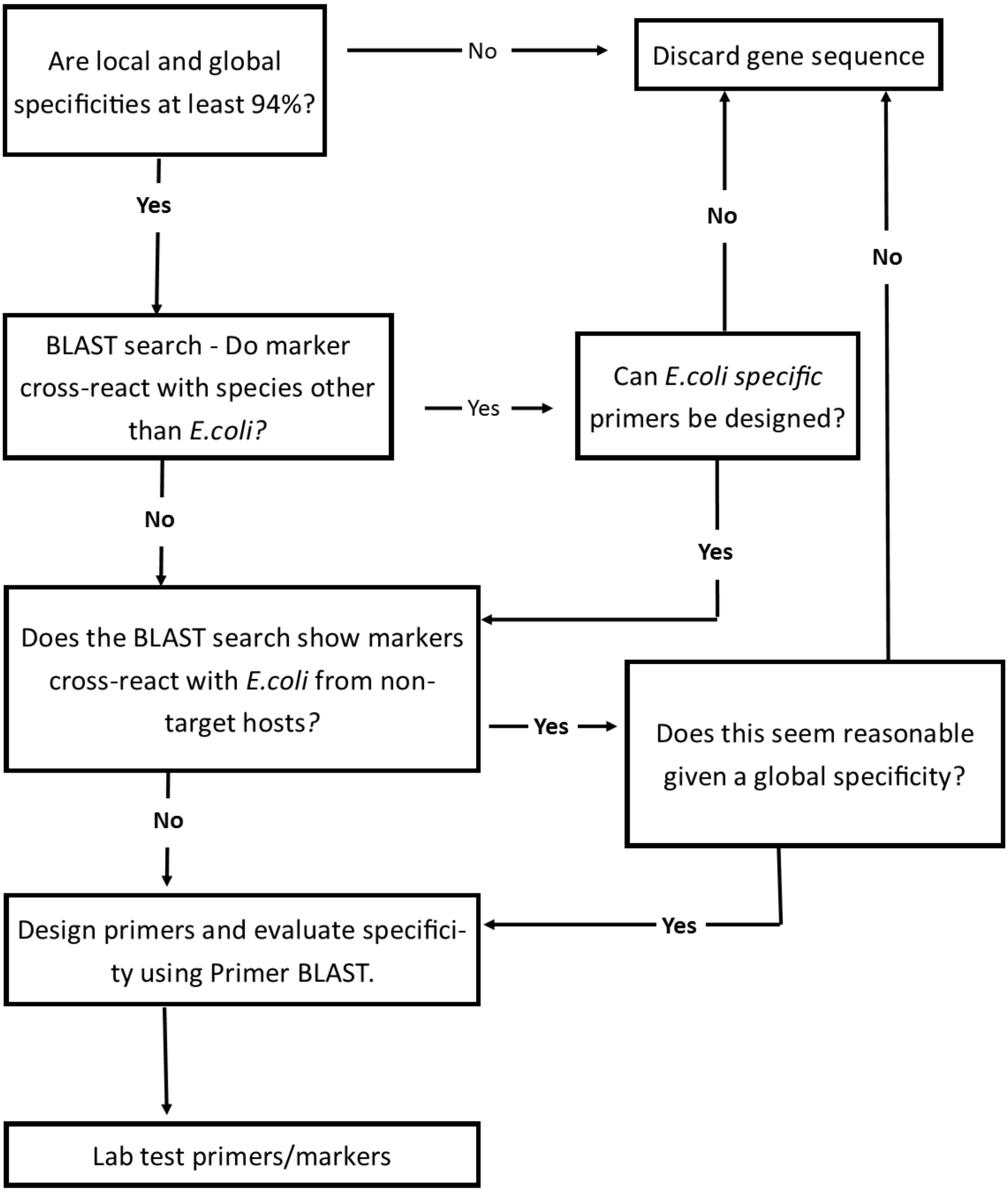

Are local and global specificities at least 94%?
No
Discard gene sequence
Yes
No
No
BLAST search - Do marker cross-react with species other than *E.coli?*
Yes
Can *E.coli specific* primers be designed?
No
Yes
Does the BLAST search show markers cross-react with *E.coli* from non-target hosts*?*
Yes
Does this seem reasonable given a global specificity?
No
Yes
Design primers and evaluate specifici-ty using Primer BLAST.
Lab test primers/markers

Figure 4.1 The criteria used for selecting markers to lab test from LS-BSR outputs

4.2.6 Marker abundance in local wastewater treatment plants

To evaluate the abundance of each marker, SYBR green-based qPCR assays were used to evaluate the abundance of markers in locally sourced sewage. QPCR assays were carried out as previously described (4.2 Quantitative PCR (qPCR)) and optimised by varying the concentration of each primer between 300nM and 900nM, and selecting the concentration giving the lowest cycle quantification (Cq) value with no other issues such as unwanted amplification. The limit of detection (LOD) was defined as the lowest gene copy number where 2 out of 3 reactions are positive. Similarly, the limit of quantification (LOQ) was defined as the gene copy number where 2 out of 3 reactions are positive with Cq values within ± 0.5 (Symonds *et al*., 2016; Hughes *et al*., 2017).

Table 4.4. Details of primers and the target genes used in this study

Primer Name	Primers		Primer Conc. (nM)	Fragment size (bp)	Annealing Temp. (ºC)	Melt Temp. (ºC)	Predicted Gene Function	Reference
Hu_56	Forward	GATGCTTGCAGTTGTCCGAA	ND	226	59	ND	Hypothetical Protein	This Study
	Reverse	CCTTTTCGATTGTGTTTCTGACC	ND					
Hu_100	Forward	ACGGTTATCAGCTCACGTCG	500	98	60	82.00	Hypothetical Protein	This Study
	Reverse	TCGCCCCTCGAAAAGCATTA	500					
Hu_112	Forward	CCCTCAAGCCCCTGATTTCT	ND	155	60	ND	Hypothetical Protein	This Study
	Reverse	ATCTCCCAGTATGCCAGCAG	ND					
Hu_113	Forward	GTGACACATCCAGGCTCCAG	ND	177	53	ND	Acetylxylan esterase	This Study
	Reverse	TAGGCCACGGTACATGAGCA	ND					
Hu_117	Forward	CTCTGGGAATATCACGTTGGAC	ND	78	60	ND	Hypothetical Protein	This Study
	Reverse	ATTCCAGCGTTCAGGATTCG	ND					
Hu_9	Forward	AAGCCAATGATGATGTGGGC	300	163	60	80.50	MFS* protein	This Study
	Reverse	TAGGCCAACTTTCTACCGCA	300					
Hu_42	Forward	GGTGGAACAATAGAGGATGA	500	233	57	79.00	Hypothetical Protein	This Study
	Reverse	CCGCAAGTTTCTCCTGACTC	500					
H8	Forward	ACAGTCAGCGAGATTCTTC	500	177	60	93.00	Sodium/hydrogen exchanger precursor	(Gomi *et al*., 2014)
	Reverse	GAACGTCAGCACCACCAA	500					
H24	Forward	CTGGTCTGGCTTTATAACAC	500	229	60	82.00	Methyl-thioribulose-1-phosphate dehydratase	(Gomi *et al*., 2014)
	Reverse	ATCATTTCCACTTGTCGGG	500					
RodA	Forward	GCAAACCACCTTTGGTCG	300	157	60	85.00	shape-determining protein	(Chern *et al*., 2011)
	Reverse	CTGTGGGTGTGGATTGACAT	300					

**Major facilitator superfamily. ND = Not determined*

4.3 Results

4.3.1 Pipeline validation for markers in the UK

The LS-BSR software can be used to evaluate the presence/absence of known genes in contigs, and assembled sections of DNA (Sahl *et al*., 2014); this feature was used to predict the presence of previously identified human markers, H8, H12, H14, H24 (Gomi *et al*., 2014) in the *E.coli* isolates. LS-BSR correctly identified the presence and absence of markers in locally sequenced isolates, in-line with previous observations (Chapter 3). Coding regions for H8, H12, H14 and H24 were identified independently by the pipeline (Figure 4.1) as highly human associated, and sensitivities and specificities were similar to those previously reported (Table 4.5).

Table 4.5. Sensitivity$_{\text{isolates}}$ and specificity of markers observed in Japan (Gomi et al., 2014), Australia (Warish et al., 2015), the UK (Chapter 3), Bangladesh (Harada et al., 2018) and predicted by the database.

	Japan		Australia		UK			Bangladesh		Global Database	
Marker	**Sensitivity (isolates)**	**Specificity (isolates)**	**Sensitivity (isolates)**	**Specificity (Faecal)**	**Sensitivity (isolates)**	**Specificity (isolates)**	**Specificity (faeces)**	**Sensitivity (isolates)**	**Specificity**	**Sensitivity (isolates)**	**Specificity (isolates)**
H8	50.0	99.2	45	94.0	10.0	99.2	100.0	16.3	ND	13.0	100.0
H12	30.0	100.0	14	85.0	3.3	100.0	100.0	ND	ND	5.8	99.4
H14	30.0	98.3	ND	57.0	16.7	93.3	93.3	ND	ND	11.5	93.0
H24	36.7	99.2	ND	72.0	50.0	95.8	93.3	ND	ND	25.9	93.0

4.3.2 Human biomarkers

The coding sequences (CDSs) were ranked according to their estimated specificity and sensitivity to humans. Unfortunately, there appears to be a trade-off between the sensitivity$_{\text{isolates}}$ and specificity$_{\text{isolates}}$ of CDSs (Pearson's coefficient = -0.752, $p < 2.2\times10^{-16}$) with the sensitivity explaining around half of the variability in specificity of coding regions (Figure 4.2). Figure 4.2 shows that the ideal human *E.coli* marker, with 100% specificity and

100% sensitivity does not exist in this database. A low marker sensitivity has been suggested to limit the detection of sequences directly from environmental samples and lead to false negative results (Gomi *et al.*, 2014; Chapter 3). Therefore, markers with the highest sensitivities, with a specificity of > 95% were prioritised. The selection of 95% specificity is arbitrary, however, in MST 80% is often used as recommended value for markers (Harwood *et al.*, 2014), and 95% was used as a way to whittle down the large number of CDSs with a specificity of > 80%.

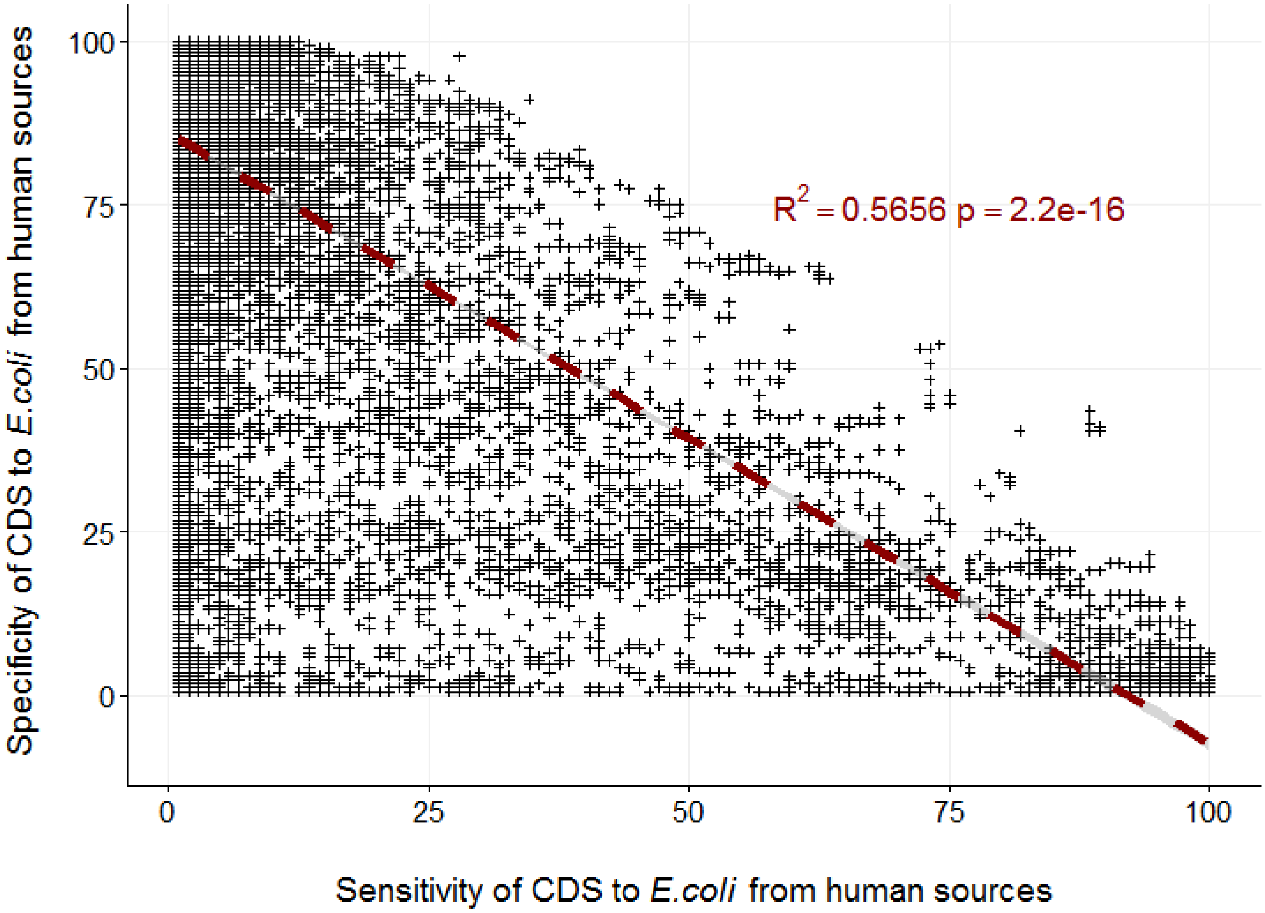


Figure 4.2. Sensitivity and specificity for each coding sequence, predicted by the LS-BSR software. Red dotted line represents a linear regression curve with the equation y=-0.93x + 85.9

Using the process outlined in Figure 4.1, 7 coding sequences were selected (Table 4.4) for testing. Of the seven, only three were deemed to show promise as practically usable markers after *in vitro* testing. These were the Hu100, Hu9 and the Hu42 markers. The Hu42 marker showed some cross-reactivity with *Salmonella,* although to a lesser extent than for H8. The

Hu56 and Hu112 primer pairs were discarded after initial testing as PCR with Hu56 led to an amplified fragment that was not of the expected length, and amplification of Hu112 led to two PCR products. The Hu117 was discarded due to low specificity, cross-reacting with dog faeces. The Hu113 primer set was removed from further analysis, as melt curve analysis revealed a second peak, which would indicate an unwanted amplification product. Two other Hu113 primer pairs were tested (data not shown) and both were unsuitable, one resulting in an unacceptable melt curve, and producing unwanted amplification during PCR; this marker was therefore removed from further analysis.

All of the selected human markers were noted to have high specificity to sewage. The Hu100, Hu9 and Hu42 markers had 95% specificity (Table 4.6). The Hu117 marker was removed from further testing as it had a specificity <95%, cross reacting with 40% of dog faeces and 10% of sheep faeces. No *in vitro* cross-reactivity was noted in the Hu113 marker.

Table 4.6. The sensitivity and specificity of markers determined in vitro and from the database.

Marker	**In vitro sensitivity (isolate)**	**(Global) Database sensitivity (isolate)**	***In vitro* sensitivity$_{\text{faecal}}$ (to sewage)**	**In vitro specificity (isolate)**	**Global Database specificity (isolate)**	**Faecal sources of cross-reactivity**
Hu9	6.7%	9.6%	83.3%	95%	94.9%	Sheep (n=2) Dog (n=1)
Hu42	3.3%	5.8%	100.0%	95%	95.8%	Chicken (n=3)
Hu100	16.70%	27.9%	100.0%	95%	97.9%	Sheep (n=1) Dog (n=2)
Hu113	23.3%	15.8%	100.0%	100%	97.5%	
Hu117	23.0%	26.7%	100.0%	92%	94.6%	Dog (n=4) Sheep (n=1)

4.3.3 Marker frequency and abundance in local wastewater treatment plants using qPCR

All markers were present in all sewage samples except the Hu9 marker which was absent (below the LOD) from two samples. The geometric means of the markers (Figure 4.3) was in the same rank order as the sensitivities derived from the database (Table 4.3), i.e., Hu100 had the highest and Hu42 the lowest abundances, respectively, with the exception of H8. There was a statistically significant difference in marker abundance between H100 and H9 (p=0.016). The concentrations of individual markers varied by about an order of magnitude between different works. In addition, there was no significant difference between the abundance of HF183 and the H8, Hu100 and H24 human markers.

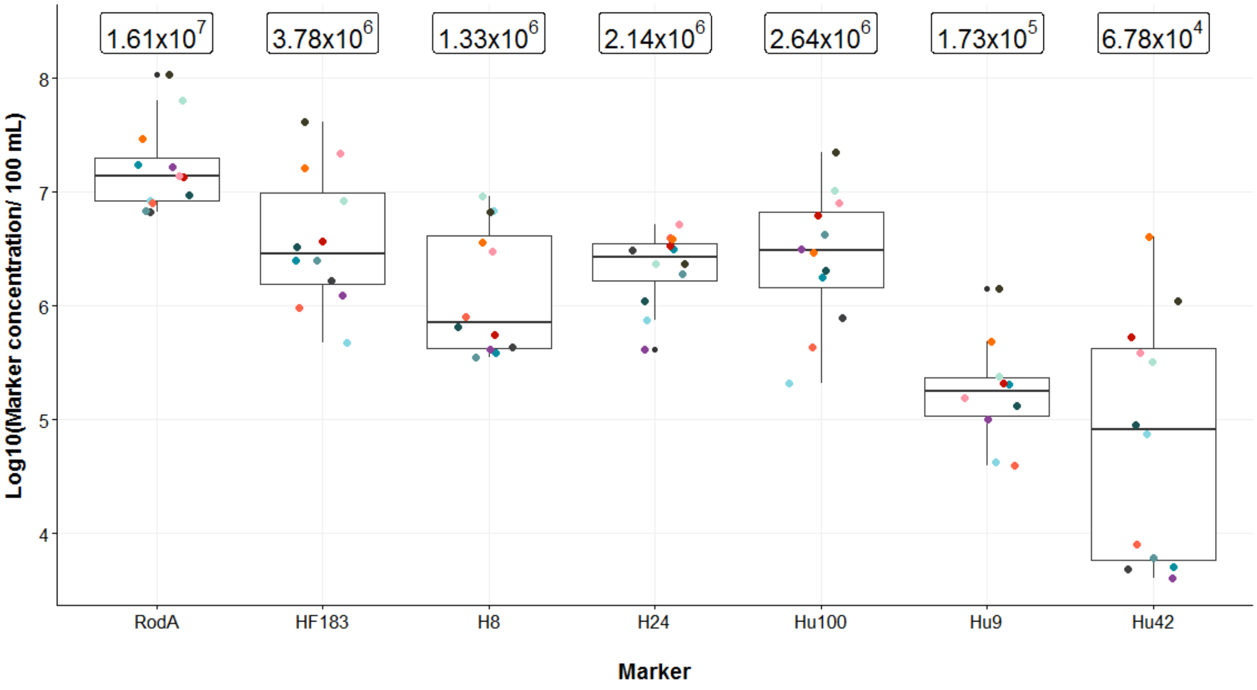


Figure 4.3. Abundance of total E.coli (RodA) and human markers in 12 small decentralised and 1 large wastewater treatment plants. Numbers at the top of the figure show the geometric mean of each genetic marker. The coloured dots represent individual data points. Two Hu9 sample points below the LOD are not shown on the graph.

While both the abundance of markers and total *E.coli* (*RodA* copy numbers) varied, the marker abundance did not appear to reflect total *E.coli* abundance between WWTPs. The range in sensitivity$_{\text{isolates}}$, the proportion of *E.coli* containing a marker, of each marker spanned an order of magnitude (Table 4.7). The highest relative abundance in the total *E.coli* population from a single WWTP was noted for H8 (81%, WWTP 3, Table 4.4), the Hu100 and H24 most commonly had the highest sensitivity$_{\text{isolates}}$ in individual treatment plants. On average, across all WWTPs, Hu100 had the highest relative abundance. Similarly to the absolute abundance (Figure 4.3), the rank order of the geometric means (adjusted $R^2 = 0.91$, $p = 0.00795$ using linear regression) and medians (adjusted $R^2 = 0.98$, $p = 0.001188$ using linear regression), was the same as predicted by the database; geometric means and medians are less susceptible to skew than the arithmetic mean ($R^2 = 0.84$, $p = 0.01858$ using linear regression). It should, however, be noted that two data points were excluded from Figure 4.3 and calculations in Table 4.7 (WWTP 1 and 7); if these points were included, for example as half the limit of detection, then this relationship would likely change for Hu9 and Hu42.

Table 4.7. Sensitivities of E.coli markers tested at each wastewater treatment plant (WWTP)

	Marker relative abundance in *E. coli* (%)				
WWTP	**H8**	**H24**	**Hu100**	**Hu9**	**Hu42**
1	6.49	45.85	11.66	ND	0.07
2	2.22	17.95	10.12	1.18	0.03
3	80.91	8.94	2.48	0.50	0.88
4	4.10	25.02	46.29	1.55	3.90
5	6.92	11.90	21.79	1.41	0.96
6	2.50	2.49	19.21	0.61	0.02
7	5.21	27.86	61.90	ND	0.09
8	9.90	48.84	5.38	0.49	0.10
9	12.17	13.11	10.01	1.65	13.79
10	21.75	37.44	57.58	1.12	2.79
11	6.13	2.17	20.80	1.33	1.02
12	14.39	3.70	16.10	0.38	0.50
Mean	14.39±21.68	20.44±16.57	23.61±20.25	1.022±0.48	2.01±3.90
Geometric mean	8.25±2.68	13.25±2.59	16.35±1.74	0.903±1.73	0.413±7.84
Median	6.71	15.53	17.66	1.15	0.69

ND = Not Determined, due to marker concentrations being below the assay limit of detection.

The large range in relative proportions of markers is also reflected in Figure 4.4, which compares the marker against the total marker abundance. Human *E.coli* markers explained as little as 1.3% (H24) and up to 73.3% (Hu9) of the variance in the abundance of total *E.coli*, measured through *RodA* gene copy numbers (Figure 4.4). The HF183 marker explained a greater proportion of the variation in total *E.coli* than all of the *E.coli* markers, except Hu9 (Figure 4.4).

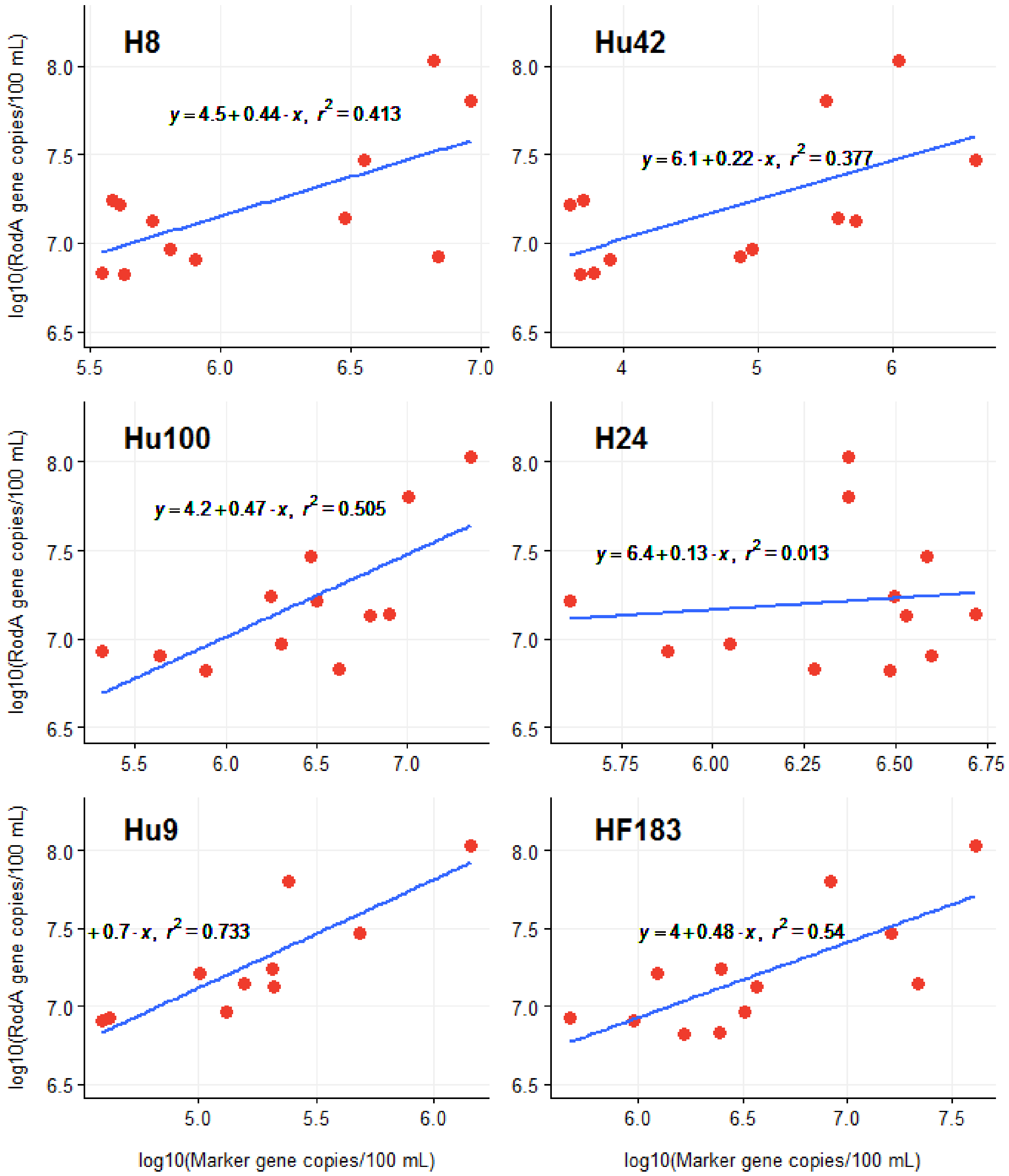


Figure 4.4. The relationship between the concentration of each marker and the total E.coli concentration in sewage samples - measured through RodA gene copy numbers.

4.4 Discussion

4.4.1 An E.coli biomarker for North East England

The motivation in undertaking this study was to identify new human-associated biomarkers, particularly those most suitable for use in North-East England. While there was no significant difference in the abundance of the Hu100, H8 and H24 *E.coli* biomarkers, the Hu100 marker, most often, had the highest sensitivity$_{\text{isolate}}$, i.e. represented the largest proportion of *E.coli*, across all 12 WWTPs and had a high specificity$_{\text{faecal}}$ (95%, Table 4.6), Hu100 may, therefore, be considered the best marker for MST in the North East of England. The abundance of the Hu9 marker explained the greatest variation in total *E.coli* abundance (R^2 =0.778), than other markers ($R^2 \approx$0.500) and might, therefore, represent total *E.coli* better; however, the abundance of Hu9 was approximately one order of magnitude lower than other markers and was absent from two wastewater treatment plants (Sensitivity$_{\text{faecal}}$ = 83.3%, Table 4.6). The sensitivity$_{\text{isolates}}$, and absolute abundance of *E.coli* biomarkers were considered critical in this study. They were noted, in a previous study (Chapter 3), to limit their efficacy as MST markers, leading to false negative results. Interestingly, while the Hu9 marker had the highest specificity in the local database, it had the lowest sensitivity$_{\text{faecal}}$ (Sensitivity$_{\text{faecal}}$ = 83.3%, Table 4.6) and a low abundance (Geometric mean = 1.73×10^5, Figure 4.3); conversely, Hu100 was not present in the local database, but had a high abundance, highlighting the importance of using a large, and ideally, global database where possible.

The global database suggests that the Hu100 biomarker may be useful on a global scale, as opposed to the regional scale evaluated here. The Hu100 biomarkers was found to perform well in subsequent studies in Thailand (Mrozik *et al.*, 2019), where it outperformed H8 in terms of its prevalence and correlation with bacteria from potentially pathogenic genera in waters used for aquaculture. The Hu100 biomarker was also used to identify sewage contamination of drinking water sources in Nepal (Acharya *et al.*, In prep). This, and other studies (Mrozik *et al.*, 2019*;* Chapter 6) suggest that Hu100 may be useful on a global scale to link MST conclusions to regulatory FIOs. However, more studies assessing the specificity of Hu100, across a wider range of faecal sources, such as rabbit faeces in which HF183 was recently identified to cross-react (Nshimyimana *et al.*, 2017), are desirable if Hu100 is to be used for regulatory and decision making purposes.

4.4.2 Comparison with HF183

The similarity in the mean abundances of H8, H24 and Hu100 to the commonly used, HF183 marker (Figure 4.3) suggests that the use of *E.coli* biomarkers for MST studies may have utility. The similarity in abundances between H8 and HF183 was also observed in Australia (Hughes *et al*., 2017), although notably, the concentrations of the H8 and HF183 markers in this study were around two orders of magnitude lower than those reported in Australia (Hughes *et al*., 2017); this discrepancy is likely due to dilution as a result of different climates i.e. winter in a combined sewer system in the UK compared to summer in Brisbane, Australia.

Surprisingly, the abundance of HF183 marker explained a similar amount of the variation in total *E.coli* abundance as the H8 and Hu100 markers. While this may suggest that there is little benefit to using *E.coli* biomarkers over HF183, the dissimilarity in the environmental persistence and behaviour of *E.coli* and *Bacteroides spp.* may make *E.coli* biomarkers more useful for studies attempting to relate conclusions drawn from MST studies to regulatory FIOs. The large amount of variability in the abundance of HF183 is also noteworthy (Figure 4.3); this was not dissimilar to Hu100, H8 or H24 suggesting that using *E.coli* biomarkers for MST and subsequent decision making is similar to using HF183. It highlights the need to use more than a single marker for the detection of sewage-borne contamination to avoid the potential of false negative results. Future studies comparing the use of Hu100, or indeed other *E.coli* biomarkers, to other common biomarkers such as HF183 in catchment studies are paramount to determining the usefulness of *E.coli* biomarkers; since a large number of markers are rarely tested in real-life applications (Harwood *et al*., 2014).

4.4.3 The stability of E.coli biomarkers

The stability of marker performance, both geographically and temporally is an important consideration in the performance of markers, particularly for decision making. The geographical stability of markers is often highlighted as unknown in MST studies (Harwood *et al*., 2014), although, it is important for regional or national decision making. For example, a marker whose sensitivity shows a low geographical stability may be present in a large number of individuals in one region, and absent from a large number in another; an identical amount of pollution entering waterbodies from each source would therefore be either overestimated or underestimated using that marker, respectively. Here, the sensitivity$_{\text{isolates}}$, the proportion of *E.coli* containing a marker, of all *E.coli* biomarkers varied greatly (Table 4.7). This large variation in sensitivity$_{\text{isolates}}$ is interesting and could be due to the irregular distribution of markers between communities or the influence of non-human faecal sources at small decentralised works. The presence of non-human faecal sources is possible given the rural nature of decentralised works and may present a challenge for MST in rural areas. However, it seems more likely to be due to the distribution of *E.coli* containing marker genes in human populations since the variance of marker abundance between plants appears to be slightly larger than that of the *RodA* gene. However, further research evaluating the presence/absence of markers in faeces, rather than sewage, would be required to support this hypothesis.

While the variation in sensitivities may be dampened in sewage from large communities (e.g., at large WWTPs), the large variation in sensitivities (Table 4.7) highlights the need for caution when using the proportion of *E.coli* from humans to compare sampling sites, particularly where small communities or small decentralised WWTPs contribute to the pollution load. For priority catchments, long-term monitoring of the proportion of *E.coli* containing a biomarker could help decision-makers overcome the variability in marker sensitivity, although this could be expensive. Future MST studies are advised to use the total abundance of different *E.coli* biomarkers to compare sample sites or catchments.

4.4.4 Monitoring E.coli biomarkers in environmental samples

Previously, *E.coli* biomarkers have been enumerated through culture-based techniques (Gomi *et al*., 2014; Warish *et al*., 2015) and PCR; while this is in-line with regulatory techniques (e.g., the BWD (2006/7/EC)), enumeration of *E.coli* biomarkers in this fashion is very expensive, due to the labour required (Chapter 7), and could not be feasibly adopted by the Environment Agency (Porter, 2016). The combined use of culturing and qPCR enumeration of *E.coli* biomarkers reduces the labour somewhat (Chapter 3), however, culturing may limit the detection of *E.coli* biomarkers especially where human sources are not the main source of pollution. Enumeration of *E.coli* by culture-based methods is limited to 100 isolates per plate and the geometric mean values of the sensitivity$_{\text{isolates}}$ of the H8, H24 and Hu100 markers was 8.25±2.68%, 13.25±2.59% and 16.35±1.74% respectively; therefore, if sewage contributes only 10% of the total *E.coli* to a water body, only a single isolate per 100 *E.coli* on each plate may be expected. Since 100 *E.coli* are rarely cultured on each plate, there is potential for false negative results. Direct enumeration of markers, using qPCR for example, may overcome these sensitivity$_{\text{isolates}}$ issues, especially since the H8, H24, and Hu100 markers had a similar abundance to HF183, currently the most commonly used MST marker (Harwood *et al*., 2014).

4.4.5 A database approach to biomarker identification

The interrogation of a database of *E.coli* genomes from known hosts was valuable in choosing which of the ~15,000 CDS to test in the laboratory. The rank-order of the sensitivity$_{\text{isolates}}$ for markers predicted by the database was the same as the geometric mean and median vales observed *in vitro*. However, lab based validation was essential to identify those regions which were most useful as MST markers, due to the variation in sensitivity between WWTPs, and the low availability of *E.coli* genomes from non-human hosts. Laboratory testing using faecal samples, as opposed to *E.coli* isolates, was also valuable to screen a large number of *E.coli* and identify unwanted amplification from non-target sequences. As the number of publically available sequencing projects increased, due to reducing sequencing costs, database approaches, for example, using LS-LSB (Sahl *et al*., 2014) to compare large numbers of genomes rapidly, are likely to become increasingly important in biomarker discovery. Recently, the interrogation of metagenomics datasets led to the discovery of crAssphagh (Dutilh *et al*., 2014), and subsequent marker genes which are potentially human-associated (Stachler *et al*., 2017); however, this involved the *in vitro* testing of 57 primer pairs to identify two potential markers, a database approach may, therefore, have reduced the costs and labour involved in laboratory testing. It is noteworthy that the poor metadata associated with many publically available genomes on NCBI genbank database (Benson *et al*., 2013) may limit the usefulness of publically available data. For example, a common issue for this study was a lack of accurate information of the host organism from which the bacterium was isolated, with a number of genomes isolated from “genomic DNA”. Nonetheless, this database style approach may lend itself to identifying host-associated genes for MST in other organisms, such as Acinetobacter or *Enterococcus spp.*, a common faecal indicator which has previously been noted to show host-specificity.

4.5 Conclusions

A database approach using 241 publicly available and 23 locally assembled *E.coli* genomes allowed efficient identification and evaluation of human-associated coding sequences. These CDSs included the identification of previously identified markers (H8, H12, H14 and H24) which supports their use as human associated markers.

A new human-associated marker, Hu100, was identified. While the Hu100 marker was not significantly more abundant than the H8 and H24 markers, it represented the largest proportion of *E.coli* in rural, decentralised treatment system in the North East of England, most often.

There was a large variability in the proportion of *E.coli* containing human-associated markers between geographically close communities which limits the usefulness of human-associated markers for decision-making. We strongly recommend that future studies do not use the marker/total-*E.coli* ratio, rather, use the total abundance of different markers since this may be more comparable between different locations.

Taking a database approach may become invaluable for identifying further markers in a cost-efficient manner. It allowed tens of thousands of potential sequences to be rapidly screened and a small number selected for laboratory validation. The further use of this approach using the genomes of other regulatory faecal indicator organisms or difficult to culture organisms which have been assembled from metagenomics data would be exciting.

Chapter 5 Evaluating the effect of library composition on community-based MST

5.1 Introduction

Microbial source tracking (MST) describes a range of methods which use microbes and their communities to identify, and often apportion, the sources of faecal pollution contaminating water bodies or food produce. Currently, the majority of MST techniques take advantage of the rapid nature of qPCR to detect genetic markers which have been shown to be highly-associated to the faeces of a particular host (Figure 1.4). Advancements and reductions in costs of high throughput sequencing (HTS) technology now allow large-scale interrogation of microbial communities in different environments. MST researchers have started to take advantage of advancements in HTS, and the dissimilarity between faecal bacterial communities of different hosts to differentiate sources of faecal pollution based on such differences (Brown *et al*., 2017; Henry *et al*., 2016). These methods have been collectively termed community-based MST (Unno *et al*., 2018).

The most common method of community-based MST involves the use of SourceTracker (Knights *et al*., 2011), a software which takes a Bayesian approach to estimate the proportion of taxa from ‘source’ communities (e.g., in faecal sources) contributing to ‘sink’ communities (e.g., in lake water). To make these estimations, SourceTracker requires the input of samples from known sources to build the sink community; these source samples are collectively termed the faecal taxon library (FTL) (Brown *et al*., 2018), a throwback to traditional MST methods which required a large library of organisms from likely faecal sources of pollution. (1.4.1 Library-dependent methods). Studies have begun to explore how the composition of the FTL can affect SourceTracker source predictions (Hägglund *et al*., 2018; Staley *et al*., 2018), although a number of aspects remain unclear or have yet to be explored.

As with previous library-dependent approaches, the number of samples required for each source of faecal pollution (i.e., the size of the FTL) remains unclear. Previous suggestions, that more than ~10 samples of each source are required to avoid false negatives (Brown *et al*. 2017) have been largely ignored, even by the same authors (Staley *et al*., 2018). Moreover, the claim that more than ~10 samples is required does not reflect other studies which have used a single sample successfully (Henry *et al*., 2016), and the same authors later suggest that fewer samples would be adequate for community-based MST (Staley *et al*., 2018). Currently there is no consensus on the number of samples required to define each source (e.g., cow faeces, sewage etc.) in a FTL; although this may depend on the variability of the microbial community within a single source since SourceTracker averages the relative abundance of OTUs in all samples from a single source prior to analysis. Indeed, SourceTracker predictions were shown to be more reliable when sources showed low intragroup variability (Brown *et al*. 2018).

In addition to the size, the appropriate composition of the FTL (i.e., the number of different sources of faecal pollution) is also unclear. Previous studies have demonstrated the ability of SourceTracker to accurately identify pollution sources when the FTL contains known sources of pollution (Henry *et al*., 2016; Staley *et al*., 2018). Brown *et al.*, (2018) also noted improved reliability when FTLs contained only sources known to be contaminants in sink samples; however, this rather negates the purpose of MST - where pollution sources are known, there is unlikely to be a need for MST. The inclusion of sources in the FTL which are not contaminating sources in the sink samples can cause false positive results (Henry *et al*., 2016; Unno *et al*., 2018). False negative results have previously been identified by calculating the relative standard deviation (RSD) of the predicted contribution of each source to each sink sample across 5 runs of SourceTracker, with samples with an RSD greater than 100% considered false positives (Henry *et al.*, 2016). In addition, reporting only predicted source contributions above a cut-off of 1% reduces the chance of false positive results (Unno *et al.*, 2018). Nevertheless, there is still a concern that sources of pollution which have similar microbial communities (i.e., share a large proportion of taxa), for example sheep and cattle faeces, may cause false positive results. Currently, there is no way to evaluate the effect of using different faecal sources with similar microbial communities; as a result, studies have either combined sources with similar communities (Staley *et al*., 2018), or just accept that the possibility of conflation of sources by SourceTracker exists (Hägglund *et al*., 2018). The conflation of the faecal sources included in the FTL and background sources has been recently highlighted as a concern (Hägglund *et al*., 2018). Hägglund *et al*. (2018) described a range of methods to take account of indigenous taxa, leading to an improved correlation between SourceTracker predictions with culturable *E.coli* counts, while Unno *et al*. (2018) simply recommend the inclusion of an environmental source in the FTL. This, however, may be impractical in the UK since it is often impossible to obtain an unpolluted environmental sample (Chapter 3). Gaining a better understanding of the impacts of this background microbiota is important to better appreciate the limitations of community-based MST, or specific FTLs.

The geographical locality of samples selected for the FTL is also of concern. Using only non-local samples in the FTL has been observed to result in either false negative results or an underestimation of the relative contribution from a source (Staley *et al*., 2018). Staley *et al*., (2018) used local samples of sewage obtained from Australia, and non-local samples obtained from the USA. Though little is understood about the biogeography and variability of microbial communities in sewage, the microbial composition of sewage has been observed to vary between cities in the USA (Shanks *et al*., 2011). Interestingly, Staley *et al*., (2018) found that the inclusion of non-local sewage samples in the FTL did not significantly ($P > 0.79$) impact the reliability of SourceTracker predictions when local sewage samples were included in the FTL. For regional water companies such as Northumbrian Water, this is important, since a single library that which can be used across a particular region, such as the North East, would reduce future MST costs greatly and allow comparison of source contributions across catchments and studies.

To evaluate SourceTracker, previous studies have focused on laboratory prepared samples (Henry *et al*. 2016; Staley *et al*. 2018). While this approach is valuable in understanding the outputs of SourceTracker for real world applications, it is difficult to assess the ability of SourceTracker to identify and quantify different sources for MST purposes; for example, to assess the suitability of an FTL. This is due to the costs associated with sequencing sources and the difficulty in mixing sources to the desired relative contributions from individual sources as bacterial cell densities may vary greatly between different sources (Staley *et al*. 2018). Furthermore, the use of laboratory prepared samples is not conducive to testing new sources rapidly, since this would require sources to be mixed in different proportions, and the DNA extracted and sequenced before these determinations could be made. This goes someway to explain why few studies have examined the ability of SourceTracker to identify different mixtures of faecal sources, particularly at low levels of contamination, commonly contributing to the pollution of environmental waters.

Understanding the behaviour of SourceTracker with different configurations of FTLs (in terms of size, composition, and locality of samples), when identifying low levels of pollution are particularly critical since it may be difficult to discriminate low levels of pollution from false positive events. A rapid and repeatable method is, therefore, required to assess the composition of FTLs, assess the ability of SourceTracker to distinguish between sources, and assess the potential for cross-reactivity between sources included in the FTL. This study aims to answer three questions:

1. Is there a discernible effect to using a background sample (e.g. sea water with no faecal contamination) as a source? (Experiment 1)
2. Can a single faecal library of sewage represent a single region (e.g., the North East of England) for use in community-based MST? (Experiment 1)
3. What is the best strategy of incorporating samples from different hosts which have similar bacterial communities? (Experiment 2)
4. Is there potential for cross reactivity between sources when using the entire FTL? In the Morland case study (Chapter 3), chicken faecal contamination was identified in a larger number of samples than expected. (Experiment 3)

5.2 Methods

5.2.1 Sample collection

Samples were collected (Methods and Methodology, 1.3 Raw sewage) and transported (Methods and Methodology, 1.1 Sample preservation and transport), from 15 wastewater treatment plants (WWTPs) in the North East of England (Table 5.1). Smaller (< 2000 PE), decentralized treatment plants were prioritized in sampling since these are likely to have the greatest variability and better reflect the problems with identifying urban diffuse pollution in catchments than larger WWTPs would.

Table 5.5.1. Size and location of wastewater treatment plants (WWTPs) sampled for experiment 1.

WWTP code	Size (Population Equivalent)	Area
A	2,211	Northumberland
B	89	Northumberland
C	199	Newcastle
D	72	Newcastle
E	262	Newcastle
F	7,148	Durham
G	79	Durham
H	128	Durham
Q	110	Durham
J	161	Durham
K	1,003,785	Newcastle
L	8,707	Northumberland
M	184	Northumberland
O	2,080	Northumberland
P	22,493	Durham

In addition to the human source, 62 potential sources of non-human pollution (Table 5.2) were collected as previously described (2.1.2 Faecal samples).

Table 5.2. Number and location of the non-human faecal sources used in this study.

Host	Area	Number of samples
Chicken	Northumberland Durham	5 5
Cow	Northumberland Durham	5 6
Sheep	Northumberland Newcastle	5 6
Pig	Northumberland Newcastle	5 5
Horse	Northumberland Newcastle Durham	4 3 3
Dog	Northumberland Newcastle	5 5

Two water samples, one sea water and a fresh water sample, were collected as previously described (2.1.4 Environmental water samples) from the Seaton Sluice catchment (Chapter 6). These water samples were selected for this experiment as both showed little faecal pollution; the sea and river water samples contained 0 and 1 *E.coli* per 100 ml, respectively, and both samples were below the limit of detection for *RodA* (total *E.coli*) and the human markers (HF183 and Hu100), and an initial SourceTracker run predicted a source contribution of less than 0.05% when SourceTracker was run with an FTL using all faecal samples from the Seaton Sluice study (Table 5.2).

5.2.2 DNA extraction and sequencing

For sewage samples, DNA was extracted as previously described (2.3.4 DNA extraction from environmental waters), with the following modifications. Between 25 and 100 mL of WWTP post-screened influent was filtered, depending on the dilution due to previous rainfall.

For faecal matter collected from non-human sources, DNA was extracted from 150-300 mg of fresh faeces as previously described (2.3.3 Extraction of DNA from faecal samples).

5.2.3 Bioinformatics

Sequence processing

Sequences were processed using the DADA2 plugin (Callahan *et al*., 2016) to the QIIME2 package (Caporaso *et al.*, 2010; Caporaso, 2018) as previously described (2.6.5 Analysis of data from Illumina sequencing). To prepare data for statistical analysis in the Phyloseq package (McMurdie & Holmes, 2013), the following modifications were made to the OTU and taxa tables exported from QIIME2. The heading in the OTU column of the OTU-table was changed from "#OTUID" to "OTUID". The heading in the feature column in the taxa-table was changed from "Feature ID" to "OTUID". The OTU-table and taxa-table were merged by the OTUID column in the R software (R Team, 2017) to produce an OTU-table similar to that output given by QIIME1. The new OTU-table was then imported to the Phyloseq package in R, according to the manual (McMurdie & Holmes, 2013). Two faecal samples, a single pig and cow sample were removed from analysis because they were dominated by a single OTU.

In all instances, SourceTracker was run five times and the relative standard deviation was calculated (Henry *et al*., 2016) to indicate the level of confidence in SourceTracker estimates (Brown *et al*. 2018).

5.2.4 Simulating sink samples

To evaluate SourceTracker, simulated-samples were made by mixing samples from faecal sources with either seawater or river water at known proportions. The code to make the simulated-samples is available in Appendix D.1. Briefly, the processed reads were imported into the Phyloseq package (McMurdie & Holmes, 2013) in R, and the selected faecal and water samples were subsampled using a probability weighting equal to the desired mixing proportions (e.g., 5% sewage sample #1, 5% sewage sample #2, 5% cow faecal sample #1, and 85% seawater sample). Samples were subsampled with replacement to a total depth of 50,000 reads. Three simulated samples were created for each desired mixing proportion to account for variation in the random sampling technique, and the potential influence of rare OTUs on the SourceTracker analysis.

5.2.5 Experiment 1 – Is a single sewage FTL adequate to represent sewage from a particular region?

Firstly, to determine whether the inclusion of a background sample, such as seawater or river water, in the FTL significantly changes SourceTracker predictions when sewage is the contaminating source, the procedure outlined in Figure 5.1 was followed. Simulated-samples were created in triplicate by mixing a sewage sample with either a sea or river water sample, at proportions of sewage between 95% to 5% sewage in 5% intervals, and 4%, 3%, 2%, 1%, 0.1%, 0.01% and 0.001% contributions of sewage to the simulated-samples. SourceTracker was then run five times using these simulated samples as the sink communities with an FTL containing only the sewage sample as a source. SourceTracker was then run another five times using the same simulated samples with an FTL containing both the sewage and the seawater or river water sample as sources. The mean predicted values of sewage contribution were calculated for each simulated sample and compared to the expected (real) proportional contribution of sewage in the simulated sink community. A pair-wise t-test was used to determine if there was a difference between the means achieved when a background sample is included in the FTL and when it is not.

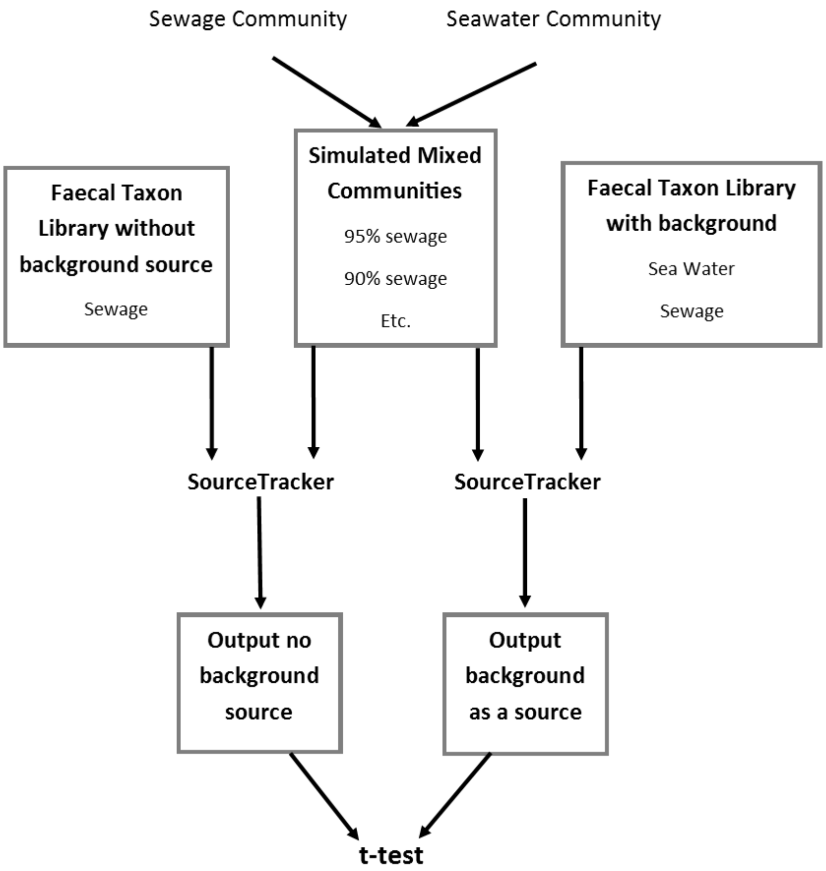

Sewage Community
Seawater Community
Faecal Taxon Library without background source
Sewage
Simulated Mixed Communities
95% sewage
90% sewage
Etc.
Faecal Taxon Library with background
Sea Water
Sewage
SourceTracker
SourceTracker
Output no background source
Output background as a source
t-test

Figure 5.1. Outline of the source-sink experiment when seawater was used as the environmental background.

To determine if an FTL could represent a region (e.g., the North East England) the Jenson-Shannon index was, firstly, used to assess the dissimilarity between samples from different faecal sources, and those from the same host. A similarity matrix was created, using taxa in the 5 most abundant phyla (12078 OTUs), using the Jensen-Shannon divergence with the Phyloseq package in R (McMurdie & Holmes, 2013). The adonis and betadisper functions in the Vegan package (Oksanen *et al.*, 2018) were used to test whether faecal communities from different hosts shared a common centroid (i.e. distance to a notional centroid of each faecal community type) and to evaluate whether intragroup variability was similar between communities from different hosts, respectively. The level of significance was assessed by performing permutation tests.

Two sets of simulated-samples were created, one using taxa from the sewage sample with the greatest dissimilarity to the other samples (Q, Table 5.1) and one set using taxa from a sewage sample which was identified as similar to most other sewage samples (A, Table 5.1) using the Jensen-Shannon divergence analysis. Simulated-samples were created in triplicate by mixing the taxa from one of these sewage samples (Q or A (Table 5.1)) with taxa from either the sea or river water sample at varying proportions between 95% sewage and 0.001% sewage (the rest of the microbial community was made up of river or sea water taxa). SourceTracker was then run five times using these simulated-samples as the sink communities with a FTL containing the background sample (sea or river water) and either i) only the sewage sample used to make the simulated-samples, ii) between two and 14 sewage samples (i.e., 13 different FTLs were tested using n = 14, 13, 12 etc. sewage samples), not including the sample used to make the simulated-samples, or iii) all 15 sewage samples (Table 5.1). The mean predicted values of the contribution of sewage for each scenario was calculated for each simulated sample, across the five runs of SourceTracker. These were compared to the expected contribution (i.e., known contribution) in the simulated sink community. A t-test was used to evaluate the differences between the means.

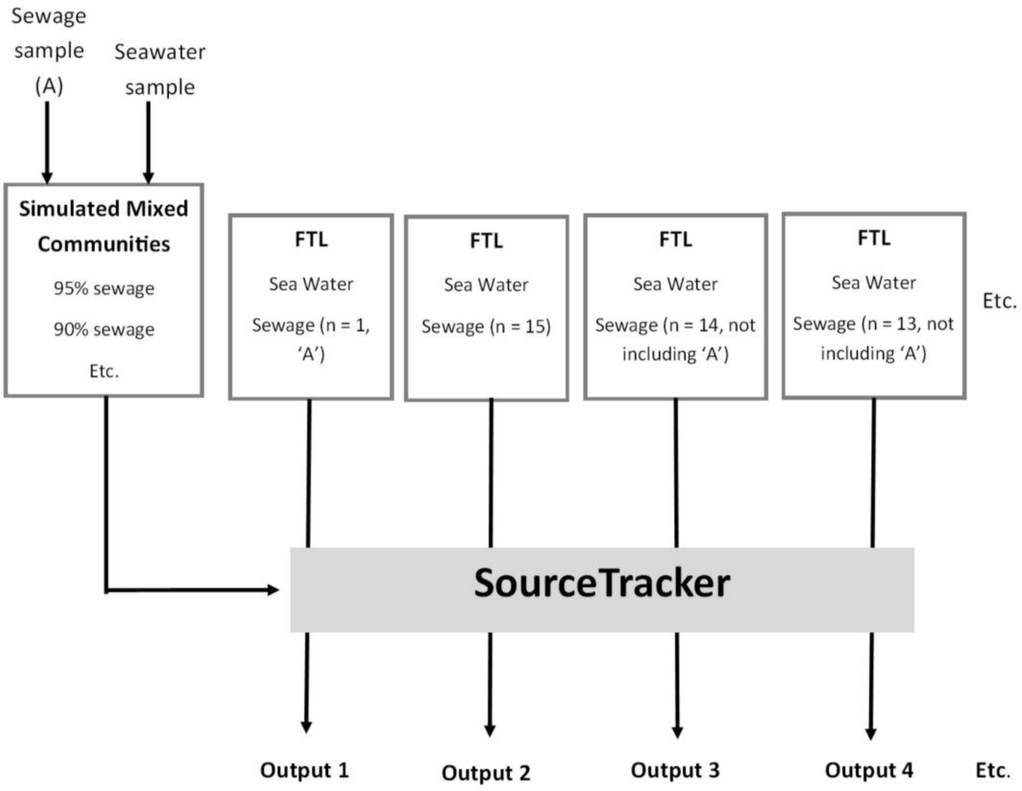


Figure 5.2. Outline of an example library size experiment using the taxa from sewage sample A (Table 5.1) and the seawater sample.

5.2.6 Experiment 2 – How to incorporate similar bacterial communities from different hosts into the FTL?

To determine the best approach to incorporate faecal sources with similar microbial communities (those which share a number of taxa) into an FTL, two sets of similar sources were selected for this experiment: i) sheep and cow faecal sources, which were identified as being similar in experiment 1 in this study and a previous study (Hägglund *et al*., 2018), and ii) sewage samples, where a single sewage sample was observed to be dissimilar to all other sewage samples in experiment 1.

Sheep and cow sources

Simulated-samples were built using either three cow or three sheep faecal samples and river water at proportions between 95% and 0.001% of faecal contributions. Firstly, simulated-samples were made, to represent sink samples, using river water and only cow sources. SourceTracker was run five times with an FTL containing river water and either i) the three cow samples used to make the simulated-samples, or ii) both the three cow samples (as a cow source) and three sheep samples (as a sheep source).

Secondly, simulated-samples were made using river water and three sheep faecal samples. SourceTracker was run five times with an FTL containing river water and either i) the three sheep samples used to make the simulated-samples, or ii) both the three sheep samples (as a sheep source) and three cow samples (as a cow source).

Finally, the simulated-samples made of either taxa from cow or sheep faeces with those from river water, were input into SourceTracker as a sink. SourceTracker was run five times with an FTL containing river water as a source and either i) the three contaminating cow or sheep samples as cow or sheep sources, respectively, or ii) all of the cow and sheep samples as a single ‘ruminant’ source. In SourceTracker, combining the sheep and cow communities into a single ruminant source within the FTL means that the OTU abundances from all sheep and cow sources were averaged together, before these source samples were used to make predictions.

Sewage samples

One sewage sample (Q, Table 5.1) was identified as having a bacterial community dissimilar to other sewage samples in experiment 1. To evaluate how best to incorporate this sample into the FTL, simulated-samples were made by either mixing taxa from sewage sample Q (Table 5.1) with those from river water, or using taxa from three sewage samples (A, B, and C, Table 5.1), with more similar bacterial communities, with taxa from river water. SourceTracker was run five times using an FTL containing either i) the three similar sewage samples (A, B, and C, Table 5.1) only, ii) the single dissimilar sample only (Q, Table 5.1), or

iii) both the three similar sewage samples (A, B, and C, Table 5.1) as a single source, and the single dissimilar sample (Q, Table 5.1) as a second, separate source.

5.2.7 Experiment 3 – Is it reasonable to use the range of faecal sources in the same FTL or is there cross-reactivity between sources?

Once the best approach to dealing with sets of samples with similar bacterial communities was determined, it was necessary to assess whether using cow, sheep, chicken, horse, pig, and sewage sources in an FTL was reasonable. To do this, separate simulated-samples were made for each faecal source; taxa from three samples from each faecal source were combined with those from river water at concentrations of 0.001%, 0.01%, 0.1%, 1%, 2%, 3%, and 4%, and at 5% intervals between 5 and 95%. Each simulated-sample contained only a single faecal source (made up of three samples) at each of the above concentrations. SourceTracker was run with a FTL which contained all available sources and samples, keeping cow and sheep sources separate. This allowed a comparison between the predicted values for each source to be evaluated, and the potential for cross-reactivity between sources, to be assessed.

5.3 Statistical analysis

The significance of the difference between SourceTracker predictions made on different samples was determined using a two sided, t-test. Where SourceTracker was run multiple times on identical samples with different configurations of FTL, a pairwise t-test was used. The alpha diversity of samples was calculated by removing OTUs not present in any sample, using the Phyloseq package in R (McMurdie & Holmes, 2013). Cohen's D was used as a measure of standardized effect size of the difference between the means for each t-test. Effect sizes are commonly described as small (Cohen's d = 0.2), medium (Cohen's d = 0.5), and large (Cohen's d = 0.8), very large (Cohen's d = 1.2) and huge (Cohen's d = 2.0) (Cohen, 1988; Sawilowsky, 2009). However, these are arbitrary values, and only used as an indicator of the effect size.

5.4 Results and discussion

5.4.1 Experiment 1

The effect of using a background sample as an additional source

Two studies have suggested that taking autochthonous taxa into account by including a background sample, such as river or sea water free of faecal contamination, in the FTL may improve the accuracy of SourceTracker (Brown *et al*. 2018; Hägglund *et al*. 2018). Here, simulated-samples, made using taxa from a single sewage sample and those from either a river or sea water sample, also suggest that including a background sample, where possible, improves the predictive accuracy of SourceTracker. Figure 5.1 shows just the simulated samples containing river water taxa, however, the same effect was observed in sea water (Appendix D.2). Including either a background sample (seawater or river water) as a source in the FTL resulted in significantly higher predictions ($p < 2.2 \times 10^{-16}$) than those made without a background source. The effect sizes were large with Cohen's D values of 1.38 and 1.36 for the inclusion of seawater and river water as background sources, respectively. The differences in SourceTracker predictions when using a background sample as a source in the FTL increased with the expected contribution (i.e., the proportions used to create the simulated-samples) of sewage increased, with a maximum difference of 6.5% occurring at an expected contribution of 30% for both seawater and river water (Figure 5.1 and Appendix D.2). At expected contributions of sewage greater than 30%, the difference between the predictions made with and without a background source decreased (Figure 5.1). In addition, if a source prediction cut-off of 1% contribution is used as a level with which to accept predicted values (Hägglund *et al*. 2018; Brown *et al*. 2017), using river as a background source in the FTL improved the sensitivity (taken here to be the expected value above the 1% predicted value used as a cut-off) of the technique. Thus, the lowest predicted contribution of 0.99% equated to an expected (true) sewage contribution from sewage of ~5% (a 4.01% difference between the true and predicted value) when a river source was not included in the FTL. This was improved to ~2% (expected contribution) when a background river source was used in the FTL (predicted contribution of 1.29%, a 0.79% difference with the true value), (Appendix D.2).

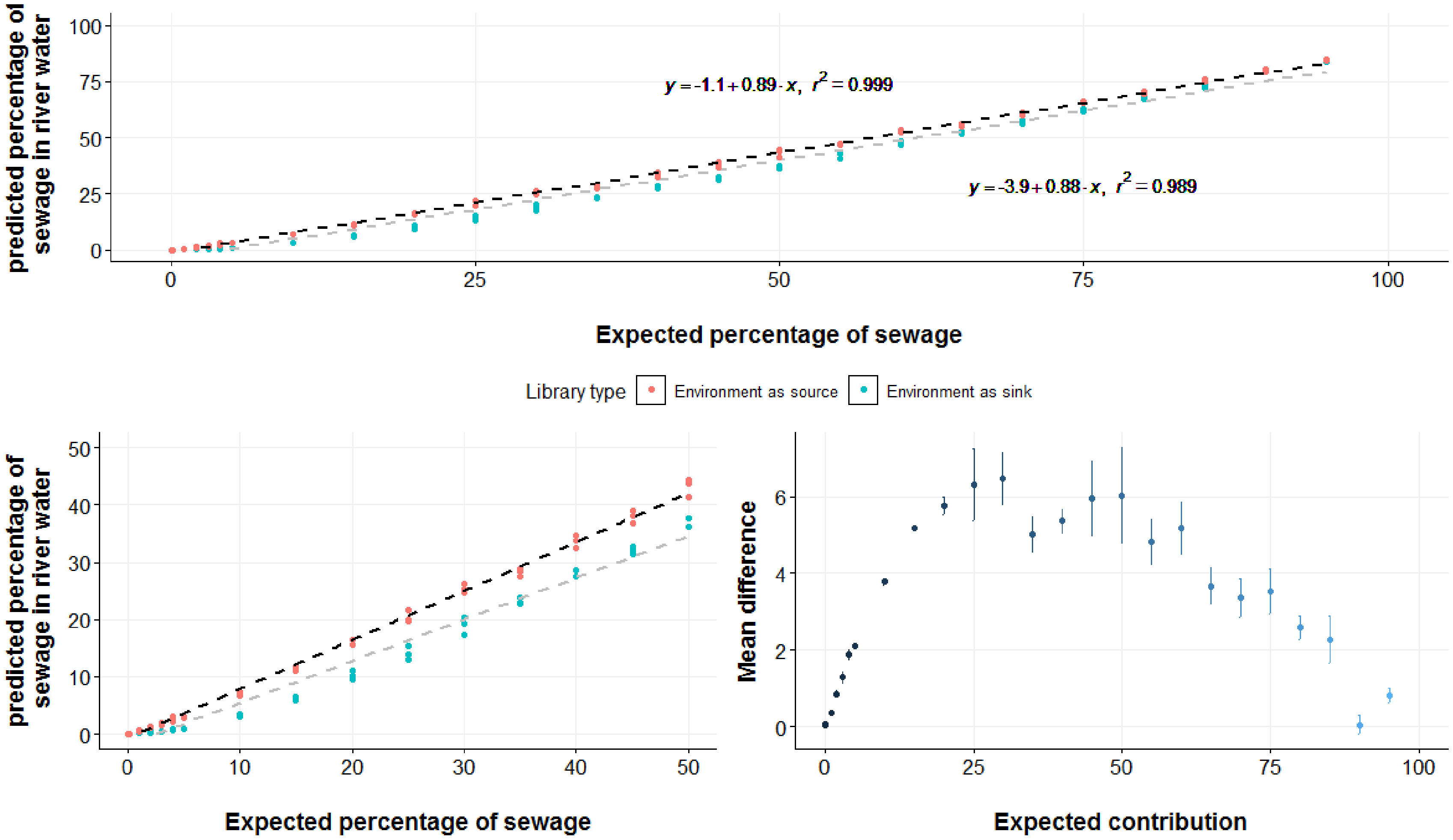


Figure 5.3. The effect of using an environmental water sample only as a sink or by including it also as a source on SourceTracker predictions when sewage is mixed with river water at proportions between 0 and 100 % (top). Proportions between 0 and 50% are shown (bottom-left) to highlight divergence of predictions at lower concentrations. The mean difference between SourceTracker predictions with river water as an additional source and as a sink only are shown (bottom-right)

This is similar to the detected range of sewage contamination reported in previous studies that has ranged between 1% and 7% (Figure B.1.3, Appendix B), and 1% and 10% (Hägglund *et al*., 2018). In addition, using a background sample to account for autochthonous taxa improved the linear relationship between the expected and predicted samples slightly (p= 0.999 compared to p = 0.989 without a background sample, Figure 5.3). The improved linearity of the predictions may explain why Hägglund *et al*., (2018) observed a better correlation between culturable *E.coli* and SourceTracker predictions after accounting for autochthonous taxa. The propensity of SourceTracker to assign reads to unknown sources, which has been previously observed (Henry *et al*., 2016), appears to be greater at larger contributions of contamination (slope of regression line = 0.989 Figure 5.3). Nevertheless, the linear relationship demonstrates that SourceTracker predictions are valid over a wide range of contamination levels from a particular source. Cases in which pristine samples cannot be obtained should be interpreted cautiously as the possibility of false negative results may be high (Unno *et al*., 2018). For all further analyses a background sample was used as a source in the FTL.

Can we build an FTL using sewage samples to adequately represent the North East of England?

The diversity of the faecal microbial communities within different chickens and dogs was less than those of cow, horse, sheep, pig, and sewage (Table 5.3).

Table 5.3. Mean Shannon diversity (± standard deviation) of faecal communities from different host environments.

Host	**Chicken**	**Cow**	**Horse**	**Dog**	**Pig**	**Sheep**	**Sewage**
Shannon Diversity	3.63 ± 0.99	6.08 ± 0.23	5.73 ± 0.36	3.24 ± 0.80	5.16 ± 0.23	5.60 ± 0.33	5.04 ± 0.18

Using the Jensen-Shannon distance as a metric for dissimilarity, samples from each faecal host source were more similar to each other, than to samples from other hosts (Figure 5.4). An adonis test showed that communities from different faecal host sources did not share the same centroid (p=0.001), suggesting that it may be possible to differentiate different host communities using SourceTracker. However, the beta dispersion (a measure of variance) is not homogenous at an alpha value of 0.01 (p = 0.043, Figure 5.4), suggesting that the results of the adonis test should be interpreted cautiously.

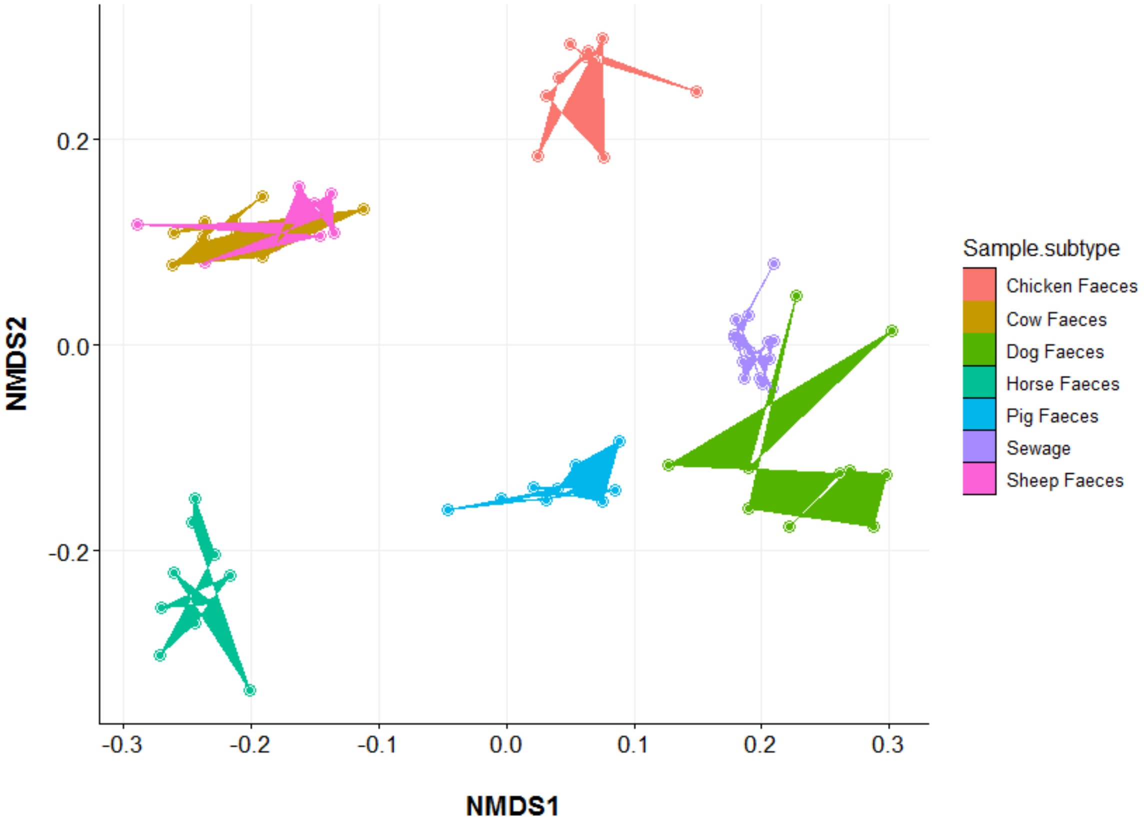


Figure 5.4. An NMDS plot visualizing the dissimilarity between bacterial communities in different faecal samples, determined using the Jensen-Shannon divergence.

Bacterial communities from the faeces of different hosts may be highly similar with a large number of shared taxa, particularly cow and sheep, and possibly sewage and dog (Figure 5.4). Interestingly, Hägglund *et al*. (2018) observed bacterial community similarities between sheep, cow and calf faeces, and also noted a large number of shared OTUs between sewage and dog bacterial communities. A single WWTP had a bacteria community that was dissimilar to all other WWTPs (Figure 5.4) and had the greatest distance to the centroid (Figure 5.5 (0.43)) compared to all other WWTPs.

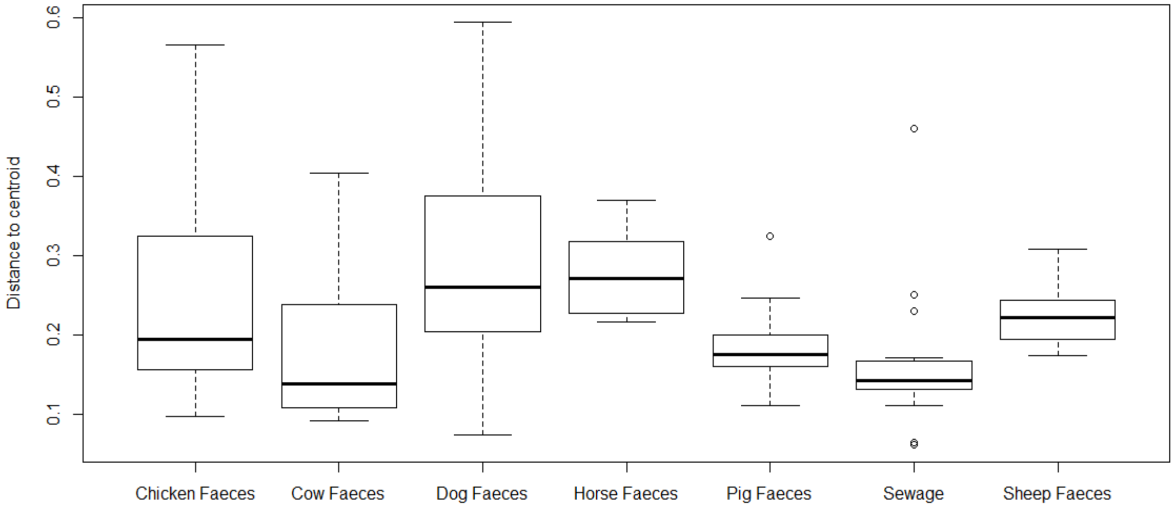


Figure 5.5. Boxplot showing the distance to centroid, a measure of the dispersion of bacterial communities between individuals within each host-type and sewage. Note the single outlying data point showing a large distance (0.43) to the centroid of all the sewage samples.

To explore whether the presence or absence of a sample in the FTL would affect SourceTracker predictions, simulated-samples were constructed using taxa from the WWTP sample (Q, Table 5.2) with the greatest distance from the centroid (Figure 5.5 (0.43)) and river water. SourceTracker was run with source FTLs, consisting of either: only the contaminating sample; all sewage samples (n = 15); or between 2 and 14 sewage samples, excluding the contaminating sample, as outlined in Figure 5.2.

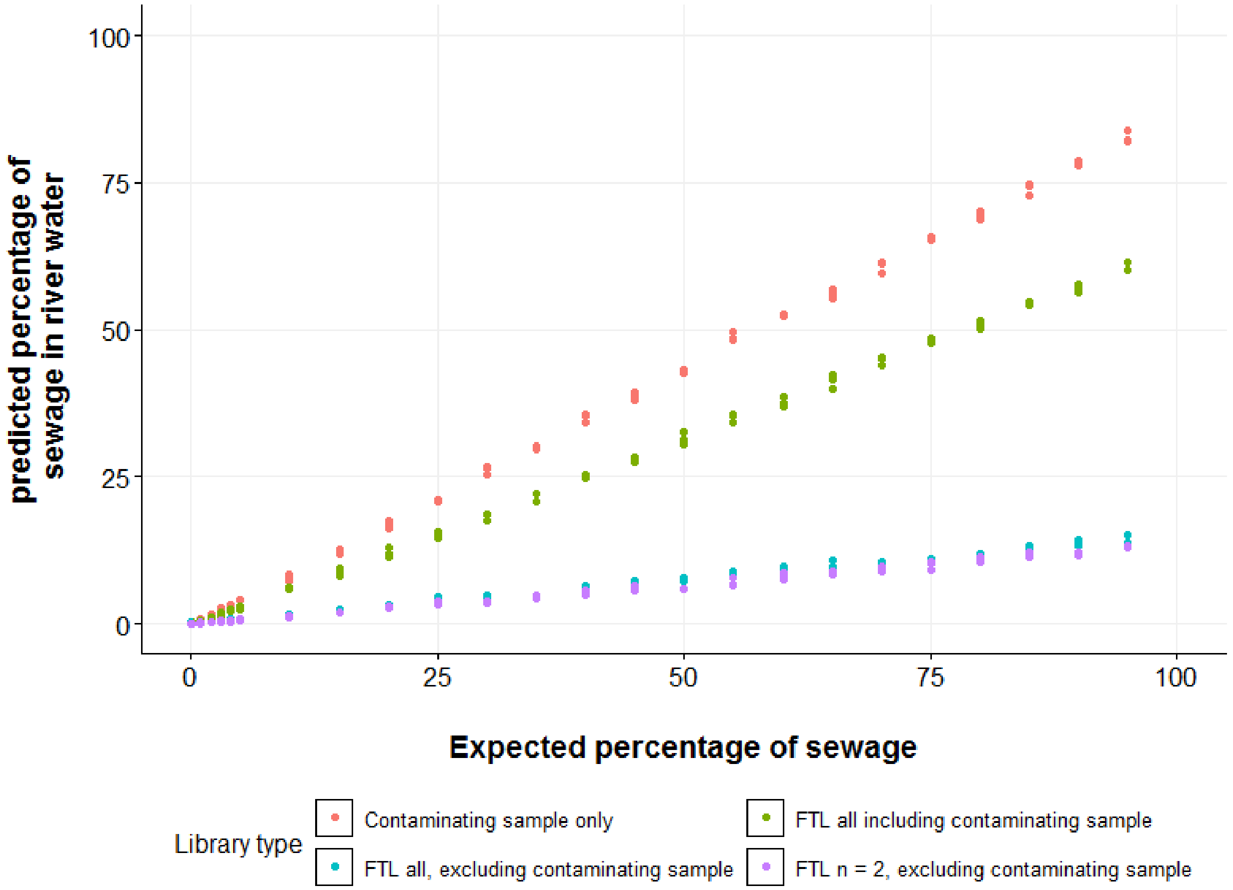


Figure 5.6. SourceTracker predictions from sewage and river water communities mixed at different proportions as a sink community, with SourceTracker runs with a FTL composed of either the single contaminating sample used to create source mixture, 14 sewage samples but excluding the contaminating sample, or all 15 sewage samples. River water was used as a contributing background source in all SourceTracker runs

Figure 5.6 shows that the predictions which were closest to those expected were achieved when the sewage FTL was comprised of only the contaminating sample (i.e., the sample used to create the simulated-samples). The addition of other sewage samples into the FTL led to a greater underestimation in the predicted proportions, compared to using only the single sample, (p= 2.57 x 10^{-16}, Cohen’s D = 1.15). Removing the contaminating sample (that from which the taxa was used to make the simulated-samples) from the FTL had the largest impact on source predictions (Figure 5.6). The size of the faecal library (i.e., the number of sewage samples) had little effect when the contaminating sample was not included in the library (Figure 5.6). Exclusion of the contaminating source (that from which the taxa was used to make the simulated-samples) from the FTL reduced the assay sensitivity from 2% expected sewage contamination (1.54% predicted with source only) to between 10% and 5% expected sewage contamination (1.70% predicted with all sources except the contaminating sample in the FTL) when using a reporting cut-off of 1% predicted contributions. Inclusion of all sewage samples in the FTL resulted in an underestimation of pollution; however, this underestimation increases with the expected proportion of sewage contamination and, therefore, only led to a small reduction in sensitivity when all sewage samples were included in the FTL (reduced from 2% to 3% expected sewage contamination (1.5% predicted contamination)).

This experiment was repeated using taxa from a sewage sample with a microbial community that was highly similar to those from other sewage samples in the FTL as the contaminating sample in the simulated sink communities to evaluate whether this phenomenon was due to the high degree of dissimilarity between the contaminating sewage sample and the FTL. Figure 5.7 shows that the effect observed in Figure 5.6 is still apparent, although to a lesser extent when the contaminating sewage sample (that from which taxa was used to make the simulated-samples) is similar to other sewage samples in the FTL. Using the single contaminating sample in the FTL is significantly better than using an FTL with all other sewage samples (excluding the sample used to make the simulated-samples) ($p = 5.681 \times 10^{-16}$, Cohen's D = 1.5) and when all sewage samples are included in the FTL ($p = 6.841 \times 10^{-16}$, Cohen's D = 1.14). However, when the contaminating sample is excluded, including a greater number of sewage samples in the FTL always improved the SourceTracker predictions compared to those with fewer sources (an FTL where n = 2 is shown in Figure 5.6 for clarity).

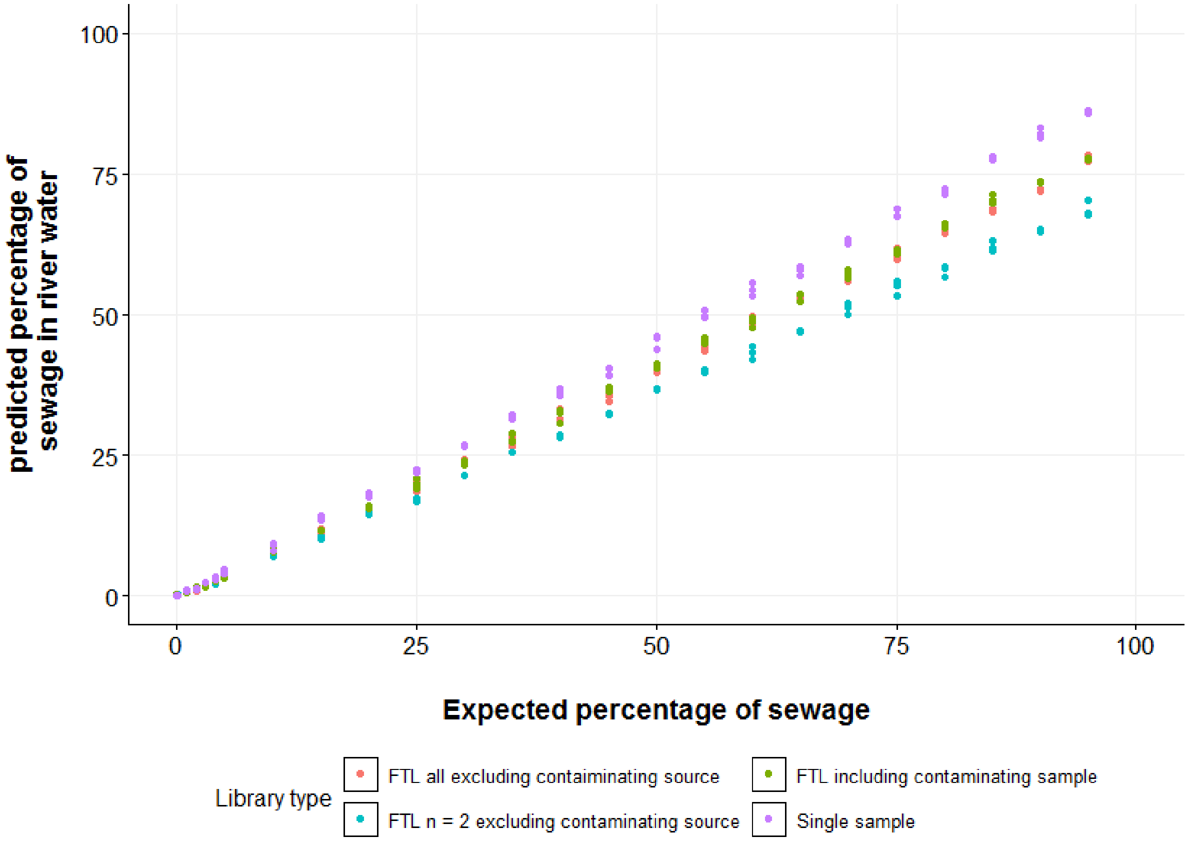


Figure 5.7. SourceTracker predictions from sewage and river water communities mixed at different proportions to create simulated-samples using a sewage sample with a similar microbial community to most other samples in the FTL. FTL libraries included with SourceTracker run with a FTL composed of either the single contaminating sample used to create simulate- samples, all sources (FTL including contaminating source), 14 sewage samples but excluding the contaminating sample (FTL all excluding contaminating source), or with fewer sewage samples – only the run containing 2 sewage samples (FTL n=2 excluding contaminating source) is shown for clarity.

Historically, the large library sizes required by library dependent microbial source tracking methods has been a limiting factor in their wide spread use. Figure 5.6 and 5.7 suggest that, where possible, using a sample from the likely contaminating source (such as a specific WWTP or farm) as a single sample in the FTL is preferable to using the entire source library. Under most circumstances, using an FTL with more samples appears to be advantageous (Figures 5.6 and 5.7), although, caution must be taken since this will lead to a slight underestimation of the contribution of faecal sources, particularly if the contaminating source community is highly dissimilar to other sources in the FTL. Brown *et al*., (2017) previously reported, following a power analyses, that more than 13 sewage samples were required to prevent false negative results. However, in practice this seems unlikely given the propensity for community analysis to produce false positive rather than false negative results, and success has been reported with fewer, or single samples (Henry *et al*., 2016; Iceton, 2018; Chapter 3). Nevertheless, the use of power analysis to inform library size requirements continue to be recommended and not used (Brown *et al*. 2018).

The inclusion of non-local samples in the FTL has previously been observed to have no impact on SourceTracker predictions using faecal samples mixed *in-vitro* (Staley *et al*., 2018). In contrast, the inclusion of sewage samples in the FTL which were not present in the sink sample led to significantly different SourceTracker predictions ($p = 1.459 \times 10^{-15}$); although, in practical terms this may have little impact on MST conclusions since the mean of the differences was less than 1% (0.661), and this would be lower still at lower levels of contamination (Figure 5.6). When the contaminating sewage sample is similar to the majority of sewage samples in the FTL, but is not necessarily in the FTL, it is unlikely to impact MST investigations (Figure 5.6). However, when the contaminating sewage sample is dissimilar to sewage samples in the FTL (having a greater distance to the centroid (Figure 5.6) then there is likely to be a large underestimation in SourceTracker predictions (Figure 5.7). Nevertheless, including non-local (potentially dissimilar) samples in the FTL is still recommended when the source of sewage is unknown (e.g., from a leaky sewer) since the exclusion of these samples is likely to reduce the sensitivity of their detection. It does, however, pose the question of how best to incorporate dissimilar samples from the same host-source into an FTL (see Experiment 2).

It appears in most cases, an FTL composed of multiple sewage samples is suitable for a particular region (in this case the North East of England). For MST studies, researchers should be aware that contamination of water bodies by sewage that has a dissimilar bacterial community to sewage samples in the FTL could lead to an underestimate of the amount of contamination, although, I found no evidence that this could lead to false negative results. One solution to prevent such underestimation would be to perform multiple SourceTracker runs with each single source sample to determine if the FTL significantly affects source predictions. In all trails conducted here, excluding the contaminating sample led to a greater reduction in predicted values, compared to adding more sources.

Repeating the analysis conducted here would be valuable for MST researchers who are developing an FTL for use across a region, particularly where small communities or decentralized WWTPs may contribute to the degradation of water quality. Moreover, repeating this analysis with other faecal sources, if these are important in a particular study, may be vital since other faecal sources such as dog and chicken samples had a greater variation in the distance to the centroid (Figure 5.5), suggesting that if dog and chicken samples are a concern, building a representative library may require greater care.

In all instances, SourceTracker was able to identify an expected sewage contamination of 1%. Below an expected contribution of 1% the RSD increased above 100% suggesting a low confidence in SourceTracker predictions (Henry *et al*., 2016), although, SourceTracker consistently identified an expected contribution of 0.1%. At 0.01%, SourceTracker reported no contribution in at least one out of three samples, suggesting that the RSD is a suitable, and potentially conservative, metric to prevent false positive results (Henry *et al*., 2016). Two recent papers have also used a 1% predicted contribution as a cut-off for reporting SourceTracker results, in addition to an RSD >100% (Brown *et al*. 2018; Hägglund *et al*. 2018). Again this is reasonable, if not slightly conservative, as an expected sewage contribution of ~2% resulted in a predicted sewage contribution of ~1%, while an expected sewage contribution of 1% yielded SourceTracker predictions between 0.2 and 0.7%. The ability to identify lower contributions does, however, depend on the level of cross-reactivity between the different sources included in the FTL (Experiment 2 and Experiment 3).

5.5 Experiment 2 – How to incorporate sources with similar bacterial communities into an FTL?

It has been suggested that host sources with similar bacterial communities could cause SourceTracker to report false positive results. Here, samples from different hosts which have been observed to be similar (Figure 5.4) were used to create simulated-samples. SourceTracker was run using these simulated sink samples and an FTL in which different host sources were either separated by source (e.g., cow and sheep), or combined into a single source (e.g., ruminants).

5.5.1 Cow and sheep sources

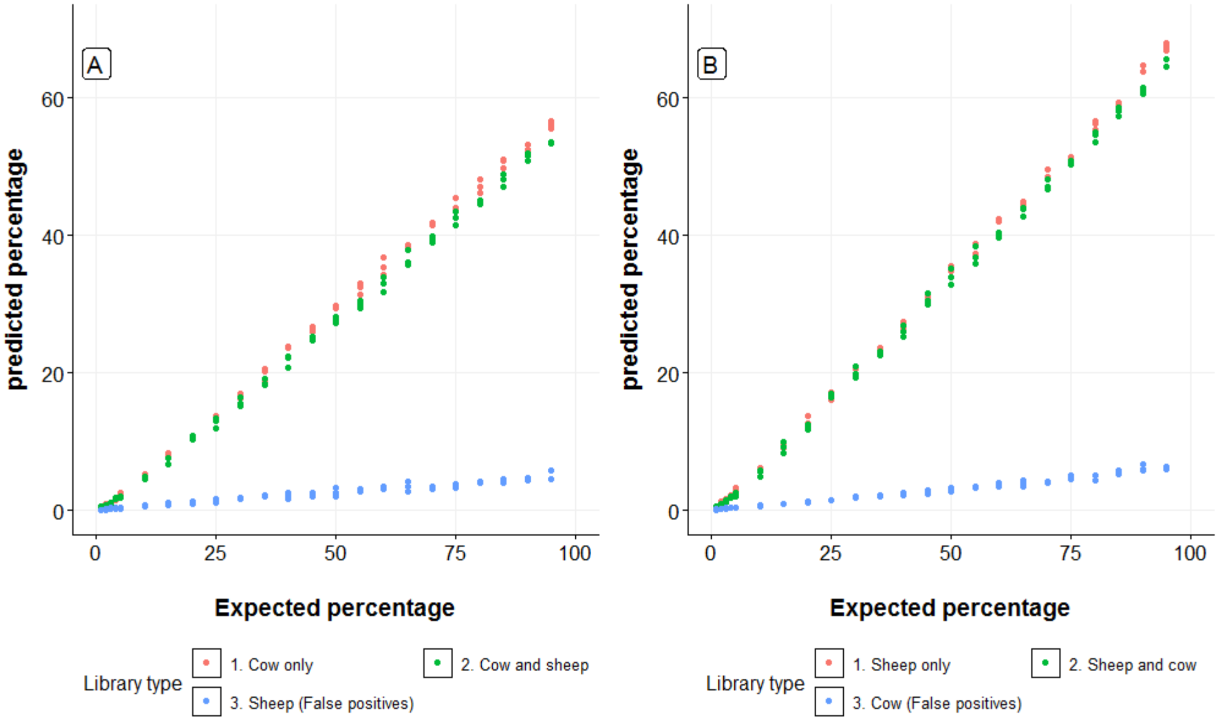


Figure 5.8. Results of SourceTracker with a simulated sink community contaminated with either cow only (A), or sheep only (B) in river water using either just the known source or both cow and sheep sources in the faecal taxon library.

Figure 5.8 shows that the presence of false positive signals at all expected contributions in river water when taxa from either cow or sheep faecal samples are not present in the sink samples, but both are present as separate sources in the FLT. However, false positive results only exceeded 1% when the contribution of sheep faeces was greater than 15%. While a large effect size (Sheep - CohensD = 0.848, p = 5.985x10^{-10}, Fig. 5.8b) was observed when using the whole dataset, when a subset of the dataset below a predicted contribution of 30% the effect size reduced (Sheep - CohensD = 0.461, P = 0.0086, Fig. 5.8 B).

Combining samples from sources with similar bacterial communities into a single source in the FTL has been suggested as a way to overcome problems arising from sources with similar bacterial communities (Staley *et al*., 2018). Figure 5.9 shows the effect of having cow and sheep faecal samples as either separate or combined (as ruminants) in the FTL. Combining sheep and cow faecal samples into a ruminant source in the FTL leads to a consistent underestimation of the contribution of cow (p = 2.029x10^{-12}, Cohen's D = 0.93, Figure 5.9) and sheep faeces (p = 1.347x10^{-15}, Cohen's D = 1.12, Figure 5.8), compared to having separate sources in the FTL. This was expected given the effect of combining dissimilar samples into an FTL (Experiment 1), although, this effect is larger for sheep faeces compared to cow (figure 5.9). Importantly, combining sources had no effect on the sensitivity of the assay when using a cut-off of 1% predicted contaminant contribution.

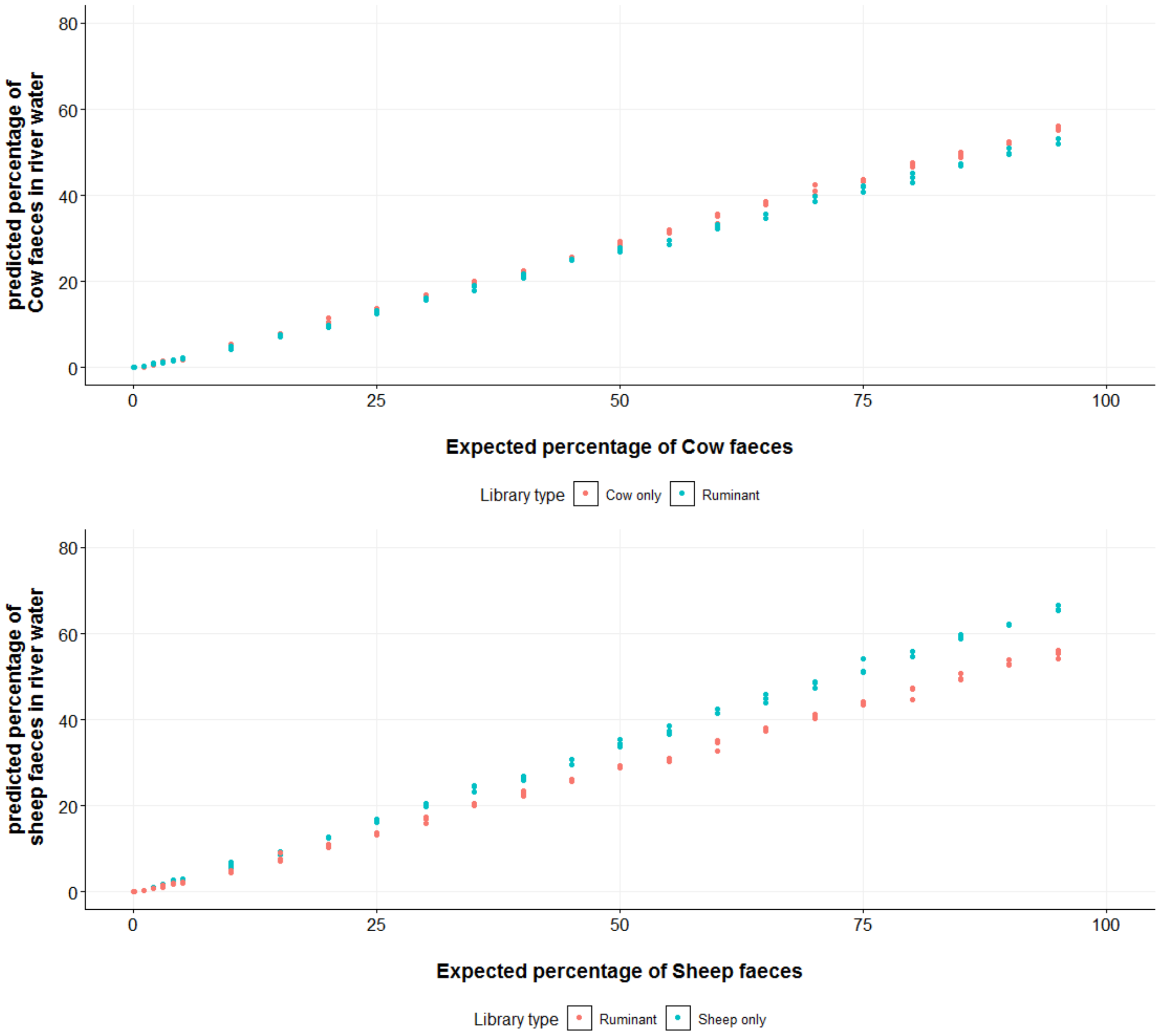


Figure 5.9. The effect on the detection of cow (top) and sheep (bottom) faeces in river water when the faecal taxon library contains both cow and sheep sources, or the cow and sheep samples are combined into a ruminant source.

Combining sources with similar bacterial communities, such as ruminants, may make the predictions between these sources more comparable. Predictions for the proportion of cow faeces in river water made by SourceTracker were consistently lower than those for sheep in river water for the same expected concentrations (Figures 5.8 and 5.9). This, and the lower predicted values of sheep and cow, compared to that of sewage, is discussed more below (Experiment 3); however, it is noteworthy that where contamination from both cow and sheep sources is expected, combining these sources in the FTL may make conclusions more useful, although sheep pollution may be significantly underestimated, particularly compared to other sources (see experiment 3).

Non-local sewage samples

To assess the best way to incorporate a single sewage sample with a bacterial community which is dissimilar to that of other sewage samples (Figure 5.5) into an FTL, a similar procedure as above (cow and sheep) was followed. Two sets of simulated communities were made (Table 5.4). For the first set (numbers 1 – 4, Table 5.4) simulated-samples were made by mixing taxa from sewage sample Q (dissimilar bacterial community) with those from river water at proportions between 0.001% -100% Q, with the addition of either 0%, 1%, 10%, or 20% taxa from other sewage samples with similar bacterial communities (A, B, and C, Table 5.4). For the second set (numbers 5 – 8, Table 5.4) simulated-samples were made by mixing taxa from sewage samples A, B & C (similar bacterial communities) with those from river water at proportions between 0.001% -100% sewage, with the addition of either 0%, 1%, 10%, or 20% taxa from sewage sample Q (Table 5.4). SourceTracker was run on all sets of simulated-samples with an FTL with sample Q separated from the other sources (i.e., the FTL contained a “sewage” source (A, B, and C) and a “Q” source). For clarity, only samples containing 0% and 20% of additional sources (numbers 1, 4, 5, and 8, Table 5.4) are shown in Figure 5.10.

Table 5.4 The mixtures of samples (percentages given in brackets) used to construct simulated-samples for experiment 2.

Simulated - sample number	Samples (Proportion contained in simulated-sample)		
1	River (0% - 99.999%)	Q (100% - 0.001%)	A, B, & C (0%)
2	River (0% - 98.999%)	Q (99% - 0.001%)	A, B, & C (1%)
3	River (0% - 94.999%)	Q (95% - 0.001%)	A, B, & C (10%)
4	River (0% - 79.999%)	Q (80% - 0.001%)	A, B, & C (20%)
5	River (0% - 99.999%)	Q (0%)	A, B, & C (100% - 0.001%)
6	River (0% - 98.999%)	Q (1%)	A, B, & C (99% - 0.001%)
7	River (0% - 94.999%)	Q (10%)	A, B, & C (95% - 0.001%)
8	River (0% - 79.999%)	Q (20%)	A, B, & C (20%)

The addition of up to 20% sewage with similar bacterial communities in the simulated-samples had no significant effect on the predicted value of Q (p = 0.1839) when SourceTracker was used to identify Q in river water (Figure 5.10, left). However, when sewage was the contaminating source and 20% of Q was added in the simulated-sample, there appears to be some conflation of taxa from these sources (Figure 5.10). In addition, the expected and predicted contributions of sewage were significantly different (p = 0.03539) at expected sewage concentrations above 25%. Exclusion of sources from the FTL that have dissimilar communities to others (e.g., Q), or their inclusion with sewage sources as a single source, could lead to a severe underestimation of any sewage sources that may be similar to the dissimilar source (Q) which may be present in a given environmental sink. A recommended approach then could be to separate Q and other sewage sources in the FTL and add these together to give a “human” source following analysis by SourceTracker. This seems sensible since sewage contributions above 25% are generally not expected, except for in highly polluted waters. This reflects and validates the approach taken in Chapter 3 in dealing with sewage and septic tank samples.

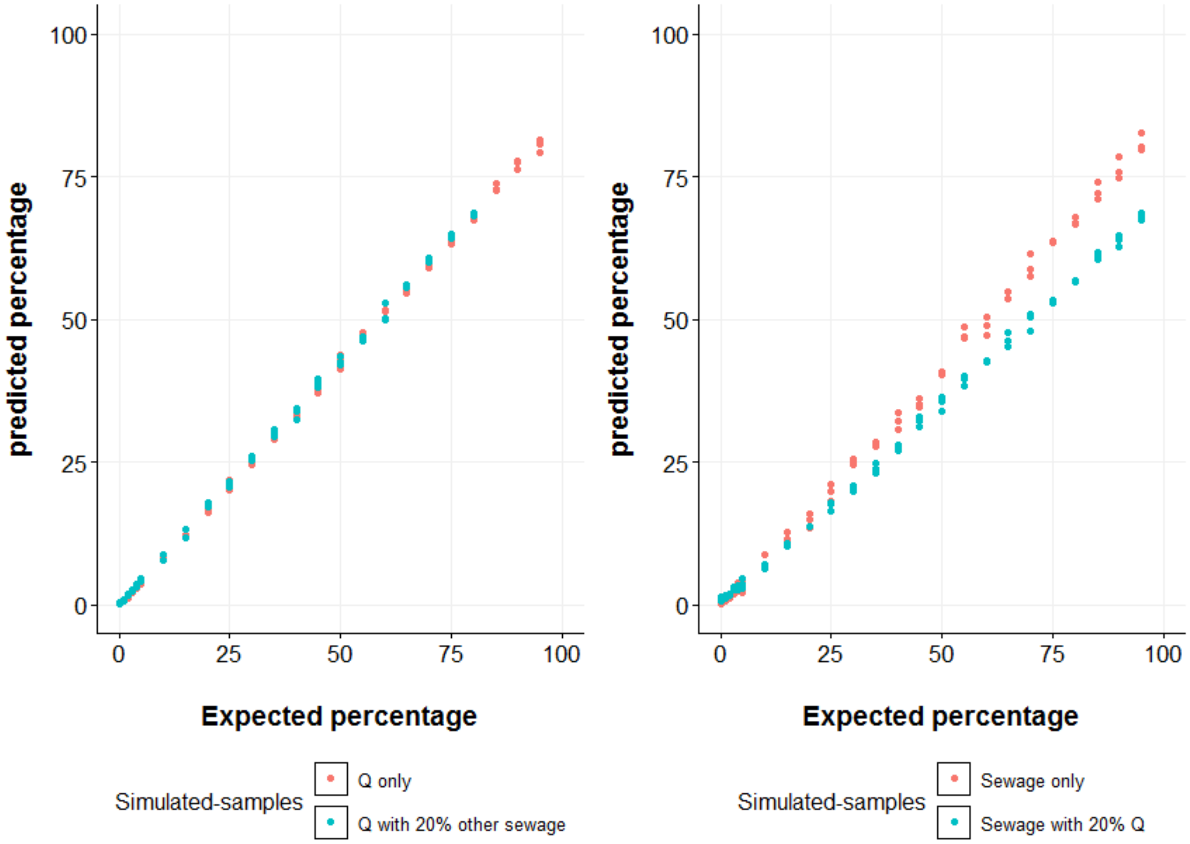


*Figure 5.10. **Left** – The detection of a sewage source Q with a bacterial community dissimilar to other sewage sources. "Q only" represents simulated-samples containing taxa from the Q and river water samples (simulated-sample number 1, Table 5.4). "Q with 20% other sewage" represents simulated-sample number 4 (Table 5.4). **Right** - The detection of a sewage source that has a bacterial community similar to other sewage sources. "Sewage only" represents simulated-samples made from taxa from the A, B, C, and river water sample (simulated-sample number 5, Table 5.4). "Sewage with 20% Q" represents simulated-sample number 8 (Table 5.4). The FTL used in all instances contained river water as a background source, sewage (Samples A, B and C), and Sample Q as separate sources.*

5.6 Experiment 3

Sets of simulated-samples were created, representing contamination from each single host source (e.g., cow), using taxa from individual sources with those from river water. These simulated samples were input into SourceTracker with the entire FTL (i.e., containing all available samples of cow, sheep, dog, horse, pig, and sewage sources). Cow and sheep were kept as separate sources in the FTL, since only one source was in each set of simulated-samples. Figure 5.11 shows that the predicted values for a given expected value were different depending on the source. The cow and horse sources showed the greatest underestimation. This is due to the larger proportion of taxa that were assigned to the ‘unknown’ source for cow and horse faecal sources, compared to other sources. The horse faecal bacterial community did not appear to be highly similar to bacterial communities from other sources (Figure 5.5), and no cross-reactivity was observed between horse and sheep or cow host sources (Table 5.4). It is difficult to determine a cause for this underestimation, although, it could be related to the mean diversity observed in the samples (Table 5.3), with the greatest underestimation observed in samples which have the largest diversity. Overcoming the disparity between predictions related to sources may be difficult for MST researchers, however, gaining an understanding of this disparity by conducting similar investigations is recommended and will help to inform conclusions drawn from MST studies.

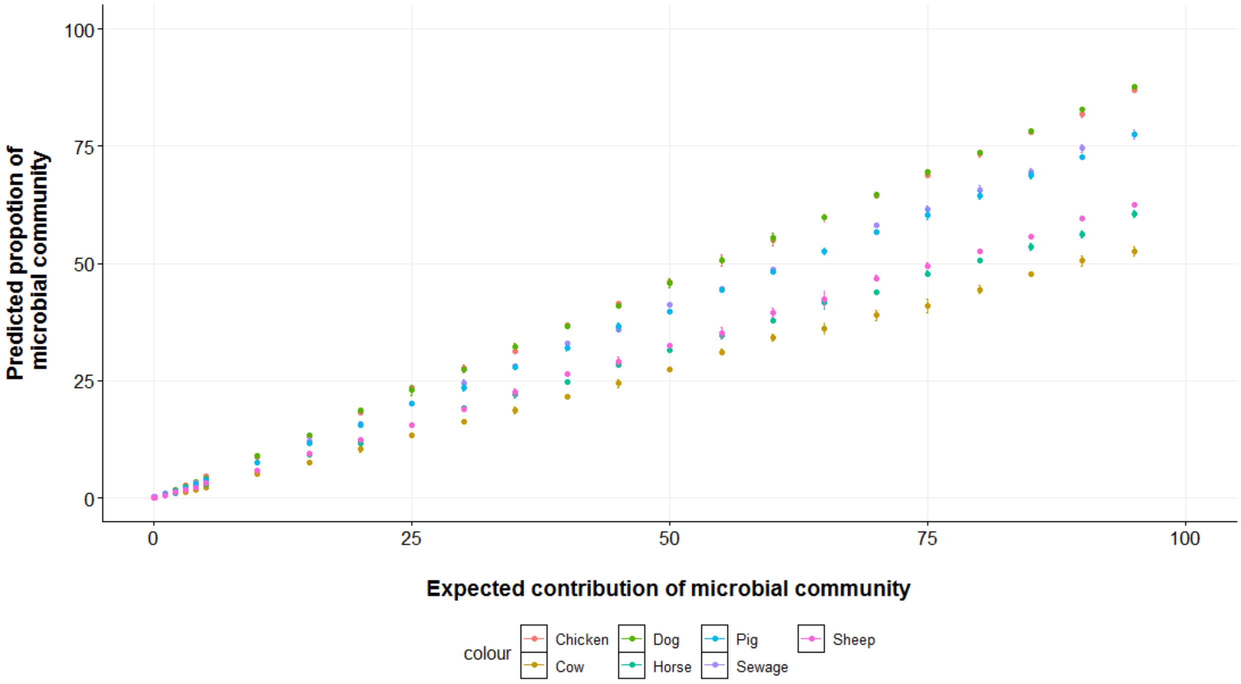


Figure 5.11. Comparison of multiple source tracker runs to predict individual faecal source contributions to simulated sinks of each source type in river water using a FTL containing all sources. Error bars represent the standard deviation of 3 sets of simulated sink microbial communities at each expected contribution, although standard deviations are very small compared to the means.

.

These underestimates only reduced the sensitivity of cow assays slightly compared to other faecal sources (3%, Table 5.4). The sensitivity of sheep and horse sources was not affected. Cross-reactivity between dog and sewage, and sheep and cow sources was observed, although, at a slightly lower level of contamination (10%, Table 5.4), than reported above. This could be a result of the variation in random sampling of microbial communities, or an effect of using an FTL containing a larger number of sources. An important note is that only using an RSD of 100% to identify false positive results would result in false positives being observed at lower concentrations. A 1% cut-off, therefore, is important to prevent false positive results.

The FTL comprised of all sources seems suitable for use in the UK, particularly with an awareness of the potential for false positives. However, caution is required when comparing chicken, pig, or sewage predictions with those from cow, sheep or horse, particularly at higher concentrations.

Table 5.5. The sensitivity and observed cross-reactivity when using the entire FTL.

	Chicken	Cow	Dog	Horse	Pig	Sewage	Sheep
Sensitivity*	2%	3%	2%	2%	2%	2%	2%
Cross-reactivity**	None	Sheep > 10%	None	None	None	Dog > 25%	Sheep > 10%

*The percentage of expected contributions where all three simulated microbial communities had a predicted percentage contribution >1%.

**Defined as at least 1 out of 3 samples containing unexpected sources at concentrations > 1%.

5.6.1 Summary

This study supports the use of community-based MST using the SourceTracker software (Knights *et al.*, 2011) for MST investigations. Community-based MST was able to consistently identify expected sources and differentiate these from other sources.

Building the FTL has been highlighted as one of the most important factors, which, can affect SourceTracker predictions and therefore MST investigations. Here, simulated-samples were used to inform the development of an FTL. A key finding of this study was that a change in the approach of MST researchers is necessary when using community-based MST.

Library size is not the best predictor of accuracy, the similarity of the samples in the FTL to those contributing to the contamination of environmental samples is more important.

Here, the FTL composed of 14 sewage samples from across North East England appears to be suitable for the source tracking of most sewage sources. However, caution is required. Separating one sewage source which was dissimilar to other sewage sources (Figure 5.3) in the FTL was recommended. By creating simulated-samples of sink microbial communities, differences in the predicted contributions of human sources to the microbial community when the FTL consisted of the dissimilar sewage source grouped with all other WWTP samples, and when this source was separate from other sewage sources (Figure 5.9). Previous suggestions to combine similar sources should be approached with caution; combining samples from different host-sources may increase the diversity of these sources, and lead to greater underestimation in SourceTracker predictions. Here, combining cow and sheep sources led to the underestimation of sheep contribution, although, combining cow and sheep faecal sources led to more comparable SourceTracker predictions. When sheep and cow faecal sources were used in the FTL alone, the predicted values for cow faeces were less than that for sheep, for the same expected values.

5.7 Conclusions and recommendations

The FTL has the potential to contribute significant bias in community-based MST. Researchers using community-based MST should focus their efforts on developing local, source targeted FTLs, rather than concentrate efforts on increasing the size of the FTL. The use of the script provided (Appendix D.1) allows researchers to evaluate the composition of their FTL either prospectively, or retrospectively to support conclusions drawn from community-based MST investigations. When building an FTL, care is required when deciding whether to combine sources with similar bacterial communities since the effects of separating or combining sources vary depending on the faecal sources. The potential for cross-reactivity was noted between sheep and cattle faeces, combining these sources in the FTL could allow for a better comparison of these sources, although, may reduce the sensitivity of detection for sheep faeces. Previous reports of conflation of sewage and dog samples (Hägglund *et al.*, 2018) does not appear to be a major issue in the FTL used in this study, although some cross-reactivity was observed when high concentrations of dog faeces was present. Using a cut-off of 1% was useful in reducing the false positive results, however, some false positives still remained at higher expected contributions (>10% for cow and sheep). It is recommended, therefore, that future studies repeat this, or similar analyses with their own datasets and FTL to inform future studies, since the predictions made by SourceTracker will depend greatly on the background samples (if provided), and the composition of the FTL.

An important and unexpected observation here was that the SourceTracker predicted values of cow and horse faecal sources were less than those from other sources, using the FTL built here. The difference between the expected and predicted proportions of the microbial community appear to differ with different faecal sources. This could be an issue for MST researchers when attempting to compare the relative contribution of faecal sources, although, the differences are larger at higher contributions (Figure 5.10). Further studies into why the predictions for some sources are under estimated would be worthwhile for MST studies. In addition, comparing FTLs for sources to cover wider geographical regions would be useful, particularly if community-based MST is going to gain wide spread acceptance in the water industry.

Chapter 6 Seaton Sluice case study

6.1 Introduction

Despite improvements in water quality over the past 30 years, the quality of environmental waters across Europe remains a concern (EEA, 2018b). Our ability to cost-effectively improve environmental water quality are limited by difficulties in identifying pressures on a catchment and applying a suitable programme of measures to monitor and reduce these pressures (Voulvoulis *et al*., 2017), which are primarily due to diffuse pollution sources (EEA, 2018a). Microbial source tracking (MST) has the potential to identify and, to some extent, apportion sources of faecal contamination. While numerous methods exists, there are few examples of 'real-life' MST investigations (Harwood *et al.*, 2014), particularly on a catchment scale. For example, community-based MST and *E.coli* biomarkers have only been used in four (Brown *et al*., 2017; Hägglund *et al*., 2018; McCarthy *et al*., 2017, Chapter 3), and three catchment investigations (Gomi *et al*., 2014; Katarżytė *et al*., 2018; Chapter 3), respectively. Those studies have typically been conducted in the USA or Australia where MST has been more widely adopted compared to the UK. Conducting catchment-wide MST investigations is, therefore, not only useful on a local scale for informing investment and management decisions, it is useful on national and international scales to inform future studies into the use of these MST techniques.

6.1.1 Selection of the Seaton Sluice catchment

The Bathing Water (BWD, 2006/7/EC) and Water Framework (WFD, 2000/60/EC) Directives are two key drivers for the improvement of environmental water quality. Under the BWD, all bathing waters must meet the "Sufficient" standard, and Northumbrian Water aims for all bathing waters to be classified by at least "Good" by 2020, and "Excellent" by 2029. the Seaton Sluice bathing water was classified as a "Good", although given a 93.60% chance of not achieving the "Excellent" classification in future bathing water seasons following an assessment by the Environment Agency (EA) to establish the likelihood of bathing waters achieving each BWD classification (Table 1.1; (Pinner, 2014)). An initial desk-study, undertaken in 2013 (Pinner, 2014), observed that regulatory bathing water samples that failed to achieve “Good” status, coincided with permitted discharges from combined sewer overflows (CSOs) some of the time. However, several failures also occurred when discharges from CSOs were not occurring. While development of the sewer infrastructure assets could improve the bathing water quality, there is a historic lack of support from customers to spend money on bathing water quality improvements (Pinner, 2014). A different approach to improving the robustness of the bathing water classification at Seaton Sluice, and indeed, all bathing waters is required. Pinner’s (2014) initial investigation recommended further research to identify the sources of pollution entering a bathing water site); this presented an opportunity to undertake a novel case study, to identify opportunities to improve water quality and improve the robustness of the ‘Excellent’ classification. Improving waters of good quality is likely to become increasingly important as water quality slowly improves and the ‘easy wins’ of problematic infrastructure assets are mitigated. It also provides an opportunity to test the potential limitations of MST methods, since a high sensitivity may be required to identify lower levels of pollution in environmental waters.

Investigation of the Seaton Sluice catchment also gave an opportunity to use MST to investigate a number of the pressures leading to low classifications under the WFD (WFD, 2000/60/EC). Several waterbodies in the Seaton Sluice catchment fail to achieve “Good” status, although the EA aim to improve them to “Good” status by 2027 (Environment Agency, 2018a). The EA highlights reasons why water bodies in the Seaton Sluice catchment fail to achieve “Good” status (Table 6.1); among these reasons, pollution from wastewater is presumed to impact multiple elements of the WFD classification. Urban diffuse pollution is emerging as a serious obstacle to water bodies achieving good status since it is often difficult to detect and difficult to differentiate from rural diffuse pollution sources. Many pressures on waterbodies therefore remain “suspected” (Table 6.1), which limits the remediation of these sources, although the RBMP for the Northumbria region suggests that mitigation of urban diffuse pollution could lead to improvements in water quality in the Seaton Sluice catchment, and three other catchments in the district (Environment Agency, 2016). Moreover, selection of the Seaton Sluice catchment gave an excellent opportunity to investigate the pressures leading to WFD failures in the catchment.

Table 6.1. Classifications and reasons for not achieving good status of water bodies in the Seaton Sluice catchment. Data from the EA data explorer (Environment Agency, 2018a).

Waterbody	Type	Classification status	Classification element	Category	Certainty	Activity	Issue
Seaton Burn	River	Moderate	Invertebrates	Water Industry	Probable	Sewage discharge (intermittent)	Pollution from wastewater
		Moderate or less	Mitigation Measures Assessment	Urban and transport	Confirmed	Not in the list	Physical modifications
		Moderate	Invertebrates	Urban and transport	Probable	Drainage - mixed	Pollution from towns, cities and transport
		Moderate	Fish	Sector under investigation	Not applicable	Unknown	Unknown
Big Waters Reservoir	Lake	Poor	Total Phosphorus	Agriculture and rural land management	Suspected	Mixed agricultural	Pollution from rural areas
		Moderate	Phytoplankton	Domestic General Public	Suspected	Un-sewered domestic sewage	Pollution from wastewater
		Moderate or lower	Mitigation Measures Assessment	Recreation	Confirmed	Other	Physical modifications
		Poor	Total Phosphorus	Domestic General Public	Suspected	Un-sewered domestic sewage	Pollution from wastewater
		Moderate	Phytoplankton	Sector under investigation	NA	Unknown	Unknown
		Poor	Total Phosphorus	Sector under investigation	NA	Unknown	Unknown
		Moderate	Phytoplankton	Agriculture and rural land management	Suspected	Mixed agricultural	Pollution from rural areas

NA – Not applicable

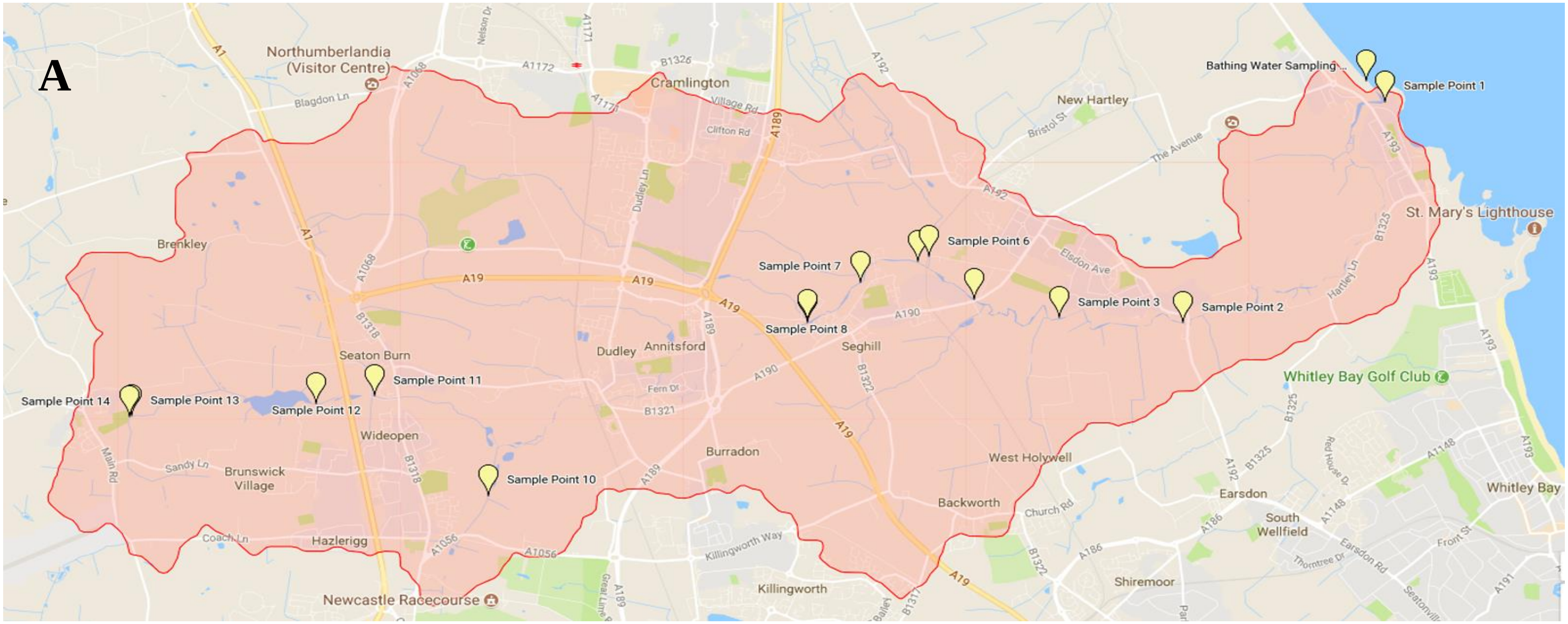
6.1.2 Catchment background

The Seaton Sluice catchment is located on the coast of North East of England, bordering two counties: Northumberland and Tyne and Wear. The catchment covers an area of 51 km^2 (Figure 6.1) and is primarily rural with over 65% of land used for arable (53%) and livestock farming (15%) (Table 6.2).

Table 6.2. Summary of land use in the Seaton Sluice catchment (2015).

Land use	**Total area (m^2)**	**Percentage land use**
Arable and horticulture	27229947	52.97
Suburban	9653288	18.78
Improved grassland	7631841	14.85
Broadleaf woodland	4016425	7.81
Urban	1269445	2.47
Inland rock	563785	1.10
Coniferous woodland	414442	0.81
Freshwater	221508	0.43
Neutral grassland	204987	0.40
Acid grassland	82171	0.16
Supralittoral sediment	47829	0.09
Saltmarsh	38026	0.07
Littoral sediment	24610	0.05
Littoral rock	5218	0.01
Total	51403522	100

Suburban and urban land makes up over 20% of land use in the catchment. The catchment drains the Seaton Valley, which contains roughly five areas of urbanised land: Seaton Sluice, Seaton Delaval, Holywell, Seghill and Dinnington (Figure 6.1). While Dinnington is outside the catchment drainage area, housing developments are currently ongoing there, and the land surface drains and sewers drain to, and run through the catchment, respectively. A combined sewer system drains the majority of the catchment (~75%), while more recently developed areas have a separate system (Figure 6.1). The land surface drains discharge into a watercourse which changes name throughout the length of the catchment; however, the watercourse is known locally in its entirety as Seaton Burn. Figure 6.1 shows the extent of the catchment which feeds into the Seaton Sluice bathing water.


A
Northumberlandia (Visitor Centre)
Cramlington
New Hartley
Bathing Water Sampling ...
Sample Point 1
St. Mary's Lighthouse
Brenkley
Sample Point 6
Sample Point 7
Sample Point 3
Sample Point 2
Sample Point 8
Seghill
Dudley
Annitsford
Seaton Burn
Sample Point 11
Sample Point 14
Sample Point 13
Sample Point 12
Whitley Bay Golf Club
Wideopen
Burradon
West Holywell
Brunswick Village
Sample Point 10
Earsdon
Backworth
South Wellfield
Whitley Bay
Hazlerigg
Killingworth
Shiremoor
Newcastle Racecourse
A1
A19
A189
A190
A192
A193
A1068
A1056
A1171
A1172
B1318
B1321
B1322
B1326
B1325
A186
A1148
A191
Blagdon Ln
Clifton Rd
Village Rd
Dudley Ln
Elsdon Ave
The Avenue
Bristol St
Hartley Ln
Fern Dr
Sandy Ln
Coach Ln
Main Rd
Killingworth Way
Church Rd
Great Lime Rd
Red House Dr
Earsdon Rd
Thorntree Dr
Front St

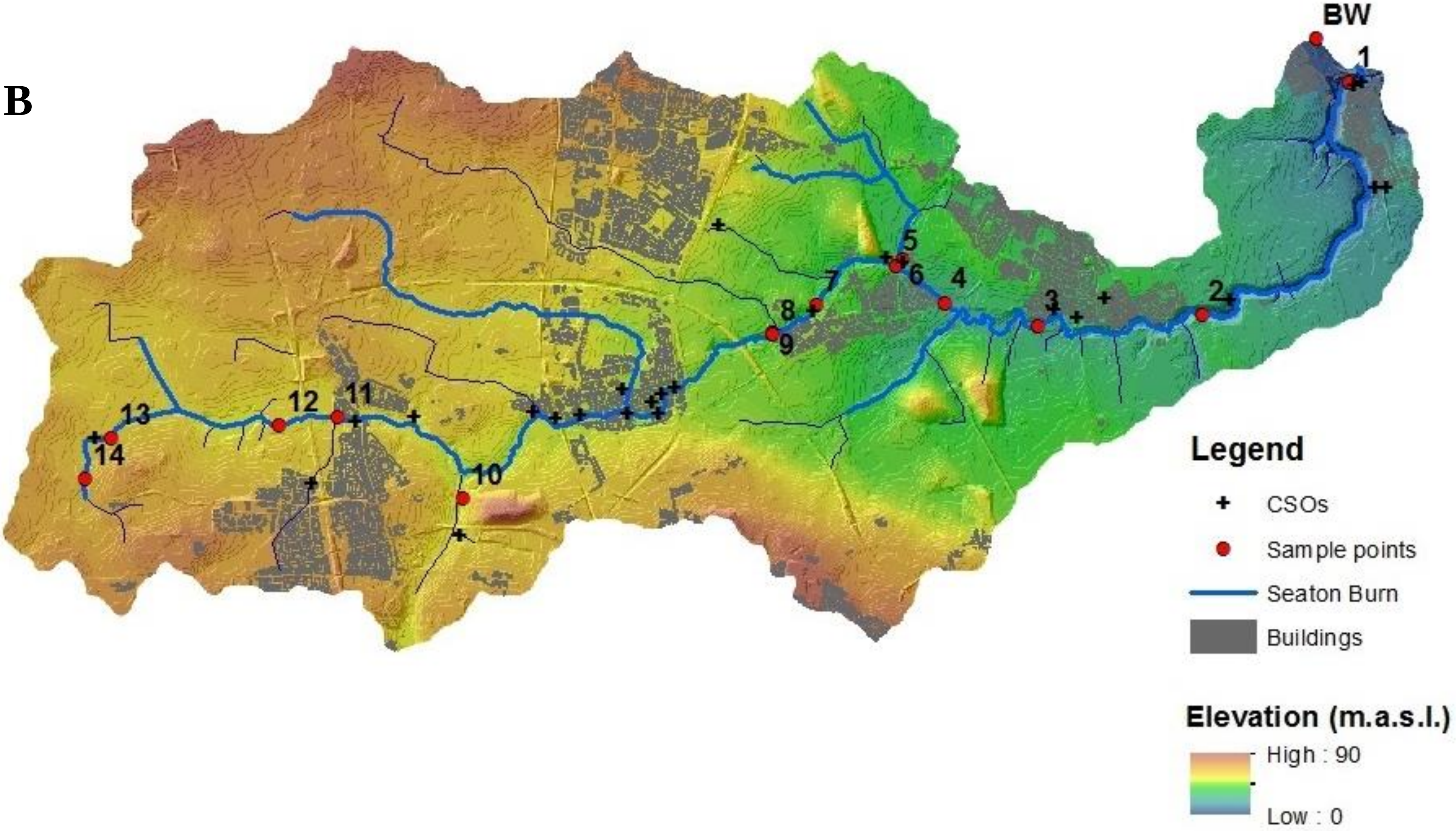


Figure 6.1. Open street map (A) and digital elevation map (B) of the Seaton Sluice catchment showing sampling points and combined sewer overflows (CSOs) (B only) in the areas.

There are a variety of potential sources of pollution throughout the Seaton Sluice catchment. There are 26 CSOs, which may impact water quality in the Seaton Burn and the bathing water quality at Seaton Sluice, 19 CSOs discharge directly into the Seaton Burn, and eight into tributaries entering the Seaton Burn (Figure 6.2). All CSOs are equipped with monitoring equipment to warn Northumbrian Water when a CSO is discharging and the duration of each discharge > 15 minutes (an example of this data is available in Appendix E.5). There are also two consented, trade effluent discharges, which arise from landfill sites in the Seghill area. These are consented as intermittent discharges and should only discharge at periods of heavy rainfall. Since bathing water failures also occurred when CSOs were not discharging, the role of diffuse pollution cannot be discounted. Sheep and cattle graze at the top and towards the bottom of the catchment, respectively and there are five stables within the catchment and others that are close to the catchment that exercise their horses throughout the catchment and along Seaton Sluice beach throughout the year. In addition, there is a public bridleway/path that runs almost the length of the Seaton Burn, which attracts horse riders and dog walkers, and a nature reserve containing some migratory birds, rabbits and deer, which could also contribute to sources of pollution. In addition, there are 38 surface water outfalls, which carry surface water directly into the Seaton Burn or its tributaries (Figure 6.1). There is, therefore, the potential for misconnections, where household wastewater plumbing is incorrectly connected to a land surface drain carrying sewage directly to a watercourse.

6.1.3 Selection of MST techniques

While a range of MST techniques are available, community-based MST has only previously been tested once in the UK (Chapter 3) and was noted to be more sensitive than certain marker-based methods using culturable organisms. It was therefore hypothesised that community-based MST would be useful where low levels of pollution are expected. One limitation of MST methods is their poor relationship with regulatory faecal indicator organisms (FIOs), which may limit the applicability of MST methods to regulatory frameworks such as the BWD (BWD, 2006/7/EC). A poor relationship between the human proportion of the bacterial community predicted by community-based MST and culturable *E.coli* was observed in a previous case study (Chapter 3). This could be due to the differential die-off of culturable organisms compared to DNA (Wanjugi *et* *al*., 2016). Using qPCR to target the *RodA* gene may overcome some of these limitations and improve the relationship between MST methods using nucleic acid detection methods and regulatory FIOs; this also supports the Environment Agency's (EA's) call for more evidence on the use of DNA-based techniques (Rhodes, 2016). In addition, the recently developed human-associated markers in the genomes of *E.coli* (Gomi *et al*., 2014; Chapter 4) may link MST methods to regulatory parameters. The Hu100 marker had the highest average abundance among sewage in the UK (Chapter 4), and was selected, as it has never been used in a catchment investigation. The HF183 marker was also selected for use in the Seaton Sluice catchment since it is the most common human-associated marker (Harwood *et al*., 2014), and therefore forms a reasonable baseline on which to compare the performance of the Hu100 marker and community-based MST.

6.2 Aims of the study

This investigation aimed to identify the likely sources and areas of pollution that are reducing river and bathing water quality in the Seaton Sluice catchment, ideally when CSOs were not discharging to the river. In doing so, the objectives set out were:

1. Conduct a sampling campaign of the catchment to identify areas with high concentrations of FIOs.
2. Compare the use of the *RodA* gene to the enumeration of *E.coli* using regulatory, culture-based methods to give a more rapid method to monitor *E.coli* concentrations in river and sea-water samples.
3. To use community analysis techniques to identify potential sources of pollution throughout the catchment.
4. Compare community analysis to library dependent approached using the Hu100 *E.coli* biomarker and HF183 marker or human pollution.

6.3 Methods

6.3.1 Study design

The sampling regime followed the EA’s bathing water sampling regime (Appendix E.1) to align results from the MST investigation with regulatory bathing water results. The weekly sampling day is randomly assigned, reducing any unintentional bias in the sampling regime. In addition, sampling was conducted on two extra days (07/11/2016 and 22/11/2016) to capture two rainfall events (Appendix E.3), when CSOs were overflowing since no significant rainfall events were captured during the sampling regime above. Fifteen sample locations were identified throughout the catchment (Figure 6.1) including the EA’s bathing water sampling location. Sampling locations were chosen to be upstream and downstream of urban areas and river confluences. In addition, a sampling location just above the tidal limit (Sample location 2, Figure 6.1 and Figure 6.3) was chosen to define the output from the catchment, and a sample location at the harbour (sample location 1, Figure 6.1) was chosen to capture all land surface drains and CSOs between sample location 2 (Figure 6.1) and the harbour entrance. Sample locations 10, 9 and 6 (Figure 6.2) represent streams entering the main river (Seaton Burn), which is made up of sample locations 14, 13, 12, 8, 7, 5, 4, 3, and 2 with sample location 1 at the harbour entrance.

In total, sampling took place on 18 different days during the bathing water season, between 4 May 2016 and 20 September 2016. In addition, two additional sampling days were collected during rainfall events on the 7 and 22 November 2016. A total of 299 samples (20 bathing and 279 river water) were collected and used for faecal indicator organism

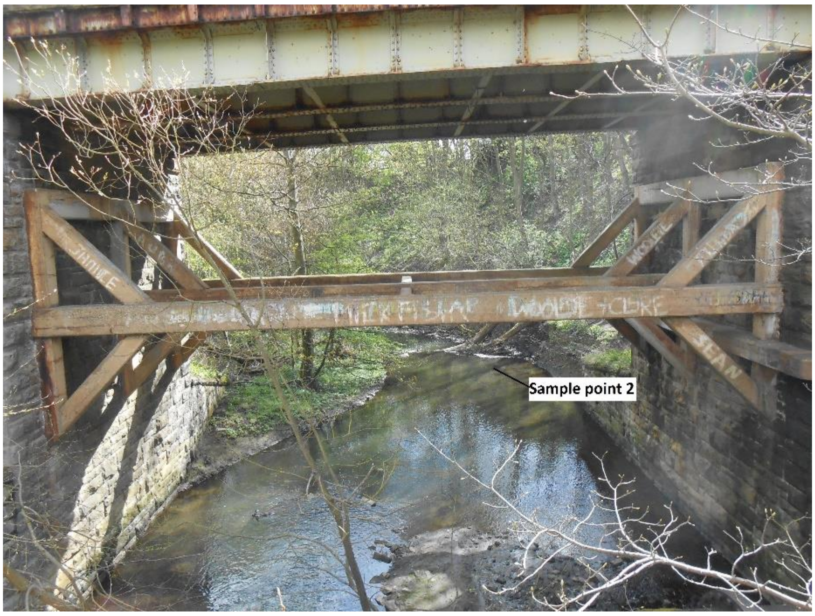


Figure 6.2. Sample location 2, at the bottom of the catchment above the tidal limit.

analysis. One sample (22/11/2016, sample location 2) could not be collected as high river levels prevented safe access to this location.

6.3.2 Overall microbial source tracking strategy

A tiered approach to MST was taken whereby the lower cost methods were conducted first on a large number of samples before more costly methods were conducted on a targeted subset of samples. The concentration of culturable *E.coli* in all samples (n = 299) was used to select twelve sample days (n = 179) for analysis of the *RodA* gene and the Hu100 and HF183 human markers by qPCR. From these twelve sample days, six (n = 92) were then selected for community analysis. Sample days were selected to give a mixture of days when samples had higher or lower levels of *E.coli*, and across different levels of rainfall (Appendix E.3) and which were either impacted or not impacted by CSO discharges (Appendix E.4).

6.3.3 Sample collection and transport

Catchment samples (Sample locations 1-14, Figure 6.1) were collected in three 250 mL, autoclaved, and acid washed bottles. Samples were collected using a sampling pole from 30 cm below the surface, where possible, of flowing water without disturbing the bed of the river. The EA sampling team kindly collected duplicate seawater samples during regulatory sampling at the Seaton Sluice bathing water. On two occasions, not on bathing water sampling days, seawater samples were collected in the same manner (2.1.4 Environmental water samples) to capture high-rainfall events that would be otherwise missed. One sample was not collected (Sample location 2, 22/11/2016) when the river was in flood, as the sample location was not safely accessible. Sediment samples were not collected as the river-bed is mostly rocky and it would be difficult to collect enough sediment to get a representative sample.

6.3.4 Enumeration of faecal indicator organisms

Samples were transported on ice, returned to the laboratory within 3 hours of collection and once at the laboratory, stored at 4 ^{o}C until processed (< 3 hours). *E.coli* was enumerated using membrane filtration (2.2 Enumeration of faecal indicator organisms). A single sample was not reported (Sample location 10, 07/11/2016) as all plates appeared un-readable, producing no distinct colonies.

6.3.5 Isolation of DNA for qPCR and sequencing

DNA was isolated from 250 mL and 800 mL of river and seawater, respectively. On one occasion (22/11/2016), 100 mL of river water was used due to high turbidity in the samples. These volumes are greater than previously used as it was suggested that this could improve the accuracy of SourceTracker analysis (Sassoubre, *et al.*, 2015) and higher volumes improve the assay sensitivity of qPCR.

6.3.6 Quantitative PCR (qPCR)

QPCR was carried out as previously described (2.4.2 Quantitative PCR (qPCR)). The HF183 marker was chosen because it is the most commonly used MST marker for human sources and therefore is useful as a comparator for other source tracking markers and methods. For river and seawater samples the limit of detection (LOD) and limit of quantification (LOQ) correspond to theoretical values of 24 and 8, and 60 and 19 gene copies (gc)/100 mL, respectively (2.4.2 Quantitative PCR (qPCR)). However, in practice, LOD and LOQ values are likely to be higher than these due to other quality controls (Table 2.3). For example, at low concentrations the relative standard deviation of gene copy numbers is more likely to exceed 25% than at higher concentrations. Moreover, this value relies on all processes, such as DNA extraction and PCR amplification being 100% efficient which is unlikely (Kralik & Ricchi, 2017). Salmon DNA and the Sketa primers were used to test for low DNA extraction efficiency and PCR inhibition (Chapter 2.4.2 Quantitative PCR (qPCR)). No tested samples exceeded the control for DNA extraction efficiency (>3 Ct difference between standard, Table 2.3); however, a single sample (Sample location 14, 12/08/16) was removed from testing as the concentrations of the *RodA*, HF183 and Hu100 markers all appeared to be inhibited and dilution resulted in gene copy values below the limit of detection.

6.3.7 Bioinformatics

Bioinformatics analysis was conducted as previously described (2.6.5 Analysis of data from Illumina sequencing) using the DADA2 algorithm (Callahan *et al.*, 2016). Two faecal samples (one cow and one pig) were removed from the analysis as the entire sample was composed of a single taxon. A single environmental sample (14/07/2016, sample location 14) was removed from analysis due to a low read count (64). The final faecal taxon library (FTL) used as input into SourceTracker, was composed of 10 pig, 10 cow, 10 sheep, 10 dog, 10 horse, 10 chicken, and 14 Sewage samples (Chapter 5). The FTL was run by combining all sewage sources as previously recommended and separating cow and sheep into individual sources. Separating cow and sheep sources was reasonable as all cow or sheep contamination was less than 10%, the proportion at which no false positives were observed (Chapter 5). SourceTracker was run five times, sources with a relative standard deviation greater than 100% were considered false positive results and removed (Henry *et al*., 2016). Sources with a predicted contribution of less than 1% were removed to reduce the likelihood of false positives, as previously recommended (Chapter 5).

6.3.8 Rainfall and GIS data

Daily rainfall data was obtained from the Met-office integrated data archive system (MIDAS) land and marine surface stations dataset (Met-office, 2012). The dataset was subset for rainfall data from a weather station located at Blyth (Latitude: 55060, Longitude: -1611), which is slightly outside of the catchment, but is the closest Met-office weather station.

Land use data was extracted from the most recent (2015) land cover map (Rowland *et al*., 2017).

6.3.9 Statistical analysis

Seventeen and four percent of samples analysed for Hu100 and HF183 by qPCR, respectively, had values between the LOD and LOQ (Appendix E.4). These values were included in further quantitative statistical analysis since using zero values or half the LOQ would under-represent the level of human pollution, and using an RSD of 25% (2.4.2

Quantitative PCR (qPCR)) improves confidence in these values. Nonetheless, it should be appreciated that the gene copy number quoted for these values has a lower level of confidence associated with it. All assays below the limit of detection were reported as 0 gc/ 100 mL, since this reflects how the EA report regulatory *E.coli* counts for the bathing water directive (BWD, 2006/7/EC).

The normality of each dataset was assessed visually though histograms and quantitatively using the Shapiro-Wilk test for normality (Wilk & Shapiro, 1965). Log-transformation of the data reduced skewness, however, all datasets; the culturable *E.coli*, *RodA* gene, HF183 marker, and Hu100 marker concentrations, remained significantly different to a normal distribution ($p < 0.01$), due to a large number of zero values in the datasets. The data were log-transformed for linear regression since normality is not an assumption of linear regression and to provide a comparison with other studies (Hassard *et al*., 2017; Noble *et al*., 2010), some of which use linear regression on non-normal data. However, all hypothesis testing was conducted using non-parametric tests. All correlations were evaluated by determination of the Spearman's rank correlation coefficient (r_s) (Spearman, 1904) and the Wilcoxon rank-sum test (Wilcoxon, 1945) was used to determine whether groups of data arise from the same population, i.e. to evaluate the difference between groups. The Wilcoxon rank-sum test was chosen since it does not assume independence between samples, which cannot be guaranteed in the catchment study.

6.4 Results

6.4.1 E.coli and RodA concentrations

6.4.2 Relationship between culturable E.coli and RodA gene concentrations

Culturable *E.coli* concentrations were significantly correlated with *RodA* gene copy concentrations ($r_s = 0.843$, p = $<2.2\text{x}10^{-16}$) and Figure 6.3 shows a strong, positive linear relationship between these variables (adjusted $r^2 = 0.666$, p = $<2.2\text{x}10^{-16}$). The *RodA* gene could be expected to be close or below the LOD for 12/14 samples, where the *RodA* gene is below the LOD and culturable *E.coli* (> 1 CFU/100 mL) are observed (Figure 6.3).

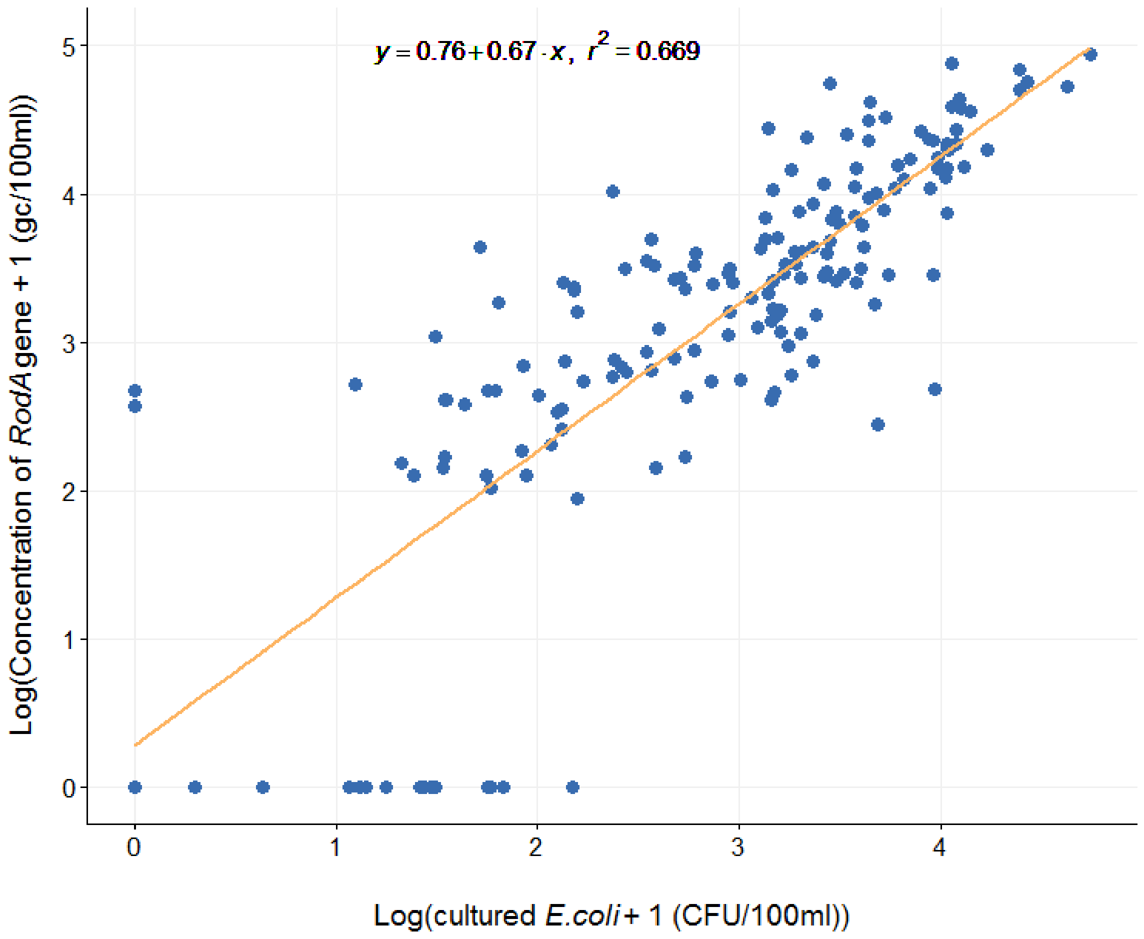


Figure 6.6.3. Relationship between culturable E.coli and RodA gene concentrations in 180 seawater (n= 12) and river water (n = 168) samples. Zero values are shown here. Removing these values changes the fit of the regression to y=0.12+0.86x with an r^2 =0.665.

On one occasion (14/07/2016), no culturable *E.coli* were detected, however, 370 ± 12 gc/100 mL of *RodA* gene were noted in a bathing water sample (Appendix E.4). This was the single seawater sample where culturable *E.coli* and the *RodA* gene did not fall within the same BWD classification (Appendix E.4). For 96% of samples analysed, using the *RodA* gene would result in the same or a lower level of classification according to the BWD (Table 6.2). On 4% of occasions, the *RodA* gene resulted in a lower classification. It should be noted that the standards for transitional (marine) and coastal waters were used, as these are more stringent than those for inland waters (<500 and <1000 for ‘Excellent’ and ‘Good’ status, respectively), and are closer in value, making it more likely that any difference between the culturable *E.coli* and *RodA* concentrations will result in a difference in classification.

Table 6.2. The number and percentage of water samples in which the culturable E.coli and RodA gene concentrations would result in the same or different bathing water classification, according to the BWD

	Number	**Percentage**
Total number of samples	177	100%
Samples with identical classifications	137	77.4%
Samples with different classifications	40	22.6%
E.coli classification more stringent than *RodA*	7	4.0%
RodA classification more stringent than *E.coli*	33	18.6%

BWD classifications for *E.coli* concentrations determined from a single sample are 0 – 249 CFU/100 mL "Excellent", 250 – 499 CFU/100 mL "Good", and >499 "Poor". The "Sufficient classification was not assigned since differentiation between this and the "Poor" classification is based on the percentile values of a four year data set.

6.4.3 E.coli concentrations in the catchment

The concentration of culturable *E.coli* and *RodA* genes in the catchment ranged between 0 (< 24 and 8 gc/100 mL for river and sea water, respectively) and 0 (< 24 and 8 gc/100 mL for river and sea water, respectively) and 86,572 gc/100 mL, respectively (Figures 6.4 and 6.5).

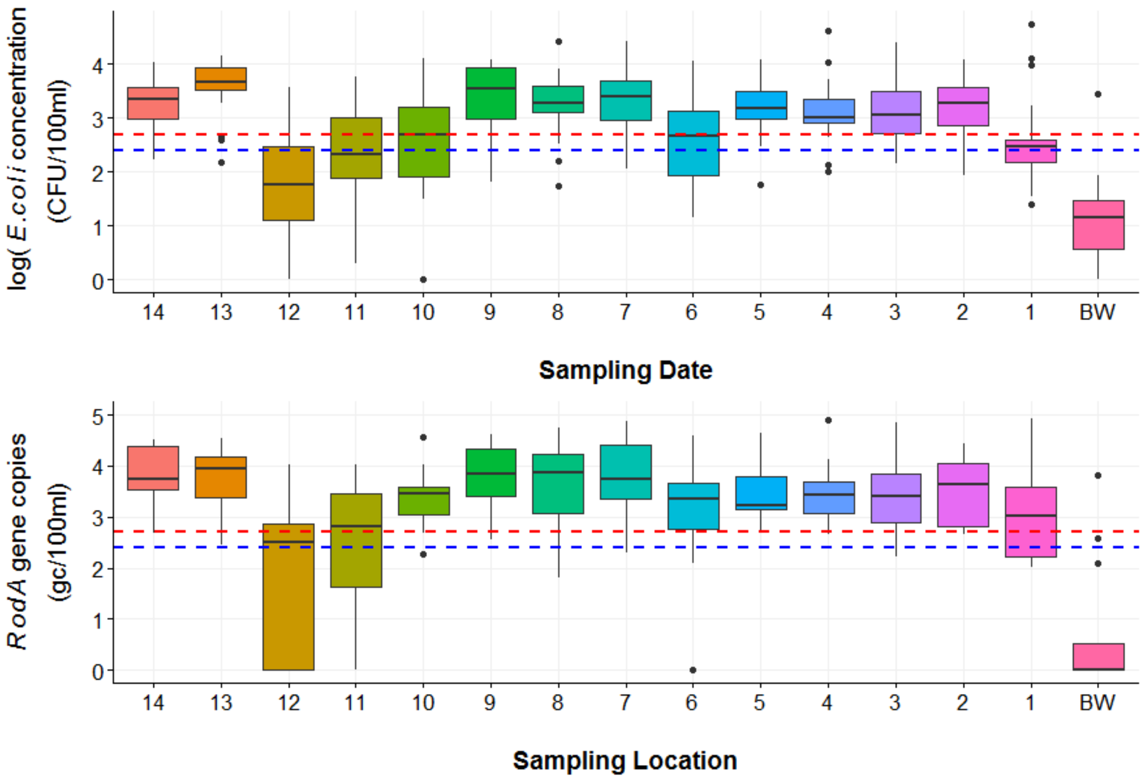


Figure 6.6.4. Culturable E.coli (Top) and RodA gene (Bottom) concentration at each sample location down the catchment (from left to right). Blue and red lines indicate the concentrations relating to the classification of bathing waters as Excellent and Good, respectively for marine and coastal waters. Culturable E.coli data is from 22 sampling days (n=328), and RodA data is from a subset of 12 of those sampling days (n=178). Colours used to identify individual sampling locations.

Sample locations 14 and 13, at the top of the catchment were consistently high (Figure 6.4). While there was a significant increase in the culturable *E.coli* between sample location 14 and 13 (p = 0.001209), There was no significant difference between the *RodA* gene concentrations (p = 0.8501). There was a significant decrease in the culturable *E.coli* (p = 1.335x10^{-05}) and *RodA* (p = 0.0093) concentrations between sample locations 13 and 12; this is likely due to the Big Waters nature reserve (Figure 6.1), which contains a large subsidence pond. The concentrations of culturable *E.coli* and *RodA* in the Bathing Water (BW, Figure 6.4) are significantly lower than at the harbour entrance (Sample location 1, Figure 1) (r_s = 1.907 x10^{-06} and p = 0.0002, respectively), likely due to the effect of dilution.

6.4.4 Human pollution sources

Relationship between E.coli, human markers, and community analysis.

At least one human marker was found in 68% of samples tested (Table 6.3). In 79% of samples, both human markers were in agreement, i.e., both HF183 and Hu100 were present or absent which provides strong evidence for the presence of human pollution in the catchment. The HF183 marker was identified more often compared than the Hu100 marker, and in samples when both markers were identified; HF183 was significantly more abundant ($p = 9.406e^{-14}$).

Table 6.3. Comparison of the presence/ absence of HF183 and Hu100 among river and sea-water samples detected through qPCR.

	Number	**Percentage**
Number of samples	179	100
At least one marker present	122	68
Both present or absent	142	79
Just HF183 present	20	11
Just Hu100 present	18	10

Culturable *E.coli* concentrations were positively correlated with HF183 ($r_s = 0.603$, $p < 2.2e^{-16}$) and Hu100 ($r_s = 0.656$, $p < 2.2e^{-16}$) marker concentrations (Table 6.4), suggesting that human sources are, at least in part, responsible for the elevated concentrations of *E.coli* in the catchment. Surprisingly, the correlation between the *RodA* gene and human markers appears to be slightly weaker than that between culturable *E.coli* and human markers (Table 6.4). Nevertheless, all of the relationships between measures of *E.coli* and human markers are positive and highly statistically significant (Table 6.4).

The contribution of human sources to the microbial community was estimated using SourceTracker (D Knights *et al*., 2011). There was a stronger correlation between the predicted contribution of human sources, culturable *E.coli* and the *RodA* gene, compared to the human (HF183 and Hu100) markers with culturable *E.coli* and *RodA* (Table 6.4), which, may be due to the higher sensitivity of SourceTracker compared to qPCR (Table 6.5).

Table 6.4. Spearman's rank correlation coefficients for the relationship between culturable E.coli, the RodA gene, HF183 and Hu100 human marker concentrations and the proportion of the microbial community arising from human sources as predicted by SourceTracker.

	Culturable *E.coli*			
RodA* gene	0.843 (p < 2.20x10^{-16})	***RodA* gene**		
HF183 marker*	0.603 (p < 2.20x10^{-16})	0.554 (p = 1.27x10^{-15})	**HF183 marker**	
Hu100 marker*	0.656 (p < 2.20x10^{-16})	0.613 (p < 2.20x10^{-16})	0.695 (p < 2.20x10^{-16})	**Hu100 marker**
SourceTracker predicted human contribution**	0.693 (p = 9.91x10^{-14})	0.665 (p = 1.65x10^{-12})	0.686 (p = 1.70x10^{-13})	0.467 (p = 4.45x10^{-06})

*n = 178, **n=89

SourceTracker identified human pollution in 95% of samples tested compared to human markers; for which at least one marker was detected in 82% of the same samples (Table 6.5). There were no occasions where markers detected human pollution and SourceTracker did not detect human pollution (Table 6.5). The weaker correlation between SourceTracker human predictions and the Hu100 marker compared to the HF183 marker concentrations is likely to be due to the greater number of samples that Hu100 was below the LOD, compared to that of HF183 (29 compared to 19, respectively) in the dataset used for community analysis.

Table 6.5. Indicating the co-occurrence of human pollution indicated by SourceTracker and human markers determined through qPCR.

	Number	**Percentage**
Number of samples	88	100%
SourceTracker positive for human sources	84	95%
At least one marker and SourceTracker in agreement (present or absent)	72	82%
Only SourceTracker positive	16	18%
Only markers positive	0	0%

Human sources in the catchment

Due to the considerable variation in marker concentrations and SourceTracker predictions at each sample location between sampling days (Figure 6.5 and Figure 6.6), determining significant differences between sample locations across the entire dataset was difficult. Nevertheless, SourceTracker predictions of human faecal contamination for the bathing water sampling location (BW, Figure 6.5) were significantly lower than that for sample location 1 ($p = 0.02$), and sample location 12 had significantly lower human contributions than sample location 13 ($p = 0.001$). The SourceTracker predictions at sampling locations 13 and 14 were not significantly different to each other ($p > 0.05$), although on 4 of 5 occasions SourceTracker predictions increased between sample locations 13 and 14 (Figure 6.5).

The HF183 and Hu100 markers show a similar pattern to the SourceTracker human pollution predictions, indicating high human pollution at the top of the catchment. There was a significant decrease in HF183 and Hu100 between sample locations 13 and 12 ($p = 0.0003$), although, there was only a significant decrease in the HF183 concentration between sample location 1 and the bathing water samples (BW) ($p = 0.03$, Figure 5). Differences between other, adjacent sampling locations were not significant ($p >0.05$).

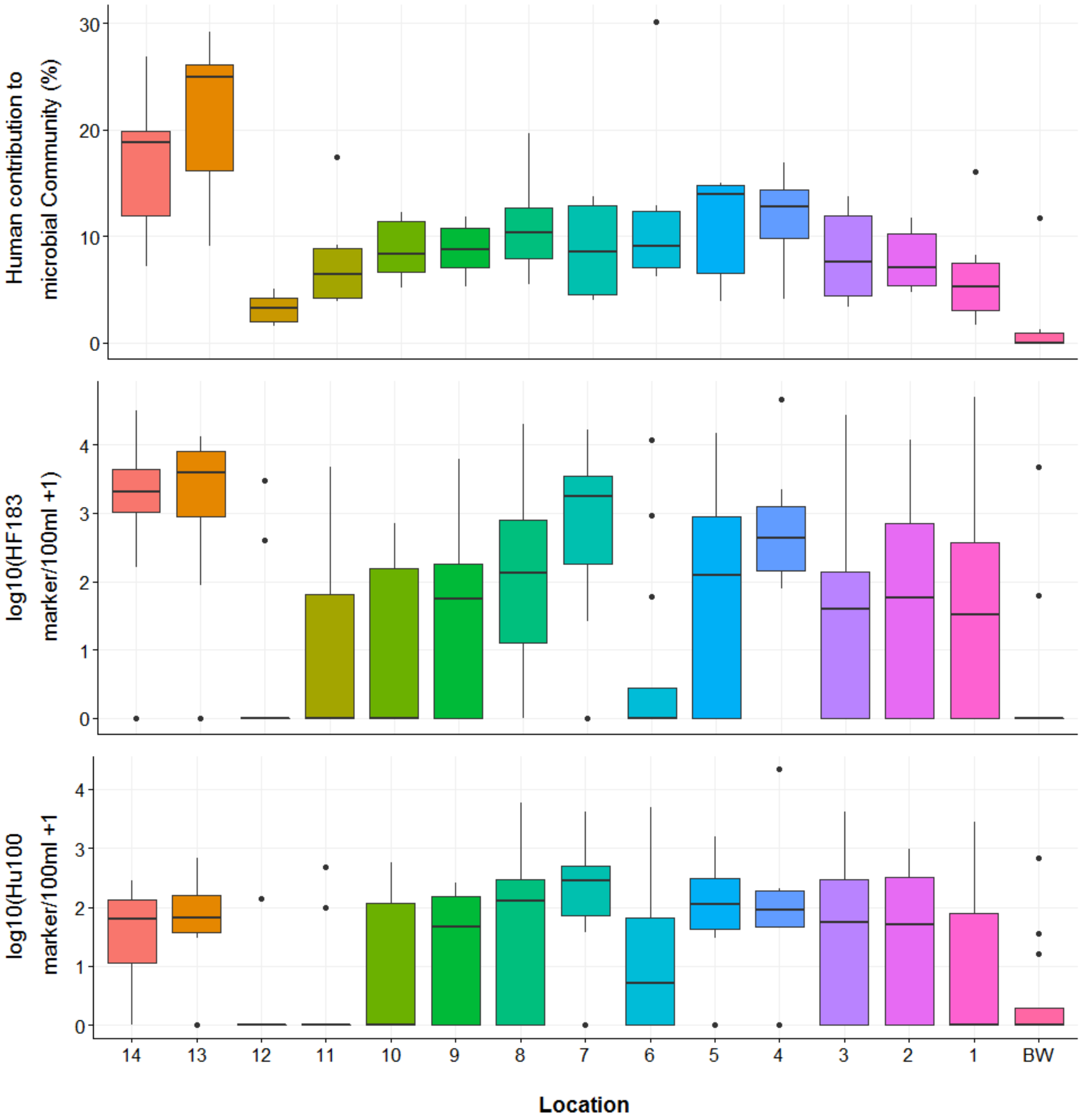


Figure 6.6.5. SourceTracker predictions of the human contribution to the microbial community for each sampling location over six days. Colours used to identify individual sampling locations.

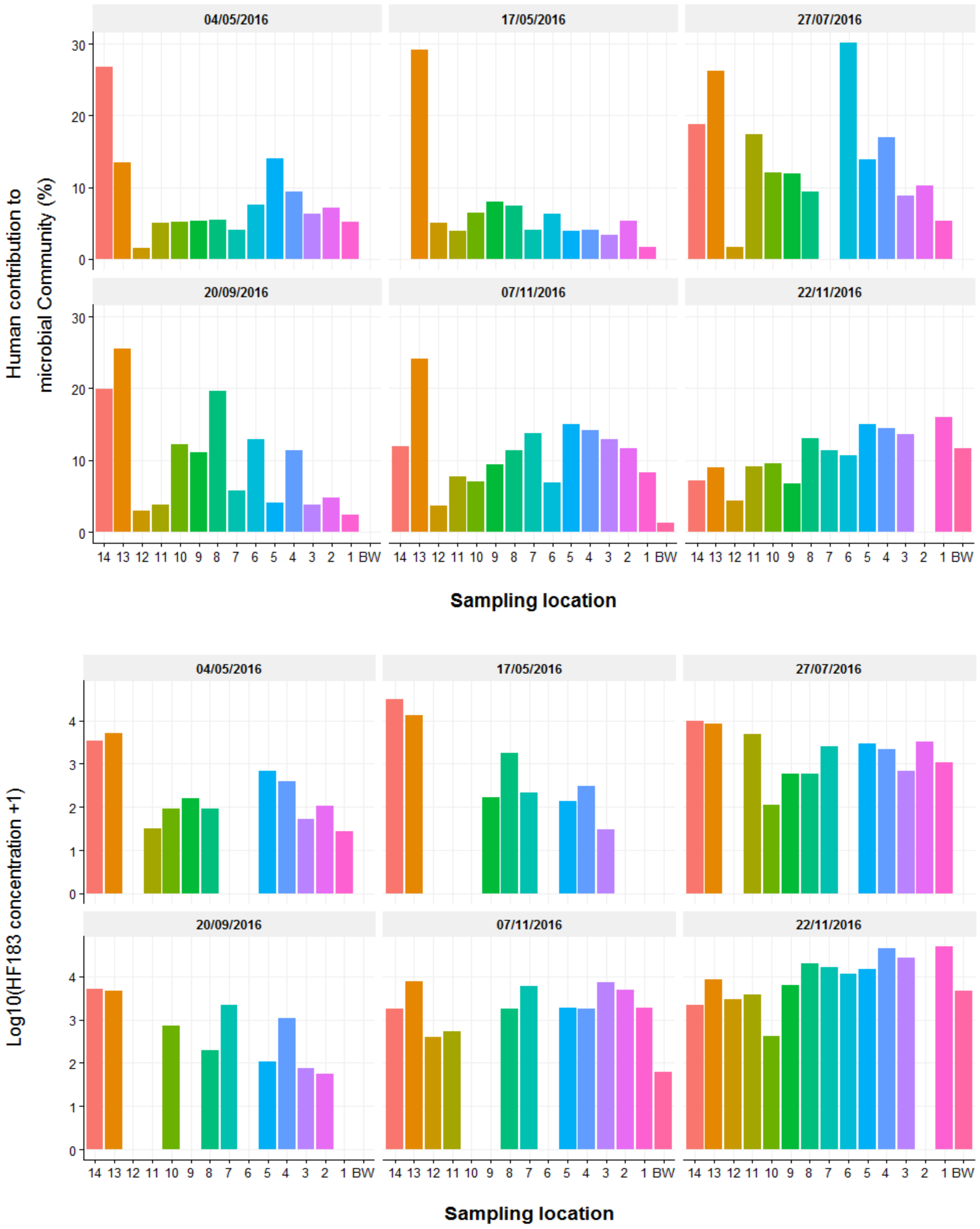


Figure 6.6. SourceTracker predictions of human contribution to the microbial community at each location on each sampling day (Top). HF183 concentrations at each location on each sampling day for which community analysis was undertaken (Bottom). NB Sample 14, 17/05/2016 (Top) was removed due to low read counts following sequencing, and sample 2 22/11/2016 was not collected due to inaccessibility of the sampling location. Colours used to identify individual sampling locations.

Combined sewer overflows or misconnections?

To differentiate between samples which are potentially impacted by discharging CSOs and those which are not, the dates and durations of CSO spills were obtained from telemetry data from Northumbrian Water Ltd. Three days were impacted (07/11/2016 and 22/11/2016), or partially impacted (27/07/2016) by CSO spills within 12 hours before sampling. Notably, there is no CSO above sample location 14 in the catchment (Figure 6.1) although human pollution was identified in all samples by community analysis at sample location 14, and in 92% (11/12) of samples for at least one marker.

Figure 6.7 shows the HF183 data split into samples which were potentially impacted by CSOs (n = 3) and those not impacted by CSOs (n = 9). The concentrations of HF183 on sample days which are potentially impacted by CSOs are significantly higher ($p = 6.719 \times 10^{-12}$), typically by an order of magnitude, than those for non-impacted sample days. The Hu100 marker concentrations (Appendix D.5) and community analysis predications (Figure 6, Sample days not impacted by CSOs are 04/05/2016, 17/05/2016 and 20/09/2016) also reflect the difference between samples potentially impacted and not impacted by CSO discharges; days impacted by CSOs having significantly higher concentrations of Hu100 ($p = 1.113 \times 10^{-07}$).

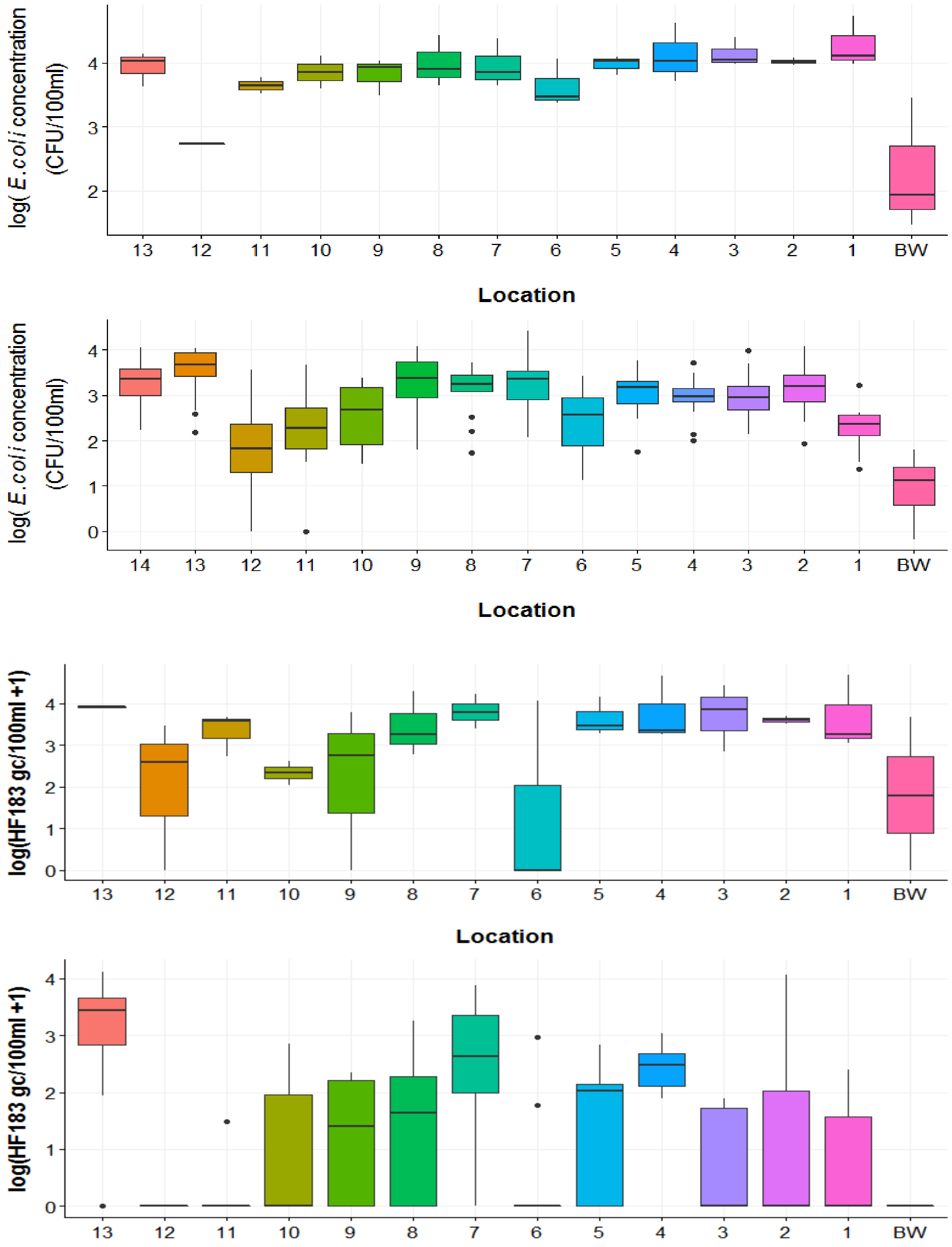


Figure 6.6.7. From top to bottom, 1) The E.coli concentration at each sample location while CSOs were overflowing, 2) E.coli concentration at each sample location while CSOs were not overflowing, 3) HF183 marker concentrations at sample locations when CSOs were overflowing and 4) HF183 marker concentrations at sample locations when CSOs were not overflowing. Colours used to identify individual sampling locations.

There appears to be a base concentration of human pollution in the catchment, particularly around sample location 13, 7, and 4 (Figure 6.6). Between sample location 14 and 13 there is a land surface drain and a CSO, and four of five occasions sample location 13 had a higher human proportion of the microbial community (Figures 6.6 and 6.7). The land surface drain and CSO were sampled on 07/11/2016 when both were flowing. As expected, the microbial community of the CSO effluent was composed of around 50% human while the land surface drain was composed of ~10% human (Figure 6.8).

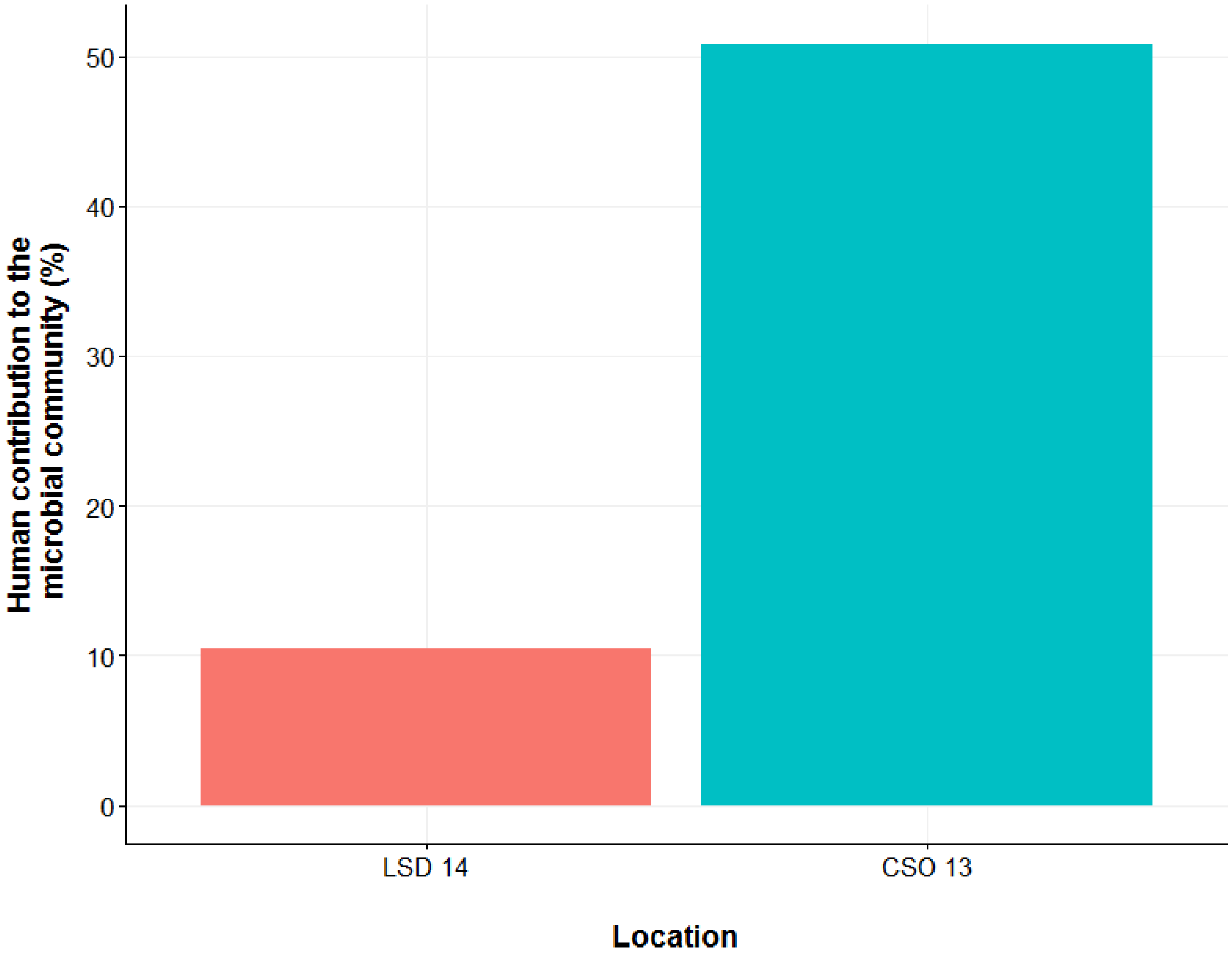


Figure 6.6.8. The human proportion of microbial community identified in a land surface drain (LSD 14) and combined sewer overflow (CSO 13) on the 7/11/16.

6.4.5 Other sources of pollution

SourceTracker was run twice with sheep and cow as separate sources, or as a combined ruminant source in the FTL. Due to the low level of ruminant pollution, the FTL with separate sources was more appropriate (Chapter 5).

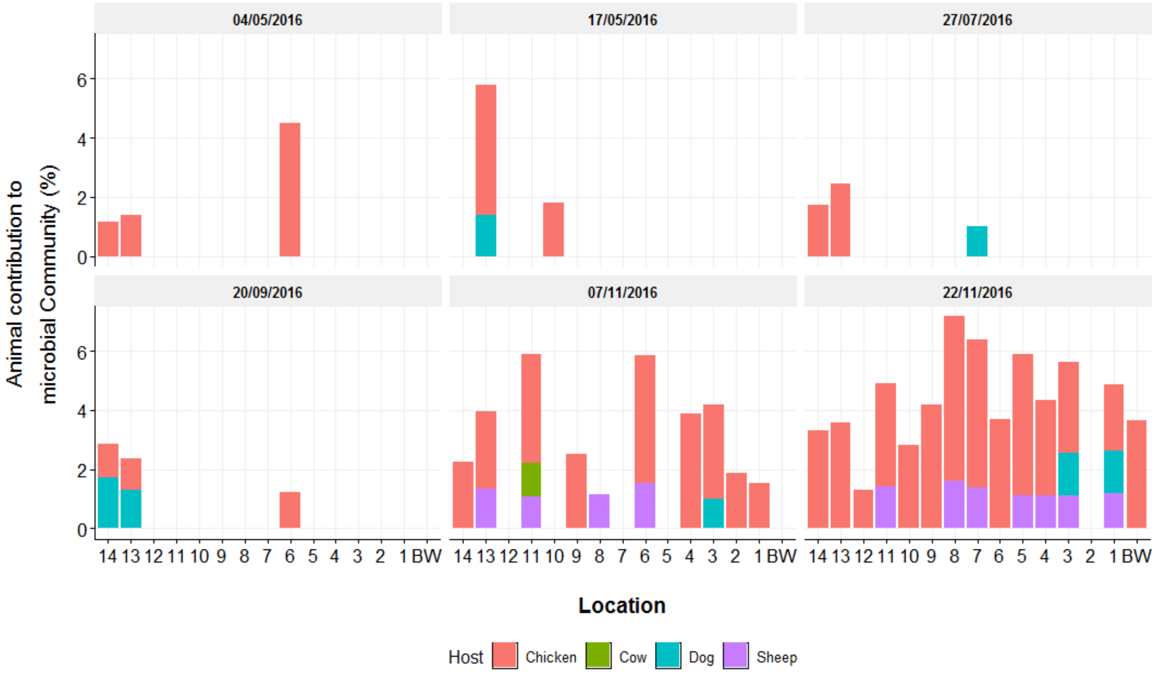


Figure 6.9. Proportion of the microbial community attributed to animal sources by SourceTracker. The human contribution is not shown since it is generally much larger, and prevents comparison between samples for animal sources. NB Sample 14, 17/05/2016 was removed due to low read counts following sequencing, and sample 2 22/11/2016 was not collected due to inaccessibility of the sampling location.

Chicken sources were the most commonly detected animal faecal source in the catchment (36% of samples, Figure 6.9), followed by sheep (12% of samples, Figure 6.9) and dog (8% of samples, Figure 6.9). Dog faecal sources in some samples (17/05/2016 and 20/09/2016, sample location 13, Figure 6.9) are potentially false positive results since the proportion of sewage is greater than 25% (Figure 6.6), previously identified as the contribution at which there is potential for false positive results (Table 5.3). Cow faeces was identified as a source in only a single sample (07/11/2016, location 11, Figure 6.9). In all samples, the predicted human contribution was larger than animal contributions (Figure 6.6 and 6.9). Animal faecal sources were identified in one of six bathing water samples (22/11/2016, figure 6.9) where chicken faeces accounted for ~ 4% of the microbial community (sewage accounted for 14%).

6.5 Discussion

6.5.1 Identification of E.coli with the RodA gene

Current culture-based regulatory assays to determine the microbiological quality of water require a 24-48 hour incubation step, limiting the timely communication of health risk and pollution events (Korajkic *et al*., 2014). As such, the EA asked for additional evidence into the use of DNA techniques to monitor water quality (Rhodes, 2016). While the USEPA accept the use of qPCR, targeting the 23S rRNA gene (23S gene) of enterococci, to monitor bathing water quality, no regulatory assays exist for the enumeration of *E.coli*. This was the first time the *RodA* gene has been used to monitor the quality of environmental waters in a catchment study. The good relationship (r_S = 0.843, p = <2.2x10^{-16}) between culturable *E.coli* and *RodA* gene concentrations (Figure 6.3) suggests that qPCR may be a useful technique to monitor *E.coli* and is similar, in terms of the strength of correlation, to relationships between culturable enterococci and the 23S gene in water samples (Noble *et al*., 2010) and culturable *E.coli* and the *UidA* gene in cattle faeces (Oliver, *et al*., 2016) and river and sea water samples during the summer. Interestingly, the relationship between culturable *E.coli* and qPCR targeting the *UidA* gene have shown a seasonal effect in previous studies (Oliver *et al*. 2016; Hassard *et al*. 2017), with an improved relationship in the UK summer in water samples (Hassard *et al*., 2017).

The high proportion of samples (96%) where the water quality classifications were the same or more stringent for the *RodA* gene compared to culturable *E.coli* would indicate that it is a conservative marker, which may be more acceptable in terms of risk for regulators. The disparity in classification may be due to the presence of viable but non-culturable organisms, the differential decay of culturable organisms and DNA, and differences in the specificity of DNA-based and culture-based assays (Hassard *et al*., 2017), or differences in sample processing. Previous studies, targeting either the 23S or *UidA* genes (Hassard *et al*., 2017; Oliver *et al*., 2016) also observed an over-estimation of *E.coli* determined through qPCR, compared to culture-based methods. While using the 23S gene may overestimate due to its multi-copy nature and lower specificity, the *RodA* gene has been suggested to be single copy and highly specific to *E.coli* (Chern *et al*., 2011). The overestimation is, therefore, likely to be due to the differential decay, and the presence of viable but non-culturable organisms. The higher (better) classification from the resulting analysis of the *RodA* gene, observed in 4% of samples, may be a concern for public health if monitoring with qPCR replaces culture-based monitoring. This overestimation may be due to the high specificity of the *RodA* gene, since the *RodA* gene can differentiate *E.coli* from other species of *Escherichia* (Chern *et al*., 2011), it is possible that other *Escherichia spp.* or other coliforms may be cultured; for example, a previous study observed that 9% of cultured isolates on a selective media were not *E.coli* (Perkins *et al*., 2014), leading to an overestimation by culture-based methods. A previous study using qPCR to target the *UidA* gene also reported a higher concentration of *UidA* gene copies than culturable *E.coli* in 29% (8/27) of the summer samples, although none of these were across a BWD classification boundary (Table 6.2) (Hassard *et al*., 2017). The possibility that differences in culture-based and DNA-based assays are due to discrepancies in the sampling and preparation procedures cannot be discounted. Samples used for culture-based analysis and DNA-based analysis were collected in separate collection bottles, there is the potential for conditions to change between these samples being collected, for example, if the river or sea bed is disturbed between each sample being taken. It may also be due to a loss of DNA during DNA extraction. While addition of salmon DNA to the lysis buffer (2.4.1 Polymerase Chain Reaction (PCR)) is used to determine the efficiency of DNA extraction, it only measures loss of material, rather than lysis efficiency, since raw DNA is added. So while DNA extraction passed quality control (within three cycle threshold values of salmon DNA), the lysis step may be inefficient, resulting in reduced *RodA* gene copies being observed. However, this is unlikely as since *E. coli* is a Gram-negative planktonic organism that is easily lysed.

The good relationship between culturable *E.coli* and *RodA* gene concentrations may also make MST conclusions draw from nucleic-acid-detection-methods more applicable to the water industry.

6.5.2 Human marker and community analysis based MST

This was the first catchment study using Hu100 in the UK. The good correlation between Hu100 and HF183 ($r_s = 0.695$, $p < 2.20\text{x}10^{-16}$, Table 6.4), and Hu100 and the predicted human proportion of the microbial community ($r_s = 0.467$, $p = 4.45\text{x}10^{-06}$, Table 6.4) supports the use of the Hu100 marker for future MST studies. The higher correlation between Hu100 and culturable *E.coli* and the *RodA* gene compared to that of HF183 could indicate that using an *E.coli* biomarker gives a better prediction of the total *E.coli* coming from human sources than the HF183 gene. Despite the better correlation exhibited by Hu100 with *E.coli* assays, both markers were necessary to identify sources of pollution in the catchment and HF183 had a significantly higher concentration when both markers were detected. The use of both markers was particularly important higher in the catchment (Sample locations 13 and 14), when Hu100 showed very low concentrations (Appendix E.6) compared to HF183 (Figure 6.8); 4 out of 24 Hu100 samples were below the LOD, and a further 7 were below the LOQ, when both HF183 and SourceTracker showed a large amount of human pollution. This is likely due to the variation in this marker between different human populations (Chapter 5), which is likely to be exacerbated since there is only likely to be a single septic tank (Sample location 14) or misconnection (Sample location 13) at this location in the catchment (there are no conurbations – see Figure 6.1).

Community-based MST was valuable in the detection of human sources of pollution, particularly at low concentrations, and improved confidence in conclusions drawn from the Hu100 and HF183 marker data. The greater sensitivity of community-based MST over qPCR, noted previously (Neave *et al*., 2014; Chapter 3), was evident with a greater proportion of samples being positive for human sources (18%, Table 6.5) where markers were likely to be below the LOD. The strong, significant correlation between the predicted human proportion of the microbial community and human-associated markers ($r_S = 0.686$, $p < 2.20\text{x}10^{-16}$, and $r_S = 0.467$, $p = 4.45\text{x}10^{-06}$, for HF183 and Hu100, respectively, Table 6.4) is essential in building confidence in MST results since the marker concentrations were often close to or below the LOQ. This highlights the potential of community-based MST in future MST studies where low levels of pollution are problematic, which, may become increasingly relevant to the UK water industry as water quality improves and identifying and tackling low-levels of diffuse sources of pollution becomes increasingly important (Figure 1.1). The strong correlation between community-based and marker-based MST was not observed in previous studies (Ahmed *et al*., 2015; Chapter 3) and may be a reflection of the dominance of human pollution in this catchment compared to non-human sources in other studies such as ruminant (Chapter 3) and bird sources (Ahmed, *et al.*, 2015). Interpretation of results from community-based MST must, therefore, be approached with caution. For example, it may not be fair to compare the human proportion between two samples where one contains only human sources and the other contains high levels of non-human sources. Where additional sources are present, the relative proportion of human pollution, as reported by SourceTracker, will be artificially compressed, reducing any correlation between the predicted human proportion and human markers. The stronger correlation observed here, may also be due to a more accurate FTL, for which the effect of including samples had been previously tested (Chapter 5). In Chapter 3, the human sources in the FTL included a single, separate septic tank and sewage from three small treatment works located outside the catchment, which may not reflect the human inputs from other septic tanks or human inputs in the catchment. In comparison, this FTL consisted of a larger number of sources so is more likely to reflect the human inputs. A different bioinformatics approach may also explain the better correlation. Error correcting algorithms such as DADA2 used here, produce less noise and erroneous sequences compared with cluster-based algorithms, such as QIIME 1, used in Chapter 3, and does not call singletons (Callahan *et al*., 2016); a lower level of noise in the outputs may lead to less shared OTUs, also observed elsewhere (Coello-Garcia, 2018), and reduce conflation between sources. In addition, in this study a sea water and river water sample was used as a background sample. Previous studies have reported improvements in SourceTracker predictions (Brown, *et al*., 2018; Chapter 4), and correlations with FIO (Hägglund *et al*., 2018) when indigenous microbiota were input into SourceTracker as a source. In Chapter 3, there were no samples free of pollution to be used as background samples, and (Ahmed *et al*., 2015) conducted their study before this advice was available. These differences in the methods used may also account for improved correlations observed in this study compared to previous studies.

6.5.3 Seaton catchment

All three MST methods highlighted widespread human pollution throughout the catchment, both during permitted overflow of CSOs and when no CSO overflows were recorded or observed (Figure 6.7 and Appendix E.5). The good correlation observed between both markers, community-based MST and *E.coli* assays support human pollution as the primary source of pollution driving high levels of *E.coli* in the catchment.

Human source of pollution contribute greatly to the poor and moderate WFD classifications for total phosphorus and phytoplankton (Table 6.1), likely leading to algal blooms in the Big Waters reservoir. The Big Waters reservoir mediates the high levels of pollution observed higher in the catchment (Sample location 12, Figure 6.6). The high level of human pollution at the top of the catchment (Sample locations 14 and 13, Figures 6.4, 6.5, and 6.6) was unexpected, particularly at sample location 14 where there were no known inputs. It is highly likely that the pollution at sample location 14 was due to a septic tank discharging into the Seaton Burn above sample location 14. There is also human pollution entering the Seaton Burn between sample location 13 and 14 from a CSO when overflowing, and a land surface drain. While dog, sheep, and chicken faeces were identified in some of the samples from these locations, they were at much lower proportions than human sources (Figures 6.9 and 6.6) suggesting that mitigation measures should focus on reducing human inputs through the identification and, ideally, removal of the septic tank and reduction in the number of misconnections from housing in Dinnington. This is likely to prove difficult due to the current development of new housing estates in this area, which could increase the number of misconnections.

Misconnections also likely contribute human pollution further down the catchment. On days when no CSOs were overflowing (04/05/2016, 17/05/2016 and 20/09/2016), human pollution was identified by community-based MST at all sample locations, except for the bathing water sampling location. The Hu100 (Appendix E.5) and HF183 (Figure 6.7), human-associated marker data support this finding. These misconnections will unquestionably contribute to WFD failures in the Seaton Burn for three measures: Invertebrates, mitigation measures assessment, and fish; although misconnections are not currently considered a potential pressure (Table 6.1). The identification and mitigation of misconnections are burdensome and cost the UK water industry around £235 million year^{-1} (Royal Haskoning, 2007), largely due to the difficulty in locating possible areas with high levels of misconnection. In the Seaton Sluice catchment, urban areas account for only 2.5% of land use (Table 6.1) but contribute almost all faecal contamination observed in the catchment. Being able to identify priority areas is, therefore, valuable to the UK water industry. In the Seaton Sluice catchment, Dinnington and Seghill (Figure 6.1) are two priority areas where misconnection mitigation efforts should be directed. At Dinnington, new developments are likely to worsen the human pollution entering the catchment above sample location 13, and the frequency of algal blooms at Big Waters reservoir may be reduced if misconnections at Dinnington are mitigated. At Seghill (Sample locations 7, 6 and 5), both human-associated markers show a general, although not statistically significant, increase between sample location 8 and 7 (Figure 6.7) and the median marker concentrations decrease between sample location 7 and the bottom of the catchment (Sample location 1, Figure 6.7).

While bathing water failures were only observed while CSOs were discharging in heavy rainfall, the wide-spread human pollution in the catchment suggest that any bathing water failures occurring when CSOs were not overflowing must be due to either urban diffuse pollution sources (misconnections) or animals, such as dogs and horses which are exercised regularly, defecating directly on the beach. Although misconnections are unlikely to lead to a failure according to the BWD on a given day, a ‘perfect storm’ scenario where FIOs from misconnections survive for a prolonged period in sediments (Craig, *et al*., 2004; Anderson, *et al*., 2005), and when a rainfall event which is not large enough to cause CSOs to discharge, but could mobilize sediments and transport FIOs associated with those sediments to the bathing water sampling location. Reducing the number of misconnections entering the Seaton Burn may, in addition to improving river water quality, increase the robustness of the bathing water classification. Further work to understand the survival and transportation of *E.coli* and pathogens in sediments would be beneficial. This would allow future studies to apply modelling efforts to determine the potential for misconnections to contribute to bathing water failures in this manner.

6.6 Conclusions and recommendations

The *RodA* gene gave a reasonable correlation to the culturable *E.coli* counts. Further testing of this gene in environmental samples is necessary before it could be considered as a regulatory replacement for culturable counts. Nevertheless, it appears to be a good target for MST studies looking to relate MST methods to FIOs rapidly.

Future studies using marker-based MST should consider the use of human-associated biomarkers alongside other markers, such as HF183, particularly where low levels of pollution are expected since the use of a single marker resulted in 10% more false positive results.

Community-based MST using SourceTracker (Knights *et al*. 2011) was noted to be more sensitive than marker-based MST methods and the proportion of the microbial community attributed to human sources showed a good correlation with *E.coli* assays (cultured *E.coli* $r_s = 0.693$, $p = 9.91\text{x}10^{-14}$, and *RodA* gene $r_s = 0.665$, $p = 1.65\text{x}10^{-12}$). The good, significant correlation between both HF183 and Hu100 and community-based MST ($r_s = 0.686$, $p = 1.70\text{x}10^{-13}$, and, $r_s = 0.467$, $p = 4.45\text{x}10^{-06}$, respectively) supports the use of community-based MST in future studies, particularly where low levels of pollution may limit the usefulness of markers.

Although urban areas account for only 2.5% of land use, urban pollution accounted for almost all of the faecal pollution observed. Misconnections were identified as being wide-spread in the Seaton Sluice catchment. Community-based and marker-based MST was used with CSO monitoring data to identify human pollution throughout the catchment when no inputs were expected. Misconnections should be added to the list of pressures in the Seaton Sluice catchment, and river basin management plans should be updated accordingly.

Chapter 7 Embedding MST in the UK water industry

7.1 Introduction

The UK water industry is in need of tools to inform decisions about the investment and management of environmental waters. The quality of environmental waters has improved significantly since the implementation of the Water Framework (2000/60/EC, European Commission, 2000) and Bathing Water Directives (2006/7/EC (CEU, 2006)). However, there is still work to be done. Only 62% of bathing waters are classified as 'Excellent', significantly less than the European average of 85% (EEA, 2015). Similarly, the quality of UK surface waters, graded by the WFD (2000/60/EC), remains concerning with only 36% of UK rivers achieving good status (Priestley, 2015), a figure which has remained stagnant for ~ nine years. On the part of the water industry, previous improvements to water quality have largely been achieved through capital investment in sewage infrastructure (Figure 1.1) for example: diverting sewage away from smaller, problematic treatment plants; installing increased storage to prevent spillages from combined sewer outfalls (CSOs) and pumping stations; and maintaining of CSOs. Investing in capital assets, however, is likely to result in diminishing returns on investment. As we remove the major problems, other more diffuse sources such as septic tanks and misconnections are likely to present greater challenges. Identifying and mitigating diffuse pollution remains difficult, and could exacerbate the level of investment required to achieve water quality improvements.

Microbial source tracking could be an important tool to help inform investment decisions and more clearly determine the pressures on a waterbody to efficiently improve water quality. However, microbial source tracking is currently used little in the water industry, and MST with sequencing not at all. There may be a number of reasons for this. The technology and expertise required to conduct MST, such as qPCR and DNA sequencing, are not readily available to the water industry, expertise is required to interpret MST data correctly, and little is known about the economic costs and benefits of MST making it difficult to build a business case or understand how best to integrate MST into daily workflows.

The aims of this chapter are to:

- Evaluate the options and costs associated with integrating MST into the daily operations of Northumbrian Water (NWL).
- Identify the market opportunities for MST to help build a business case for the use of MST in the water industry.
- Consider how the water industry could take full advantage of MST methods in the future.

To achieve these aims the market opportunities for MST are discussed. The costs of using the MST techniques developed and evaluated during this thesis are then used to compare options to integrate MST into Northumbrian Water's operations.

7.2 Methods

Options for and costs associated with the integration of MST methods into Northumbrian Water operations.

Three options to integrate MST methods into NWL were explored (Table 7.1), each with a different degree of internalisation into NWL operations. While there are a large number of options to export different techniques to outside laboratories, the three options which represent full internalisation of methods, exporting all methods and partially outsourcing methods were explored.

Table 7.1 The three options considered in this study to integrate MST into Northumbrian Water workflows

Option	Description	Requirements
Fully internalize all MST methods	This would involve NWL carrying out all aspects of an MST investigation including planning, sampling, sample processing, MST techniques and analysis and reporting of results.	Capital investment in laboratory equipment Increased operational costs from labour and equipment maintenance. Investment in training of staff in laboratory techniques and MST analysis/reporting.
Partially outsource MST methods	This would allow NWL to undertake the planning, sampling and sample preparation before exporting the sequencing and qPCR/ PCR methods to a commercial laboratory. NWL would then maintain responsibility for analysis and reporting of data.	Capital for the necessary equipment for sample preparation is currently owned by NWL Investment in training and expertise in MST analysis and reporting.
Fully outsource MST methods	This would either involve NWL completing sampling and a contractor completing the rest of the MST investigation. All laboratory, analysis and reporting are undertaken by an external service provider.	No additional capital or operational requirements.

For each option, the costs were estimated for three MST methods used during this thesis (Figure 7.1). The culturing and gene detection method is not recommended for the water industry due to the large variation in marker sensitivity (Chapters 4 and 5) This method requires considerable labour to culture and pick a large number of isolates prior

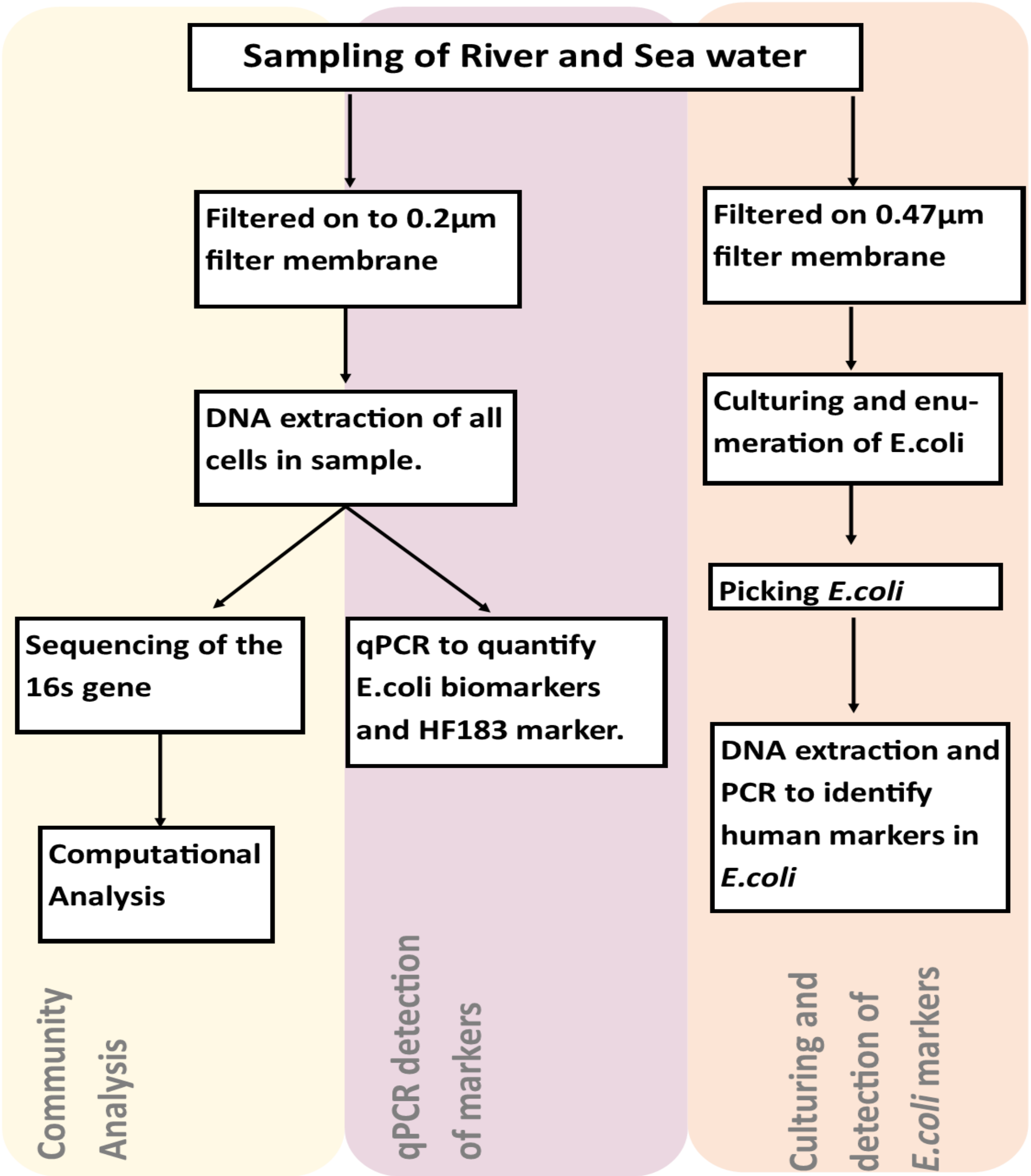

Sampling of River and Sea water
Filtered on to 0.2µm filter membrane
DNA extraction of all cells in sample.
Sequencing of the 16s gene
Computational Analysis
qPCR to quantify E.coli biomarkers and HF183 marker.
Filtered on 0.47µm filter membrane
Culturing and enumeration of E.coli
Picking *E.coli*
DNA extraction and PCR to identify human markers in *E.coli*
Community Analysis
qPCR detection of markers
Culturing and detection of *E.coli* markers

Figure 7.1 Three possible routes to identify sources of faecal pollution in river and seawater samples: Community Analysis; qPCR detection of markers; culturing and detection of E.coli markers.

to screening for markers. In addition, it was difficult to find any commercial laboratories who would offer this service. The culturing and picking technique was, therefore, not considered in the costing of options involving outsourcing of techniques.

The capital costs of integrating each technique, laboratory consumable costs and the costs of exporting laboratory techniques to service providers were estimated from quotes or supplier/provider websites. The actual quotes are not disclosed since they may contain business sensitive and/or client specific information, however, a cost sheet (downloadable here) is available which can be easily updated from quotes, allowing anyone planning an

MST investigation, or considering integrating MST into their own workflows to quickly repeat this analysis.

Consumable costs were estimated, taking into account the number of samples which could be processed by each consumable item. For example, during qPCR 24 samples can be processed simultaneously on a single 96 well plate. The cost of the plate remains the same, even if only 4 samples are processed simultaneously. The optimum number of samples for each consumable and the costs per sample calculated. Figure 7.2 shows how the consumable cost per sample (Figure 7.2, top) and total consumables cost (Figure 7.2, bottom) vary with sample size for qPCR based MST methods.

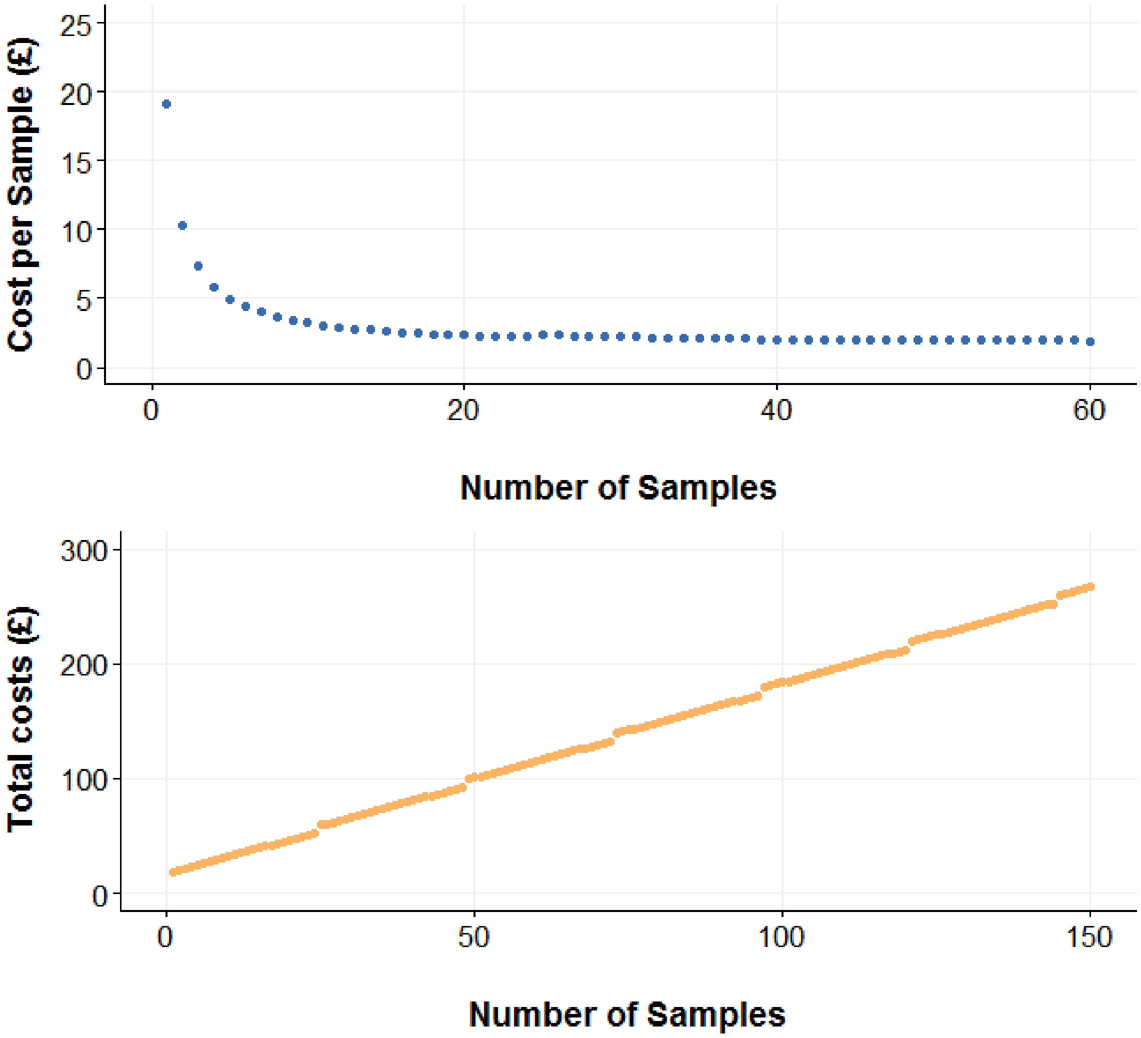


Figure 7.2 Variation in the cost per sample (top) and total consumable cost (bottom) with sample size for consumables required to perform qPCR based MST.

To calculate labour costs, the staff time required to undertake each technique were conservatively estimated based on personal experience gained during this project. An hourly wage of £10 h^{-1} (~£18,000) was assumed. The optimum number of samples was estimated by considering the maximum number of samples which could be simultaneously processed during a rate-limiting step in each MST protocol. This was used to estimate the staff cost with varying sample size. For example, in DNA extraction only 24 samples can be inserted into a machine which takes 25 minutes to run, and little else can be achieved during that time. Batches of 24 samples was assumed to be the optimal number of DNA extractions.

The options were considered based on the cost of implementing each option with each technique where possible. This approach was chosen since there may be little commercialisation value for the research and the selection of which techniques to use will vary depending on the aims of individual projects and the logistics of sampling. A number of assumptions were necessary for costing and comparing the options in Table 7.1. These include:

- The additional sampling required to undertake MST investigations is constant between all three. While the actual cost of sampling will vary with the scope and scale of each project and the selection of MST techniques, these costs will be reasonably consistent across different options.
- The laboratory staff-time is valued at £10 h^{-1}, approximately £18,000 yr^{-1}.
- There is no automation of laboratory processes. Many of the laboratory processes can be automated with greater capital expenditure. It was assumed that sample processing was all done by hand where possible.
- Additional personnel and training requirements were not taken into account since there may be staff with previous expertise in these techniques, and it would be difficult to evaluate the current staff capacity to undertake these techniques, particularly on an *ad hoc*, project by project basis.

- That 150 samples per sequencing run on an Illumina Miseq would give adequate sequencing depth (This represented the sample sizes used during this project (Chapter 6).
- The staff training required to interpret these results were not taken into account since no one offers this training in the UK, and there may already be in-house expertise which remains unknown.

7.2.1 Market opportunities

To evaluate market opportunities, willingness to pay descriptors were taken from a recent survey and the predicted value of water quality to the local economy was used.

7.3 Results and discussion

7.3.1 Option 1

To carry out all MST methods internally, Northumbrian Water would require ~£230,000 of capital investment in additional laboratory equipment (Table 7.2). The required capital was estimated based on the additional equipment required. For example, the laboratory already possesses incubators and a vacuum system, the costs of these items are therefore excluded.

Table.7.2. Summary of the capital costs required for each technique in option 1 - to fully internalize all MST methods

Technique	Capital cost (£)
qPCR	59,272
Sequencing	180,608
qPCR and Sequencing	210,452
Culture *E.coli* and identify human markers using PCR	17,727
Total (Three techniques)	228,180

It is worth noting that this does not include the additional training requirements, nor the cost of additional auditing required by the United Kingdom Accreditation Service (UKAS), to maintain these standards across Northumbrian water laboratory facilities. The addition of new methods will likely result in the need for a reassessment. Case studies provided by UKAS suggest that an initial assessment can cost between £7,000 and £15,000 (UKAS, 2018).

QPCR has the lowest operational costs independent of sample size, requiring the least staff time to run. At smaller sample sizes (< 33) *E.coli* culturing methods are less expensive than sequencing, however, if sample sizes exceed 33 the high throughput nature of sequencing makes it cheaper than the relatively low throughput culturing methods.

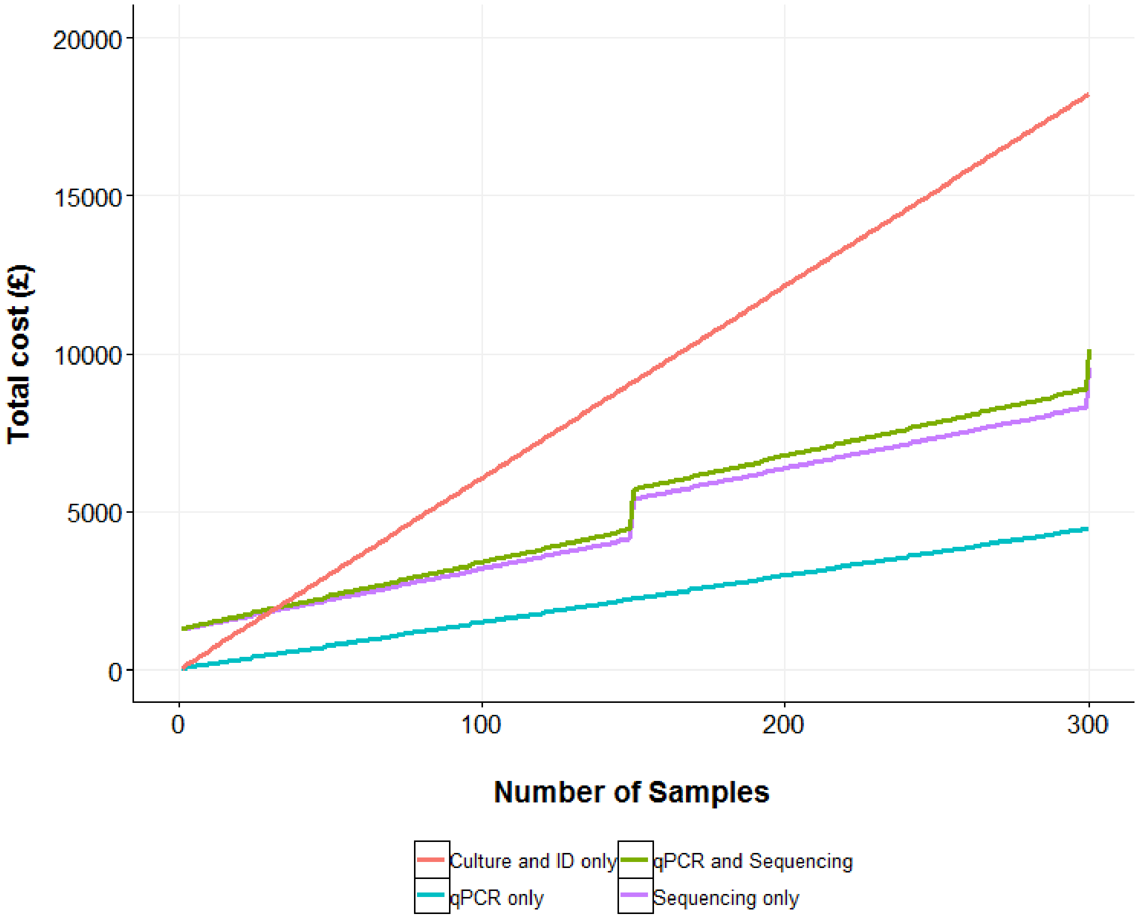


Figure 7.3 Operational costs for each source tracking technique if methods are fully internalised

While qPCR is the least expensive technique on a per sample basis, further consideration is required. A single qPCR run will only identify a single source, or up to three sources if a multiplex reaction is used, although a decrease in sensitivity can be expected. Sequencing can be used to identify multiple sources simultaneously and therefore if more than 80-100 samples are to be processed to identify three different sources of pollution, sequencing may be comparable to qPCR in terms of operational cost. It is worth noting that using a single qPCR marker, or even a single technique, is rarely advisable and sequencing alongside qPCR may be particularly useful.

Identifying human pollution with *E.coli* biomarkers by traditional culture-based techniques (as opposed to qPCR) is a lot more expensive than qPCR, and when processing >40 samples is comparable in cost to sequencing. This is due to the laborious nature of this technique.

This option, where Northumbrian Water conducts all analysis in-house, carries the largest business risk since it requires the largest capital investment. There is also business risk with the current rapid development of sequencing technology since the technology may be outdated quickly. It would also require the largest investment in personnel in terms of training and the additional time required.

7.3.2 Option 2

In order to outsource the sequencing and qPCR/ PCR techniques, Northumbrian Water would conduct filtration and DNA extraction steps (Figure 7.1), before sending the extracted DNA for analysis. There would be some capital costs associated with the equipment required for DNA extraction and additional filtration equipment.

Table 7.3. Summary of the capital requirements of option 2 – to partially outsource MST methods

Technique	Capital cost (£)
Membrane filtration	2,826
DNA extraction and quantification	26,601
Total	29,427

The capital costs of option 2 (Table 7.3) are much lower than internalising all laboratory processing (Table 7.2), the operational costs are higher (Figure 7.4). It is worth noting that the additional transport costs to ship samples from Northumbrian Water to the commercial laboratory are not taken into account here, although these are likely to be nominal relative to the cost of laboratory testing.

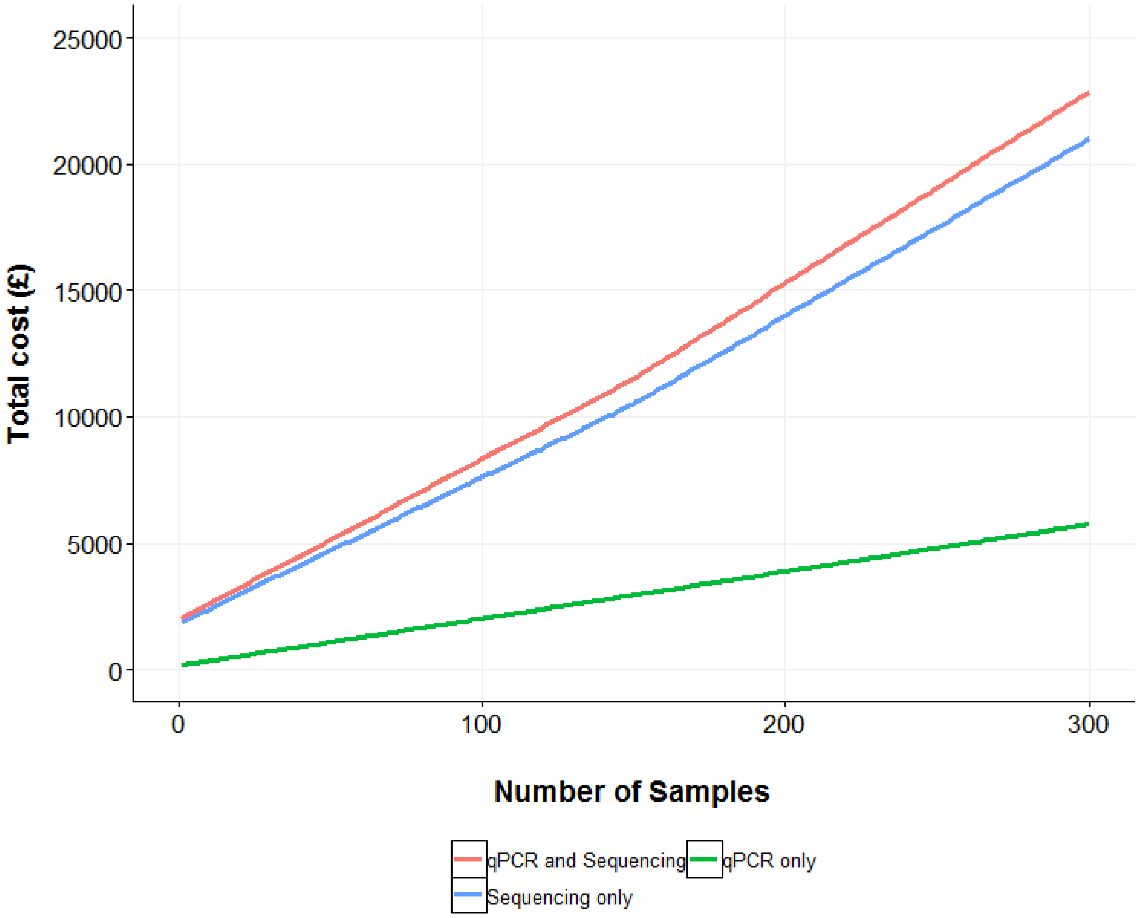


Figure 7.4 Operational expenditure (including staff time) of each technique if the main analysis is outsourced to a commercial laboratory.

7.3.3 Option 3

To fully outsource all microbial source tracking techniques, Northumbrian Water would only be responsible for the membrane filtration of samples, which is currently supported at the moment, and transport of samples to a commercial laboratory. The total expenditure (totex) costs, are shown per sample in Figure 7.5, although it is important to note that while commercial laboratories can carry out the techniques, it is unlikely that they will have expertise in conducting microbial source tracking investigations. It is also worth noting that few commercial laboratories are familiar with environmental samples, most commercial laboratories are optimized to process medical samples (blood, tissues, etc.). Only a single laboratory (Environment Agency, 2018b) employs MST in England,

and currently only use qPCR. It is also worth noting that they also do not provide any advice on their reported results. Northumbrian water will, therefore, still require some in-house expertise to interpret the results to inform decision-making.

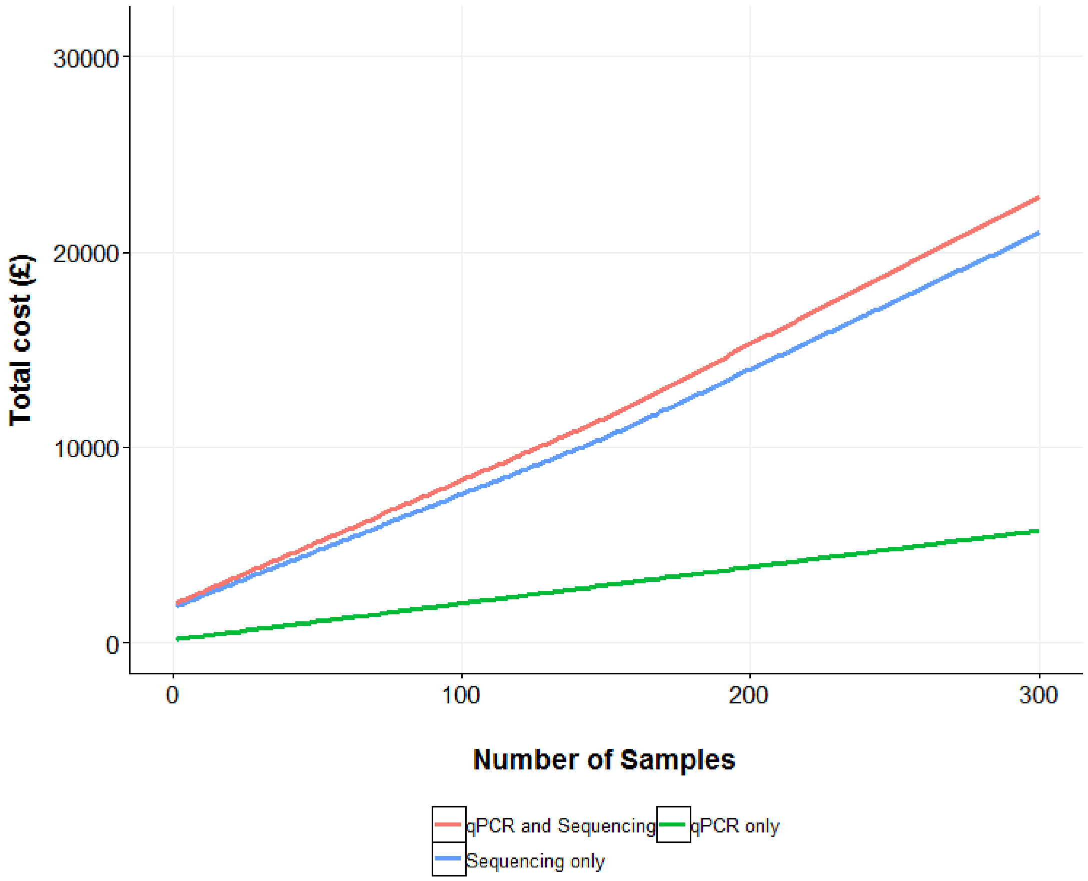


Figure 7.5 Operational expenditure (opex), including staff time, of option 3, with different sample sizes

7.3.4 Comparison of options

To compare options, the variation in totex, capital plus operational expenditure, of each option was calculated. The variation in totex with sample size was used to compare options. Sample size was chosen as the comparator, as opposed to time since this a reasonable proxy for the amount of MST work which may be undertaken by Northumbrian Water. Additionally, the cost of MST work depends directly on the size and number of projects which are undertaken and is reasonably easy to estimate.

If using only qPCR based techniques, option 2 appears the most cost-effective. After processing 374 samples, option 2 is more cost-effective than option 3 (Figure 7.6).

However, internalising all techniques (option 1) only becomes more cost-effective than option 2 when >1,300 samples are processed. This is due to the large costs associated with the staff time required to undertake DNA extraction which NWL staff undertake in both options 1 and 2.

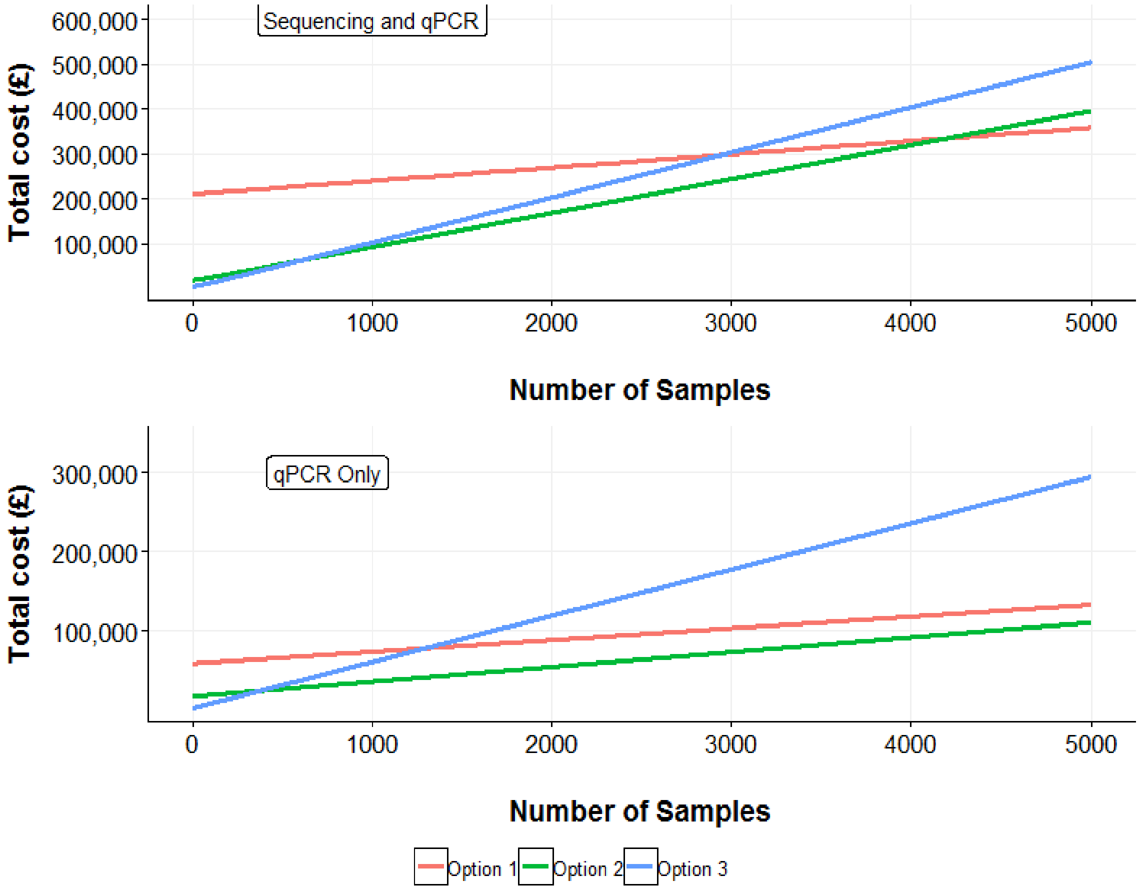


Figure 7.6 Comparison of the total expenditure (totex) for each option only using qPCR and sequencing-based techniques (top) and only qPCR techniques (bottom).

When both qPCR and sequencing techniques are considered, 618 samples are required to make option 2 more cost-effective than option 3, i.e., if Northumbrian Water will process more than 618 samples for sequencing and qPCR, 3,000 samples would need to be processed before option 1 was more cost-effective than option 2. This is due to the costs of the large amount of staff time required to undertake DNA extraction it is worth investing in DNA extraction equipment.

Internalising operations (option 1) becomes more cost-effective than options 2 and 3 when more than 4200 and 3000 samples have been processed, respectively. This is due to the higher capital investment required in option 1 (Tables 7.2 and 7.3).

7.3.5 Comparison of options in case studies

As a comparison, Table 7.4 shows the costs of conducting the Morland (Chapter 3) and Seaton (Chapter 6) case studies using each option.

Table.7.4. Estimated operational costs of the case studies undertaken in this thesis for each option to integrate MST into Northumbrian Water's workflows

Case Study	MST Methods	Number of Samples	Option 1 Cost (£)	Option 2 Cost (£)	Option 3 Cost (£)	NLS* Cost (£)
Morland (Chapter 3)	Culturing and PCR	36	2,190	ND	ND	ND
	Sequencing	46	2,140	4,470	5,600	NA**
Seaton Sluice (Chapter 6)	qPCR single marker	168	2,500	3,280	9,940	25,200
	Sequencing	169	5,790	11,850	15,992	NA**

**National Laboratory Service*

***Sequencing is not available as a service through the National Laboratory Service*

ND Not determined as no providers could be found which offer this service.

While the costs shown in Table 7.4 show clear operational savings from internalising MST methods, it does not include capital investment. It does, however, highlight how after an initial capital investment, making business cases for the use of MST becomes much easier.

7.4 Summary

The most appropriate MST methods to use depend on the intended use. However, the techniques considered here are established (qPCR) or emerging (Sequencing and *E.coli* biomarker) versatile techniques for most MST situations. The use of both qPCR and sequencing is recommended for MST studies. Using only qPCR risks false positive results since an entirely host-specific marker has not been identified. Additionally, the variability of markers between communities reduces the confidence in their use. Sequencing is able to detect a greater range of potential pollution sources without the development of new methods and may be more sensitive to pollution sources. However, using sequencing alone only gives us a qualitative understanding of the contribution of faecal sources, when a quantitative value is often required to make investment and management decisions. Using both sequencing and qPCR, therefore, appears to be a better option, particularly since using qPCR in addition to sequencing results in a relatively small increase in operational costs.

Determining the best option for Northumbrian Water to integrate MST into their operations is difficult and depends on the scale and scope of future projects. If MST is to be used only occasionally, for example when a bathing water sample has failed (only three samples in 2018), then option 3, outsourcing all laboratory techniques, is likely to be the most suitable option. If larger, catchment characterisation style projects are foreseen then option one or two will be most valuable.

Table 7.5. Other risks and benefits associated with each option to integrate MST into Northumbrian Waster's workflows

Option	Capital Costs	Opportunities	Risks
1	£228,180	Opens new technology to other areas of the business Staff development Commercialisation opportunities	Rapid advancement in sequencing technology can make investments out of date quickly. Potentially unexpected costs from increase laboratory inspections.
2	£29,427	Staff development	Difficult to maintain knowledge level in staff if techniques are not used regularly.
3	£0*		Lack of consultancies offering MST reduces the choice of service and could lead to poor quality/ expensive service. Difficult to maintain knowledge level in staff if techniques are not used regularly.

**This does not include the necessary staff time.*

Considering totex, Northumbrian Water would need to process more than 3,000 samples to make investing in qPCR and sequencing equipment cost-effective (i.e., selecting option 1). MST investigations involve a range of sample sizes of 24 (Hughes *et al*., 2017) to over 400 (Nguyen *et al*., 2018), depending on the type of MST investigation, the size of the water body or catchment and the resources available. Northumbrian Water would, therefore, need to conduct roughly between 10 and 100 MST projects before option 1 becomes cost-effective. However, there are a number of factors other than cost to consider (Table 7.5). While option 1 carries the largest business risk, requiring the largest capital investment, this risk is small due to the small level of capital investment required (£228,180, Table 7.5). Investing in laboratory techniques such as qPCR and sequencing makes this technology available to other areas of the business, not just for MST. Other areas of NWL could benefit from qPCR and sequencing technology, for example, using these techniques to compliment flow cytometry in assessing the quality of drinking water or to assess the ability of activated sludge plants to effectively remove nutrients. In addition, regulatory monitoring of bathing and drinking water may make use of PCR based methods in the future, for example, the USEPA has accepted the use of qPCR to enumerate enterococci in environmental waters since 2012 (USEPA, 2012). Having expertise in these areas prior to changes in regulation could be invaluable to Northumbrian Water, giving Northumbrian Water greater insight and sway in any consultations. This could also be an opportunity for Northumbrian Water to establish themselves as industry leaders in the use of MST and molecular techniques would bring reputational advantages and fit well with the NWL vision of being the leading provider of water and wastewater services (NWL, 2018).

Although option three presents the lowest financial risk, this option becomes less appealing when considering which subcontractor could do the work. Currently, only the National Laboratory Service (National Laboratory Service, 2018) offer MST using qPCR as a commercial service in the UK. This service gives a report for the total abundance of a marker in each sample and no advice is given regarding what these values mean. In addition, option 3 has the highest operational costs (Figure 7.5) and no sequencing service is offered. It would, therefore, be difficult to conduct meaningful MST investigations through option three at present. This highlights the potential of option 1 to develop MST as a commercial service within Northumbrian Water.

7.5 Market Opportunities for MST

7.5.1 Market Opportunities for the water industry (wastewater)

Evaluating the market opportunities for MST is challenging. MST techniques are often used to inform decision making and, therefore, some of the benefits of MST come from money not invested as much as money invested. Challenges to value MST also stem from the difficulty in valuing any environmental improvements which occur from decisions taken as a result of MST. Approaches to value environmental improvements, such as ecosystem service approaches, are often limited since they depend on assessing stakeholder's perception of the value of environmental improvements for example, through willingness to pay surveys.

For recreational waters, studies have taken a health-based approach, estimating the disease burden in terms of healthcare costs and lost work days, however, there is a lack of these studies in the UK. A disease burden of > $3.3 million per year (Dwight *et al*., 2005) was estimated for users of two California beaches, although visitor numbers and the cost of healthcare are likely to be much less in the North East of the UK. These studies also fail to take into account the positive health benefits of using recreational waters (1.2 Benefits of improving water quality) which could increase the value of bathing waters. What is possibly the most comprehensive economic evaluation in the UK (Phillips *et al*., 2018), reported beaches in Scotland to be worth between £0.8 million and £4 million per year to the local economy (Phillips *et al*., 2018) and while the quality of bathing water did not seem to affect the frequency of beach visits, poor water quality did diminish the quality of a beach visit. A previous willingness to pay study seems to support this value. Southern Water customers, both household and business, valued an improvement in a single bathing water site to good or excellent as between £642,000 and £1,048,000 per year (Accent, 2013). In the UK, of the 626 designated bathing water sites, only 62% achieve excellent, while 20 sites remain poor. This suggests that there are around 238 sites which could benefit from the use of MST which may be worth more than £190 million yr^{-1} to local economies, assuming a worth of £0.8 million yr^{-1}.

The value of surface water is more difficult to ascertain. Across Europe, over 50% of water bodies failed to achieve good status by 2015, a key milestone of the Water Framework Directive (WFD). An assessment of the implementation of the WFD highlighted the importance of correctly identifying the correct pressures on a water body (EC, 2015). Voulvoulis *et al*., (2017) report that 21 out of 27 Member States showed no clear links between pressures suggested to be impacting a waterbody and the programme of measures to monitor and alleviate these pressures. A poor understanding of the pressures on a water body could lead to wasted investment in attempts to remediate these pressures. For the Northumbria river-basin area alone, this investment is predicted to be around £820 million over the next 37 years, with the Northumbrian Water taking £440m of this financial burden. The efficient identification of pollution and its source is, therefore, critical to ensure that investment and management decisions are economically justified and evidence-based. The 79% of surface waters failing to achieve good status, therefore, also provide a market opportunity for MST. Southern Water customers valued an improvement in river water quality to good or high status between £14,500 and £23,110 per km of river (Accent, 2013).

For local water and wastewater companies, the use of MST techniques to improve water quality also provides opportunities to improve their reputation. Water companies are intricately linked to water quality, and often blamed, fairly or otherwise, for poor environmental water quality. There is, therefore, an opportunity to improve the reputation of Northumbrian Water by improving environmental water quality. Although sufficient water quality is adequate on a regulatory level and the fines are typically low where bathing water quality fails due to sewage pollution, poor or sufficient bathing water quality presents a risk to water companies' reputation.

For the water industry, justifying the use of MST to target bathing water quality improvements appears to be easier than river water quality improvements. Customers value the quality of bathing waters much more than river waters. However, typically, complex catchments discharge to bathing water sites, and the use of MST investigations will identify sources which affect both bathing and river water quality. Justifying the use of MST to improve bathing water quality from poor/satisfactory to good/excellent appears to be straight forward. However, customer willingness to pay for maintaining water quality, or improving the resilience of water quality is less certain.

7.5.2 Other markets for MST

There appears to be a large, currently untapped, potential for the use of MST to reduce the load on drinking water abstraction points or identify the sources of pollution entering groundwater sites. This could be through the management of catchments feeding abstraction points of the identification of leaky sewers, for example.

A variety of market opportunities exist, outside of the water industry, for MST including surveying food items for faecal contamination (Li, 2014), identifying the source of contamination of marine waters with non-native species, and identifying the sources of pollution in groundwater aquifers.

7.5.3 Current availability of MST in the UK

Currently, only two UK organisations carry out MST, the Environment Agency and the Scottish Environmental Protection Agency (Personal communication with Nathan Critchlow-Watton, April, 2018). The Environment Agency offer a commercial service, through their National Laboratory Service (National Laboratory Service, 2018), to detect some sources of faecal pollution through the detection of markers by qPCR (Figure 7.1) and charge an in-house rate of ~£150 per sample (Personal communication with Hannah Westerby, Environment Agency 23/5/18), suggesting there is some level of demand for MST.

Currently, the Environment Agency and the Scottish Environmental Protection Agency undertake MST using qPCR based techniques. Both agencies only undertake this work where bathing water quality is poor. The Environment Agency take additional samples which are tested using MST if a sample has a high bacterial count; the Scottish Environmental Protection Agency target catchments, which are feeding poor quality bathing waters, with a wider sampling campaign. Only the latter method is likely to result in joint benefits to bathing river water quality.

7.6 Conclusions and recommendations

The use of both qPCR and sequencing techniques is recommended. QPCR can identify a single source with a single marker and the qPCR method must be repeated if more than ~3-4 different sources are to be used. In comparison, sequencing allows the identification of multiple sources simultaneously, and for the identification of sources where a suitable marker for using qPCR has not been identified.

There is a range of options to incorporate qPCR and sequencing for MST into the daily operations of Northumbrian Water. The option to fully internalising MST techniques seems the most advantageous since, although this presents the largest business risk, the capital expenditure is <£300,000, and this option presents a range of other opportunities such as allowing other areas of the business to access these techniques to enhance innovation. Undertaking between 10 and 100 MST studies would be required to recoup the capital expenditure for option one through reduced operational expenditure, compared with option two.

Evaluating the benefits of MST in monetary terms is difficult. Further research into the monetary benefits of improving water quality from good to excellent, for example, would be beneficial to the water industry. Willingness to pay surveys show customers valued improvements in the quality of coastal bathing water over river waters used for bathing. It is recommended that Northumbrian Water, therefore, use a catchment-based approach, carrying out MST on a catchment basis, when investigating the sources of reduced bathing water quality. This will allow the simultaneous identification of sources impacting river water quality.

Chapter 8 General discussion

The overall aim of this research was to evaluate the performance of two emerging MST techniques, *E. coli* biomarkers and community analysis, for use by the UK water industry, and assess the feasibility of their incorporation into workflows for Northumbrian Water Ltd.

8.1 *E.coli* biomarkers

The Hu100 biomarker, developed using a database approach (Chapter 4), is the best *E.coli* marker for use in the North East of England since, although the absolute mean abundance was not significantly different to other markers, it most often had the highest average sensitivity$_{\text{isolate}}$. The similarity in absolute abundance between the H8, Hu100, and H24 *E.coli* markers and the HF183 marker reflects previous findings (Hughes, *et al*., 2017), although, the low proportion of *E.coli* containing H8 (8.25±2.68%, Table 4.7) does not support suspicions that the high abundance is due to cross-reactivity with similar sequences found in *Yersinia* and *Klebsiella* spp.. The large variability in the abundance of markers (Figure 4.3) and proportion of *E.coli* containing a marker (Table 4.7) suggest that there is no single, ideal marker for global or national use; instead markers should be evaluated before use in a catchment study. This variability is likely to be exaggerated across small, decentralised WWTPs, which were targeted in this study. Nevertheless, this reflects problems likely to be encountered in catchment studies, where small communities contribute to WWTPs and contaminate a water body; i.e., septic tanks or overflows from CSOs serving sewage from small populations. The large variability is likely to be the reason why the Hu100 marker was around an order of magnitude less abundant than the HF183 marker in the Seaton Sluice catchment study (Chapter 4). The variability in the sensitivity$_{\text{isolate}}$ of the H8 marker, which has been observed at an international scale (Table 3.1), is also evident on a local scale (Table 4.7), suggesting that all markers will exhibit a similar variation wherever they are used. Indeed, further studies indicate some global variation through the use of the Hu100 and H8 markers in Tanzania, Thailand and Nepal (Acharya *et al., In prep*; Mrozik *et al., 2019*). Local variability is problematic for the water industry if MST is to be used as a quantitative tool to prioritise catchments or areas for investment. In the Morland catchment, the first time that the H8, H12, H14 and H24 markers (Gomi *et al*., 2014) were used in the UK, the use of the H8 marker alone would have resulted in a large number (28/31) of false negative results, which meant that conclusions relied on the H24 and H14 markers, with the lowest specificity (93%) and the highest sensitivities.

The interrogation of a database, built with 263 *E.coli* genomes (Chapter 4), supported the order of marker sensitivities determined *in vitro* (Chapter 3), namely, H24 > H14 > H8 > H12; and also suggests that the trade-off in sensitivity and specificity noted in Chapter 3 is inherent to CDS within the *E.coli* genome (Figure 4.2). Interestingly, phage infecting *Enterococcus* and *Bacteroides* hosts exhibit similar performance characteristics to those observed in *E.coli* CDS, with an inverse relationship between the specificity and sensitivity (Purnell, 2011) and geographical variability in performance (Payan *et al*., 2005). Interrogation of the *E.coli* database and the trade-off between sensitivity and specificity suggest that while other human-associated CDS are present in *E.coli* genomes, there are unlikely to be any significantly better MST markers; Hu100 had the highest sensitivity (among CDS with a specificity >95%), and sensitivity, rather than specificity, has been noted as limiting the efficacy of MST markers (Chapter 3) or organisms (Purnell, 2011) where this sensitivity-specificity trade-off exists.

The low sensitivity$_{\text{isolate}}$ of *E.coli* markers (3 – 50%, Table 4.7) limits approaches which can be used for their detection. Using qPCR to quantify the abundance of marker genes from DNA extracted from environmental samples is, therefore, preferable to culture-based techniques, although, molecular methods such as qPCR also have limitations. Using current regulatory approaches, which are culture-based, to identify *E.coli* biomarkers, such as picking cultured isolates for PCR or qPCR (Chapter 3) could lead to false negative results unless a range of biomarkers are used since the sensitivity of all four biomarkers was 69%. Although the method used in Chapter 3, culturing and picking *E.coli* isolates before qPCR, sped up the process slightly, it still proved very expensive (Chapter 7), and would probably be unfeasible for use in a regulatory context due to the labour required (Porter, 2016). For the water industry, qPCR detection of FIOs is unlikely to be included in the next Bathing Water Directive following advice from the WHO (WHO, 2018). This mismatch of techniques used for MST and regulatory monitoring may be a source of concern for the water industry; and the WHO reported concerns about discrepancies between faecal indicator counts determined with culture-based and molecular techniques (WHO, 2018). These discrepancies are likely to be due to the difference in degradation rates of culturable organisms and DNA in environmental waters (Brown and Boehm 2015; Brooks and Field 2016); however, they may also be due to a lack of specificity or the multi-copy nature of the common gene targets (e.g., the 23S rRNA gene), leading to an overestimation by molecular methods (Chern *et al*., 2011). Nonetheless, several studies have observed a good relationship between culturable and molecular methods to quantify FIO (Noble and Weisberg, 2005; Oliver *et al.*, 2016; Hassard *et al.*, 2017). In my study, there was a good relationship (Spearman's ρ = 0.846, p = <2.2x10^{-16}) between the *RodA* gene and culturable *E.coli,* which, can increase confidence in the link between conclusions drawn from MST assays quantifying DNA targets, such as qPCR and sequencing, and sources of FIOs. Future studies are advised to use correlations between the concentrations of FIO determined through qPCR and regulatory techniques to improve confidence and acceptance of MST results. This is especially pertinent when the link between MST conclusions and culturable FIO is paramount, such as under the BWD (2006/7/EC) or epidemiological studies.

8.2 Community-based MST

Overall, this work supports the use of community analysis as an MST technique, particularly for use in the UK water industry. Expected sources were consistently identified and differentiated from other sources. Faecal taxon libraries (FTLs) form the basis of community analysis with SourceTracker; assessing the ability of an FTL to detect and differentiate sources of pollution is paramount to improve confidence in conclusions drawn from community analysis, particularly where investment and management decisions may be based on these conclusions. Simulating microbial communities allowed the effect of including, excluding and combining sources within the FTL to be evaluated. Building an FTL for the North East of England using 14 sewage samples seems reasonable; however, caution is necessary when using the FTL to detect sewage. The dissimilarity of microbial communities from potentially similar sources, such as sewage (Figure 5.10), could lead to a substantial underestimation in the predicted levels of contamination. Previous suggestions to combine similar sources (Staley *et al*., 2018) should be approached with caution; while combining similar sources can reduce the likelihood of false positive results, it may also result in an underestimation of some sources (e.g., sheep sources, Figure 5.8). Assessment of an FTL is vital when comparing sources; the differences observed in predicted values of different sources (Figure 5.10) should be a concern where investment and management decisions are based on findings. For example, a 25% predicted contribution of cow and dog sources, using the FTL in this study, would actually equate to an expected contribution of ~23% and 45% of dog and cow sources, respectively. Future MST studies are, therefore, advised to use similar methods to those in this study (Chapter 5), to evaluate the impact of their FTL on comparisons between sources. The importance of using local sources of pollution for FTL construction suggests that a change in the approach of MST researchers is necessary when using community analysis methods. While the traditional approach to library dependent MST, where a bigger library size is better, has been transposed to community-based MST (Brown *et al*. 2017; Staley *et al*. 2018), this approach does not necessarily translate to community-based MST. A larger library should, instead, be seen as only necessarily where samples of the actual sources of pollution cannot be obtained.

8.3 Comparison of biomarkers and community-based MST

The catchment studies presented here (Chapter 3 and Chapter 6) are currently two of only three studies which have used human-associated markers in conjunction with community analysis and, while more case studies are required, the good agreement between community analysis and biomarker MST results is encouraging. Moreover, the high sensitivity and specificity of community analysis techniques (Chapter5) were essential in confirming the presence of human pollution, especially where conclusions would otherwise rely on markers with a lower specificity to human hosts (e.g., the reliance on the H14 and H24 markers in Chapter 3) that would otherwise miss such pollution.

Community-based MST is limited to reporting the relative contribution of sources to the microbial community which is a limitation for studies assessing health risk, through QMRA, for example. Nevertheless, the good correlations between marker and community-based MST (Table 7.4) support the use of community-based MST as a decision support tool.

There was a reasonable agreement between community analysis and *E.coli* biomarkers. When human sources dominated, and markers were detected through qPCR, a good and significant relationship was observed ($\rho = 0.467$, $p < 2.2 \times 10^{-16}$, Table 7.4); where non-human sources were dominant, and marker detection involved a culturing step, the relationship was, only marginally, non-significant (ρ = 0.32, p = 0.0577, Appendix B.4). The non-significant relationship in the Morland catchment is also likely to be influenced by the differential decay rates of culturable *E.coli* and DNA in the environment (Brown & Boehm, 2015; Wanjugi *et al*., 2016; Korajkic *et al*., 2014), as biomarkers were identified from cultured *E.coli* isolates (Table 3.4). The higher sensitivity of community analysis compared to other markers observed in both catchment studies (Chapter 3 and Chapter 6) is particularly useful for the UK water industry who are likely to be dealing with increasingly lower levels of pollution as environmental standards are tightened in the future. Moreover, this improved sensitivity is essential for identifying urban diffuse pollution which can be challenging to detect.

8.4 Urban diffuse pollution

The ubiquity of urban diffuse pollution in both catchment studies (Chapters 3 and 6) should be a concern for the water industry, environmentalists, and policymakers. Urban pollution was observed in all samples taken in a largely rural catchment (Morland, Appendix B.2, Figure B2.3), and was the dominant source of pollution (Figure 7.6 and Figure 7.9) in a semi-rural catchment where urban areas accounted for only 2.5% of land use (Table 6.1). In the Seaton Sluice catchment, this pollution was attributed to sewer misconnections, since combined sewer overflows (CSOs) are monitored in the catchment, and observations during sampling did not reveal any problems with the CSOs; however, the widespread sewage pollution could also be attributed to leaky sewers, or unknown and poor performing septic tanks. Identifying and prioritising the search area for locating pollution sources, particularly diffuse pollution, is valuable (Ellis & Butler, 2015), not least to the water industry - misconnections are estimated to cost the water industry £235 million each year (Royal Haskoning, 2007). Modelling efforts suggest that elimination of misconnections from toilets and reducing misconnections from other appliances to less than 2% may be enough to improve the biological oxygen demand and ammonia elements of water quality, although, phosphorus would likely still remain a problem unless all misconnections could be eliminated (Ellis and Butler, 2015). However, this requires identification of priority areas where estimates suggest that up to one in five UK properties have misconnections (Environment Agency, 2007). Similar difficulties exist in estimating the contribution of phosphorus to freshwater resources from septic tanks. A study by Natural England (May *et al.*, 2015) notes that while a non-negligible proportion of phosphorus in catchments in England arise from septic tanks, efforts to estimate the actual phosphorus contribution from septic tanks are hampered due to a lack of information on their number and location.

The ubiquity of urban diffuse pollution observed in these studies suggest that water companies and catchment and beach managers need to be conscious of urban diffuse sources when planning water quality improvement initiatives. Some studies have suggested that predictive modelling may be beneficial to manage and improve bathing water quality (Oliver *et al.*, 2016). While statistical modelling using long-term data to predict bathing water quality may be useful, the efficacy of physical-based models to prioritise and predict sources of pollution could be undermined by urban diffuse pollution sources since they are unpredictable, are difficult to map, and there is little data available on them.

8.5 Implementation of MST into Northumbrian Water

Currently, MST is underused in the UK water industry, potentially due to the difficulty in formulating a business case for its use. This difficulty stems from the fact that:

- MST does not directly result in improvements to water quality; rather, it directs future projects;
- The value of MST often comes from not investing money where it is not needed;
- The value of improvements to water quality is difficult to determine, particularly over short-term periods and in monetary terms;
- There are no studies which have attempted to value the contribution of MST to water quality improvements directly.

Using economic valuations or pseudo-monetary valuation, such as perceived willingness-to-pay or ecosystem services, may be useful for determining the potential value of, and building a business case for MST. For example, Southern Water could justify spending £14,000 annually on improving each km of river to a high standard in their area, according to a willingness-to-pay survey. However, using pseudo-monetary valuations does not overcome other factors limiting the use of MST. A way forward for water companies may be to set a percentage of the estimated value of the project as a cut-off on which to decide whether to use MST or not.

Using both qPCR and sequencing is recommended due to the limitations of using each technique alone for MST studies, and the relatively small increase in cost in using qPCR in addition to sequencing. Three options for the use of MST by Northumbrian Water were explored. Processing >3000 samples (10 – 100 MST projects depending on scale) for qPCR and sequencing would make investing the ~£230k in capital costs to bring all MST techniques in-house the best option. While this is a large number of samples, the business risk remains small, and bringing these technologies in-house has several benefits, such as encouraging innovation by opening the use of these technologies to other areas of the business, development of laboratory staff and commercial opportunities.

The areas where MST is likely to be most valuable to the UK water industry are in reducing pollution loads to drinking water abstraction points and identifying the correct pressures on a waterbody for the development of river basin management plans.

Chapter 9 Conclusions, recommendations and future work

9.1 Conclusions

- Community-based MST, with high assay sensitivities and specificities, is more useful than marker-based MST to survey a waterbody or catchment for a range of potential pollution sources and shows potential for use in the UK water industry. However, since community-based MST only reports relative abundances, marker-based MST is still necessary for further analysis such as QMRA and catchment modelling.

- *E.coli* biomarkers may help relate MST conclusions to regulatory indicators, particularly in large catchments where transit times are long; however, the large variability in the proportion of *E.coli* containing a marker between different human communities means they are likely to be no better than other more abundant MST markers, or that multiple markers are required.

- Using the culture-based techniques featured in this study, detection of human-associated biomarkers is unfeasible for the UK water industry due to the high costs of labour involved.

- An FTL for the North East gave robust and accurate results for the detection of sewage in this region; however, there is a need to evaluate any bias in FTLs for individual studies to predict the effects of including particular sources, or combining sources with similar microbial communities.

- To incorporate MST techniques using qPCR and sequencing into NWL workflows, processing more than ~3000 samples (10 – 100 MST projects depending on the scale) would make investing the ~£230k in capital costs to bring all MST techniques in-house the best option.

- Urban diffuse pollution is ubiquitous, even in rural and semi-rural catchments and tackling this pollution should be a priority for water quality stakeholders.

9.2 Recommendations

- The UK water industry should adopt community-based MST to be able to rapidly identify a range of pollution sources; although the use of both marker and community based MST, where possible, is recommended to improve confidence in results.
- Northumbrian Water should invest £230k in capital to bring sequencing and qPCR technologies in-house is the best option for NWL to incorporate MST into their workflows. While a large number of samples (>3000) is required to achieve a return on investment, the business risk remains small, and bringing these technologies in-house has several benefits such as encouraging innovation by opening these technologies to other areas of the business, development of laboratory staff and commercial opportunities.

- If results from marker-based MST are to be used to prioritise catchments or regions for investment, the absolute abundance should be used and, ideally, the relative abundance of markers in sewage sources in each catchment should be determined.

- Using faecal taxon library developed here is recommended for the detection of sewage in the North East of England.

- While misconnections are the responsibility of households, it should be a priority for the UK water industry since it is an underlying and wide-spread source of pollution leading to deteriorated water quality.

9.3 Further work

- The use of the H8 and Hu100 markers in different geographic regions of the world (North East, Thailand, Tanzania and Nepal) provide some evidence for the more global nature of the localised variability observed in chapter 3. While ‘global’ style meta-analyses are not particularly useful for policymakers (e.g., the lack of use of evidence of epidemiological studies in Europe leading to contributed use of culture-based regulatory assays), a Europe or UK-wide evaluation of the variability in marker abundance may be useful in evaluating the potential of biomarkers to fulfil a regulatory role.

- The trade-off between sensitivity and specificity and the low sensitivity of host-associated CDS is interesting. Determining the functions of some of the host-associated CDSs would be an interesting study to determine if these CDS are not advantageous to *E.coli* isolates, and their apparent distribution in isolates from a single host is due to chance.

- The *E.coli* database suggests that there is unlikely to be any significantly better human-associated markers targeting coding sequences that are specific to the *E.coli* genome. Further work to identify host-specific markers should, therefore, focus on intergenic regions or SNP patterns in *E.coli*, or apply a similar, database approach (chapter 4) to enterococci populations. An interesting future study would be a genome-wide comparison between *Enterococcus* or *Bacteroides* isolates which are targeted by host-specific phage and those which are not; this could support the theory that both host-specific and cosmopolitan members of a species exist.

- During this work, a range of data was collected for the Seaton Sluice catchment, which would facilitate hydrological modelling. Catchment modelling of pollution to further isolate potential sources of pollution, and quantify the faecal loading at each sample point would be valuable for future decision-making. Work to determine the degradation rate of FIO, human markers and faecal microbial communities, especially in the UK, would be valuable for this modelling.

References

Accent (2013) *Southern Water Customer Engagement (Economic) - Willingness to Pay*.

Acharya, K., Khanal, S., Pantha, K., Amatya, N., Davenport, R.J. & Werner, D. (n.d.) A comparative assessment of conventional and molecular methods, including nanopore sequencing, for surveying water quality, and their application in the Kathmandu Valley, Nepal. *In prep*.

Ahmed, W., Harwood, V.J., Gyawali, P., Sidhu, J.P.S. & Toze, S. (2015) Comparison of concentration methods for quantitative detection of sewage-associated viral markers in environmental waters. *Applied and Environmental Microbiology*. 81 (6), 2042–2049.

Ahmed, W., Payyappat, S., Cassidy, M., Besley, C. & Power, K. (2018) Novel crAssphage marker genes ascertain sewage pollution in a recreational lake receiving urban stormwater runoff. *Water Research*. 145769–778.

Ahmed, W., Sidhu, J.P.S. & Toze, S. (2012) Evaluation of the nifH gene marker of methanobrevibacter smithii for the detection of sewage pollution in environmental waters in southeast Queensland, Australia. *Environmental Science and Technology*. 46 (1), 543–550.

Ahmed, W., Staley, C., Sadowsky, M.J., Gyawali, P., Sidhu, J., Palmer, A., Beale, D.J. & Toze, S. (2015) Toolbox approaches using molecular markers and 16S rRNA gene amplicon data sets for identification of fecal pollution in surface water. *Applied and Environmental Microbiology*. 81 (20), 7067–7077.

Aitken, M., Merrilees, D.. & Duncan, A. (2001) *Impact of Agricultural*

Practices and Catchment Characteristics on Ayrshire Bathing Waters.

Albertsen, M., Karst, S.M., Ziegler, A.S., Kirkegaard, R.H. & Nielsen, P.H. (2015) Back to basics - The influence of DNA extraction and primer choice on phylogenetic analysis of activated sludge communities. *PLoS ONE*. 10 (7), .

Amir, A., McDonald, D., Navas-Molina, J.A., Kopylova, E., Morton, J.T., Zech Xu, Z., Kightley, E.P., Thompson, L.R., Hyde, E.R., Gonzalez, A. & Knight, R. (2017) Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns Jack A Gilbert (ed.). *mSystems*. 2 (2), .

Anderson, K.L., Whitlock, J.E. & Harwood, V.J. (2005a) Persistence and differential survival of fecal indicator bacteria in subtropical waters and sediments. *Applied and Environmental Microbiology*. 71 (6), 3041–3048.

Anderson, K.L., Whitlock, J.E. & Harwood, V.J. (2005b) Persistence and differential survival of fecal indicator bacteria in subtropical waters and sediments. *Applied and Environmental Microbiology*. 71 (6), 3041–3048.

Ando, H., Kitao, T., Miyoshi-Akiyama, T., Kato, S., Mori, T. & Kirikae, T. (2011) Downregulation of katG expression is associated with isoniazid resistance in Mycobacterium tuberculosis. *Molecular Microbiology*. 79 (6), 1615–1628.

Araújo, S., Henriques, I.S., Leandro, S.M., Alves, A., Pereira, A. & Correia, A. (2014) Gulls identified as major source of fecal pollution in coastal waters: A microbial source tracking study. *Science of the Total Environment*. 470–47184–91.

Arber, W. (2000) Genetic variation: Molecular mechanisms and impact on microbial evolution. *FEMS Microbiology Reviews*. 24 (1), 1–7.

Ashbolt, N., O. K. Grabow, W. & Snozzi, M. (2001) *Indicators of microbial water quality*.

Van Asperen, I.A., Medema, G., Borgdorff, M.W., Sprenger, M.J.W. & Havelaar, A.H. (1998) Risk of gastroenteritis among triathletes in relation to faecal pollution of fresh waters. *International Journal of Epidemiology*. 27 (2), 309–315.

Auguie, B. (2017) *gridExtra: Miscellaneous Functions for 'Grid' Graphics*. [Online] [online]. Available from: https://cran.r-project.org/package=gridExtra.

Aw, T.G., Howe, A. & Rose, J.B. (2014) Metagenomic approaches for direct and cell culture evaluation of the virological quality of wastewater. *Journal of Virological Methods*. 21015–21.

Badgley, B.D., Ferguson, J., Heuvel, A. V, Kleinheinz, G.T., McDermott, C.M., Sandrin, T.R., Kinzelman, J., Junion, E.A., Byappanahalli, M.N., Whitman, R.L. & Sadowsky, M.J. (2011) Multi-scale temporal and spatial variation in genotypic composition of Cladophora-borne Escherichia coli populations in Lake Michigan. *Water Research*. 45 (2), 721–731.

Badgley, B.D., Thomas, F.I.M. & Harwood, V.J. (2011) Quantifying environmental reservoirs of fecal indicator bacteria associated with sediment and submerged aquatic vegetation. *Environmental microbiology*. 13 (4), 932–942.

Bailey, J.K., Pinyon, J.L., Anantham, S. & Hall, R.M. (2010) Distribution of human commensal Escherichia coli phylogenetic groups. *Journal of Clinical Microbiology*. 48 (9), 3455–3456.

Baker-Austin, C., Rangdale, R., Lowther, J. & Lees, D.N. (2010) Application of mitochondrial DNA analysis for microbial source tracking purposes in shellfish harvesting waters. Water Science and Technology 61 p.1–7.

Ballesté, E. & Blanch, A.R. (2011) Bifidobacterial diversity and the development of new microbial source tracking indicators. *Applied and Environmental Microbiology*. 77 (10), 3518–3525.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A. V, Sirotkin, A. V, Vyahhi, N., Tesler, G., Alekseyev, M.A. & Pevzner, P.A. (2012) SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*. 19 (5), 455–477.

Batista, A.M.M., Meynet, P., Garcia, G.P.P., Costa, S.A. V, Araujo, J.C., Davenport, R.J., Werner, D. & Mota Filho, C.R. (2018) Microbiological safety of a small water distribution system: Evaluating potentially pathogenic bacteria using advanced sequencing techniques. *Water Science and Technology: Water Supply*. 18 (2), 391–398.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Sayers, E.W. (2013) GenBank. *Nucleic Acids Research*. 41 (D1), D36–D42.

Bergthorsson, U. & Ochman, H. (1998) Distribution of chromosome length variation in natural isolates of Escherichia coli. *Molecular Biology and Evolution*. 15 (1), 6–16.

Bernhard, A.E. & Field, K.G. (2000) Identification of nonpoint sources of fecal pollution in coastal waters by using host-specific 16S ribosomal DNA genetic markers from fecal anaerobes. *Applied and Environmental Microbiology*. 66 (4), 1587–1594.

Beversdorf, L.J., Bornstein-Forst, S.M. & McLellan, S.L. (2007) The potential for beach sand to serve as a reservoir for Escherichia coli and the physical influences on cell die-off. *Journal of Applied Microbiology*. 102 (5), 1372–1381.

Biocentre, T. (2018) *TATAA Biocentre*. [Online] [online]. Available from: http://www.tataa.com/ (Accessed 1 March 2019).

Blanch, A.R., Belanche-Muñoz, L., Bonjoch, X., Ebdon, J., Gantzer, C., Lucena, F., Ottoson, J., Kourtis, C., Iversen, A., Kühn, I., Mocé, L., Muniesa, M., Schwartzbrod, J., Skraber, S., Papageorgiou, G.T., Taylor, H., Wallis, J. & Jofre, J. (2006) Integrated Analysis of Established and Novel Microbial and Chemical Methods for Microbial Source Tracking. *Applied and Environmental Microbiology*. 72 (9), 5915 LP-5926.

Boehm, A.B., Grant, S.B., Kim, J.H., Mowbray, S.L., McGee, C.D., Clark, C.D., Foley, D.M. & Wellman, D.E. (2002) Decadal and shorter period variability of surf zone water quality at Huntington Beach, California. *Environmental science & technology*. 36 (18), 3885–3892.

Boehm, A.B., Van De Werfhorst, L.C., Griffith, J.F., Holden, P.A., Jay, J.A., Shanks, O.C., Wang, D. & Weisberg, S.B. (2013) Performance of forty-one microbial source tracking methods: A twenty-seven lab evaluation study. *Water Research*. 47 (18), 6812–6828.

Bolger, A.M., Lohse, M. & Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 30 (15), 2114–2120.

Bonjoch, X., Ballesté, E. & Blanch, A.R. (2004) Multiplex PCR with 16S rRNA gene-targeted primers of Bifidobacterium spp. to identify sources of fecal pollution. *Applied and Environmental Microbiology*. 70 (5), 3171–3175.

Box, G.E.P. & Cox, D.R. (1964) An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*. 26 (2), 211–252.

Broeders, S., Huber, I., Grohmann, L., Berben, G., Taverniers, I., Mazzara, M., Roosens, N. & Morisset, D. (2014) Guidelines for validation of qualitative real-time PCR methods. *Trends in Food Science & Technology*. 37 (2), 115–126.

Brooks, L.E. & Field, K.G. (2016) Bayesian meta-analysis to synthesize decay rate constant estimates for common fecal indicator bacteria. *Water Research*. 104262–271.

Brown, C.M., Mathai, P.P., Loesekann, T., Staley, C. & Sadowsky, M.J. (2018) Influence of Library Composition on SourceTracker Predictions for Community-Based Microbial Source Tracking. Environmental Science and Technology

Brown, C.M., Staley, C., Wang, P., Dalzell, B., Chun, C.L. & Sadowsky, M.J. (2017) A High-Throughput DNA-Sequencing Approach for Determining Sources of Fecal Bacteria in a Lake Superior Estuary. *Environmental Science and Technology*. 51 (15), 8263–8271.

Brown, K.I. & Boehm, A.B. (2015) Comparative decay of Catellicoccus marimmalium and enterococci in beach sand and seawater. *Water Research*. 83377–384.

Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M.W., Shipley, G.L., Vandesompele, J. & Wittwer, C.T. (2009) The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clinical chemistry*. 55 (4), 611–622.

Byappanahalli, M.N., Nevers, M.B., Korajkic, A., Staley, Z.R. & Harwood, V.J. (2012) Enterococci in the Environment. *Microbiology and Molecular Biology Reviews : MMBR*. 76 (4), 685–706.

Byappanahalli, M.N., Shively, D.A., Nevers, M.B., Sadowsky, M.J. & Whitman, R.L. (2003) Growth and survival of Escherichia coli and enterococci populations in the macro-alga Cladophora (Chlorophyta). *FEMS Microbiology Ecology*. 46 (2), 203–211.

Byappanahalli, M.N., Yan, T., Hamilton, M.J., Ishii, S., Fujioka, R.S., Whitman, R.L. & Sadowsky, M.J. (2012) The population structure of Escherichia coli isolated from subtropical and temperate soils. *Science of the Total Environment*. 417–418273–279.

Caldwell, J.M., Raley, M.E. & Levine, J.F. (2007) Mitochondrial multiplex real-time PCR as a source tracking method in fecal-contaminated effluents. *Environmental Science and Technology*. 41 (9), 3277–3283.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A. & Holmes, S.P. (2016) DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*. 13581.

Cantwell, M.G., Katz, D.R., Sullivan, J.C., Borci, T. & Chen, R.F. (2016) Caffeine in Boston Harbor past and present, assessing its utility as a tracer of wastewater contamination in an urban estuary. *Marine Pollution Bulletin*. 108 (1–2), 321–324.

Cao, Y., Raith, M.R. & Griffith, J.F. (2015) Droplet digital PCR for simultaneous quantification of general and human-associated fecal indicators for water quality assessment. *Water Research*. 70337–349.

Cao, Y., Sivaganesan, M., Kelty, C.A., Wang, D., Boehm, A.B., Griffith, J.F., Weisberg, S.B. & Shanks, O.C. (2018) A human fecal contamination score for ranking recreational sites using the HF183/BacR287 quantitative real-time PCR method. *Water Research*. 128148–156.

Cao, Y., Van De Werfhorst, L.C., Dubinsky, E.A., Badgley, B.D., Sadowsky, M.J., Andersen, G.L., Griffith, J.F. & Holden, P.A. (2013) Evaluation of molecular community analysis methods for discerning fecal sources and human waste. *Water Research*. 47 (18), 6862–6872.

Caporaso, G.. (2018) *Qiime2*. [Online] [online]. Available from: https://qiime2.org/.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pẽa, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., *et al*. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*. 7 (5), 335–336.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J., Fierer, N. & Knight, R. (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences*. 108 (Supplement 1), 4516 LP-4522.

Carlos, C., Pires, M.M., Stoppe, N.C., Hachich, E.M., Sato, M.I.Z., Gomes, T.A.T., Amaral, L.A. & Ottoboni, L.M.M. (2010) Escherichia coli phylogenetic group determination and its application in the identification of the major animal source of fecal contamination. *BMC Microbiology*. 10.

Cersosimo, L.M., Bainbridge, M.L., Kraft, J. & Wright, A.-D.G. (2016) Influence of periparturient and postpartum diets on rumen methanogen communities in three breeds of primiparous dairy cows. *BMC Microbiology*. 16 (1), 78.

Chakravorty, S., Helb, D., Burday, M., Connell, N. & Alland, D. (2007) A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of Microbiological Methods*. 69 (2), 330–339.

Chern, E.C., Siefring, S., Paar, J., Doolittle, M. & Haugland, R.A. (2011) Comparison of quantitative PCR assays for Escherichia coli targeting ribosomal RNA and single copy genes. *Letters in Applied Microbiology*. 52 (3), 298–306.

Clermont, O., Lescat, M., O'Brien, C.L., Gordon, D.M., Tenaillon, O. & Denamur, E. (2008) Evidence for a human-specific Escherichia coli clone. *Environmental Microbiology*. 10 (4), 1000–1006.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*. 2nd Editio. New York: Routledge.

Craig, D.L., Fallowfield, H.J. & Cromar, N.J. (2004) Use of microcosms to determine persistence of Escherichia coli in recreational coastal water and sediment and validation with in situ measurements. *Journal of Applied Microbiology*. 96 (5), 922–930.

DEFRA (2018a) *Agriculture in the United Kingdon*.

DEFRA (2018b) *Farming statistics: livestock populations at 1 December 2017 – England*. [Online] [online]. Available from: https://www.gov.uk/government/statistics/farming-statistics-livestock-populations-at-1-december-2017-england (Accessed 6 October 2019).

DEFRA (2011) *Water for Life*.

Deng, D., Zhang, N., Mustapha, A., Xu, D., Wuliji, T., Farley, M., Yang, J., Hua, B., Liu, F. & Zheng, G. (2014) Differentiating enteric Escherichia coli from environmental bacteria through the putative glucosyltransferase gene (ycjM). *Water Research*. 61224–231.

Deng, D., Zhang, N., Xu, D., Reed, M., Liu, F. & Zheng, G. (2015) Polymorphism of the glucosyltransferase gene (ycjM) in Escherichia coli and its use for tracking human fecal pollution in water. *Science of the Total Environment*. 537260–267.

Denton, H. (2017) 'Mental Wellbeing and Open water Swimming', in *UK Bathing Water Conference 2017*. [Online]. 2017 p.

Ding, T., Suo, Y., Xiang, Q., Zhao, X., Chen, S., Ye, X. & Liu, D. (2017) Significance of Viable but Nonculturable Escherichia coli: Induction, Detection, and Control. *Journal of microbiology and biotechnology*. 27 (3), 417–428.

Diston, D. & Wicki, M. (2015) Occurrence of bacteriophages infecting Bacteroides host strains (ARABA 84 and GB-124) in fecal samples of human and animal origin. *Journal of water and health*. 13 (3), 654–661.

Dombek, P.E., Johnson, L.K., Zimmerley, S.T. & Sadowsky, M.J. (2000) Use of repetitive DNA sequences and the PCR To differentiate

Escherichia coli isolates from human and animal sources. *Applied and environmental microbiology*. 66 (6), 2572–2577.

Dufour, Al, Wade, Timothy, Kay, D. (2012) 'Epidemiological studies on swimmer health effects associated with potential exposure to zoonotic pathogens in bathing beach water – a review', in *Animal Waste, Water Quality and human Health*. 1st edition [Online]. Genevre: World Health Organisation. p.

Dutilh, B.E., Cassman, N., McNair, K., Sanchez, S.E., Silva, G.G.Z., Boling, L., Barr, J.J., Speth, D.R., Seguritan, V., Aziz, R.K., Felts, B., Dinsdale, E.A., Mokili, J.L. & Edwards, R.A. (2014) A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nature Communications*. 5.

Dwight, R.H., Fernandez, L.M., Baker, D.B., Semenza, J.C. & Olson, B.H. (2005) Estimating the economic burden from illnesses associated with recreational coastal water pollution - A case study in Orange County, California. *Journal of Environmental Management*. 76 (2), 95–103.

Dymond, J.R., Serezat, D., Ausseil, A.-G.E. & Muirhead, R.W. (2016) Mapping of Escherichia coli Sources Connected to Waterways in the Ruamahanga Catchment, New Zealand. *Environmental science & technology*. 50 (4), 1897–1905.

Ebdon, J. & Taylor, H. (2006) Geographical Stability of Enterococcal Antibiotic Resistance Profiles in Europe and Its Implications for the Identification of Fecal Sources. *Environmental science & technology*. 405327–5332.

Ebdon, J.E., Sellwood, J., Shore, J. & Taylor, H.D. (2012) Phages of Bacteroides (GB-124): A Novel Tool for Viral Waterborne Disease

Control? *Environmental Science & Technology*. 46 (2), 1163–1169.

Edberg, S.C., Rice, E.W., Karlin, R.J. & Allen, M.J. (2000) Escherichia coli: the best biological drinking water indicator for public health protection. *Symposium series (Society for Applied Microbiology)*. (29), 106S–116S.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 26 (19), 2460–2461.

EEA (2018a) *European Environment Agency 2018 Water Assessment*.

EEA (2018b) *European waters getting cleaner, but big challenges remain*.

Ellis, J.B. & Butler, D. (2015) Surface water sewer misconnections in England and Wales: Pollution sources and impacts. *Science of the Total Environment*. 52698–109.

Elmir, S.M., Wright, M.E., Abdelzaher, A., Solo-Gabriele, H.M., Fleming, L.E., Miller, G., Rybolowik, M., Peter Shih, M.-T., Pillai, S.P., Cooper, J.A. & Quaye, E.A. (2007) Quantitative evaluation of bacteria released by bathers in a marine water. *Water Research*. 41 (1), 3–10.

Environment Agency (2018a) *Catchment Data Explorer: Seaton Burn source to tidal limit*. [Online] [online]. Available from: https://environment.data.gov.uk/catchment-planning/WaterBody/GB103022076190 (Accessed 10 January 2018).

Environment Agency (2018b) *National Laboratory Service: Service*. [Online] [online]. Available from: http://natlabs.co.uk/our-services/ (Accessed 28 December 2018).

Environment Agency (2007) *The Unseen Threat to Water Quality*.

Environment Agency (2016) *Water for Life and Livelihoods Part 1:*

Northumbria River Basin District River Basin Management Plan.

European, E.A. (2018) *Diffuse sources*. [Online] [online]. Available from: https://www.eea.europa.eu/archived/archived-content-water-topic/water-pollution/diffuse-sources (Accessed 9 June 2019).

Federici, S., Miragoli, F., Pisacane, V., Rebecchi, A., Morelli, L. & Callegari, M.L. (2015) Archaeal microbiota population in piglet feces shifts in response to weaning: Methanobrevibacter smithii is replaced with Methanobrevibacter boviskoreani. *FEMS Microbiology Letters*. 362 (10), fnv064-fnv064.

Feng, S., Bootsma, M. & McLellan, S.L. (2018) Novel human-associated &lt;em&gt;Lachnospiraceae&lt;/em&gt; genetic markers improve detection of fecal pollution sources in urban waters. *Applied and Environmental Microbiology*.

Fewtrell, L. & Kay, D. (2015) Recreational Water and Infection: A Review of Recent Findings. *Current Environmental Health Reports*. 2 (1), 85–94.

Field, K.G. & Samadpour, M. (2007) Fecal source tracking, the indicator paradigm, and managing water quality. *Water Research*. 41 (16), 3517–3538.

Fisher, K. & Phillips, C. (2009) The ecology, epidemiology and virulence of Enterococcus. *Microbiology*. 155 (6), 1749–1757.

Flag, B. (2014) *Blue Flag Beaches*. [Online] [online]. Available from: http://www.blueflag.global/beaches2/.

Fleisher, J.M., Kay, D., Salmon, R.L., Jones, F., Wyer, M. & Godfree, A.F. (1996) Marine waters contaminated with domestic sewage: Nonenteric illnesses associated with bather exposure in the United Kingdom. *American Journal of Public Health*. 86 (9), 1228–1234.

Flores, G.E., Bates, S.T., Knights, D., Lauber, C.L., Stombaugh, J., Knight, R. & Fierer, N. (2011) Microbial biogeography of public restroom surfaces. *PLoS ONE*. 6 (11), .

Forootan, A., Sjöback, R., Björkman, J., Sjögreen, B., Linz, L. & Kubista, M. (2017) Methods to determine limit of detection and limit of quantification in quantitative real-time PCR (qPCR). *Biomolecular Detection and Quantification*. 121–6.

Fricker, E.J., Illingworth, K.S. & Fricker, C.R. (1997) Use of two formulations of Colilert and QuantiTray® for assessment of the bacteriological quality of water. *Water Research*. 31 (10), 2495–2499.

Fujioka, R.S. & Narikawa, O.T. (1982) Effect of sunlight on enumeration of indicator bacteria under field conditions. *Applied and environmental microbiology*. 44 (2), 395–401.

García-Aljaro, C., Ballesté, E., Muniesa, M. & Jofre, J. (2017) Determination of crAssphage in water samples and applicability for tracking human faecal pollution. *Microbial Biotechnology*. 10 (6), 1775–1780.

Gawler, A.H., Beecher, J.E., Brandão, J., Carroll, N.M., Falcão, L., Gourmelon, M., Masterson, B., Nunes, B., Porter, J., Rincé, A., Rodrigues, R., Thorp, M., Martin Walters, J. & Meijer, W.G. (2007) Validation of host-specific Bacteriodales 16S rRNA genes as markers to determine the origin of faecal pollution in Atlantic Rim countries of the European Union. *Water Research*. 41 (16), 3780–3784.

Gomi, R., Matsuda, T., Matsui, Y. & Yoneda, M. (2014) Fecal source tracking in water by next-generation sequencing technologies using host-specific escherichia coli genetic markers. *Environmental Science and Technology*. 48 (16), 9616–9623.

Gooch-Moore, J., Goodwin, K.D., Dorsey, C., Ellender, R.D., Mott, J.B., Ornelas, M., Sinigalliano, C., Vincent, B., Whiting, D. & Wolfe, S.H. (2011) New USEPA water quality criteria by 2012: GOMA concerns and recommendations. *Journal of water and health*. 9 (4), 718–733.

Goodwin, K.D., Gruber, S., Vondrak, M. & Crumpacker, A. (2016) Watershed Assessment with Beach Microbial Source Tracking and Outcomes of Resulting Gull Management. *Environmental science & technology*. 50 (18), 9900–9906.

Gourmelon, M., Caprais, M.P., Mieszkin, S., Marti, R., Wéry, N., Jardé, E., Derrien, M., Jadas-Hécart, A., Communal, P.Y., Jaffrezic, A. & Pourcher, A.M. (2010) Development of microbial and chemical MST tools to identify the origin of the faecal pollution in bathing and shellfish harvesting waters in France. *Water Research*. 44 (16), 4812–4824.

Green, H.C., Haugland, R.A., Varma, M., Millen, H.T., Borchardt, M.A., Field, K.G., Walters, W.A., Knight, R., Sivaganesan, M., Kelty, C.A. & Shanks, O.C. (2014) Improved HF183 quantitative real-time PCR assay for characterization of human fecal pollution in ambient surface water samples. *Applied and environmental microbiology*. 80 (10), 3086–3094.

Gregory Caporaso, J. (2018) *Qiime2View*. [Online] [online]. Available from: https://view.qiime2.org/ (Accessed 21 March 2019).

Griffith, J.F., Layton, A.B., Holden, P.A., Jay, J.A., Hagedom, C., McGee, C.D. & Weisberg, S.B. (2013) *The California Microbial Source*

Identification Manual: A Tiered Approach to Identifying Fecal Pollution Sources to Beaches.

Griffith, J.F., Weisberg, S.B. & McGee, C.D. (2003) Evaluation of microbial source tracking methods using mixed fecal sources in aqueous test samples. *Journal of Water and Health*. 1 (4), 141–151.

Grubbs, F.E. (1969) Procedures for detecting outlying observations in samples. *Technometrics*. 11 (1), 1–21.

Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 29 (8), 1072–1075.

Haas, B.J., Gevers, D., Earl, A.M., Feldgarden, M., Ward, D. V., Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S.K., Sodergren, E., Methé, B., DeSantis, T.Z., Petrosino, J.F., Knight, R. & Birren, B.W. (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research*. 21 (3), 494–504.

Hagedorm, C., Blanch, A.R.H. & Harwood, V.J. (2011) *Microbial Source Tracking, Applications, and Case Studies*. New York: Springer.

Hägglund, M., Bäckman, S., Macellaro, A., Lindgren, P., Borgmästars, E., Jacobsson, K., Dryselius, R., Stenberg, P., Sjödin, A., Forsman, M. & Ahlinder, J. (2018) Accounting for bacterial overlap between raw water communities and contaminating sources improves the accuracy of signature-based microbial source tracking. *Frontiers in Microbiology*. 9 (OCT), .

Hall, V., Taye, A., Walsh, B., Maguire, H., Dave, J., Wright, A., Anderson,

C. & Crook, P. (2017) A large outbreak of gastrointestinal illness at an open-water swimming event in the River Thames, London. *Epidemiology and Infection*. 145 (6), 1246–1255.

Halliday, E. & Gast, R.J. (2011) Bacteria in beach sands: an emerging challenge in protecting coastal water quality and bather health. *Environmental science & technology*. 45 (2), 370–379.

Harada, H., Fujimori, Y., Gomi, R., Ahsan, M.N., Fujii, S., Sakai, A. & Matsuda, T. (2018) Pathotyping of Escherichia coli isolated from community toilet wastewater and stored drinking water in a slum in Bangladesh. *Lett Appl Microbiol*. (66), 542–548.

Harrell, F.. & Dupont, C. (2018) *Hmisc R Package*. [Online] [online]. Available from: https://cran.r-project.org/package=Hmisc.

Harwood, V.J., Boehm, A.B., Sassoubre, L.M., Vijayavel, K., Stewart, J.R., Fong, T.T., Caprais, M.P., Converse, R.R., Diston, D., Ebdon, J., Fuhrman, J.A., Gourmelon, M., Gentry-Shields, J., Griffith, J.F., Kashian, D.R., Noble, R.T., Taylor, H. & Wicki, M. (2013) Performance of viruses and bacteriophages for fecal source determination in a multi-laboratory, comparative study. *Water Research*. 47 (18), 6929–6943.

Harwood, V.J., Staley, C., Badgley, B.D., Borges, K. & Korajkic, A. (2014) Microbial source tracking markers for detection of fecal contamination in environmental waters: Relationships between pathogens and human health outcomes. *FEMS Microbiology Reviews*. 38 (1), 1–40.

Hass, C N, Rose, J B, Gerba, C.P. (1999) *Quantitative Microbial Risk Assessment*. 1st edition. New York: John Wiley & Sons Inc.

Hassard, F., Andrews, A., Jones, D.L., Parsons, L., Jones, V., Cox, B.A., Daldorph, P., Brett, H., McDonald, J.E. & Malham, S.K. (2017) Physicochemical factors influence the abundance and culturability of human enteric pathogens and fecal indicator organisms in estuarine water and sediment. *Frontiers in Microbiology*. 8 (OCT), .

Haugland, R.A., Siefring, S.C., Wymer, L.J., Brenner, K.P. & Dufour, A.P. (2005) Comparison of Enterococcus measurements in freshwater at two recreational beaches by quantitative polymerase chain reaction and membrane filter culture analysis. *Water Research*. 39 (4), 559–568.

Haugland, R.A., Varma, M., Sivaganesan, M., Kelty, C., Peed, L. & Shanks, O.C. (2010) Evaluation of genetic markers from the 16S rRNA gene V2 region for use in quantitative detection of selected Bacteroidales species and human fecal waste by qPCR. *Systematic and Applied Microbiology*. 33 (6), 348–357.

Havelaar, A.H., van Olphen, M. & Drost, Y.C. (1993) F-specific RNA bacteriophages are adequate model organisms for enteric viruses in fresh water. *Applied and Environmental Microbiology*. 59 (9), 2956–2962.

He, X., Chen, H., Shi, W., Cui, Y. & Zhang, X.-X. (2015) Persistence of mitochondrial DNA markers as fecal indicators in water environments. *Science of the Total Environment*. 533383–390.

He, X., Liu, P., Zheng, G., Chen, H., Shi, W., Cui, Y., Ren, H. & Zhang, X.-X. (2016) Evaluation of five microbial and four mitochondrial DNA markers for tracking human and pig fecal pollution in freshwater. *Scientific Reports*. 6.

Heaney, C.D., Myers, K., Wing, S., Hall, D., Baron, D. & Stewart, J.R. (2015) Source tracking swine fecal waste in surface water proximal to swine concentrated animal feeding operations. *Science of the Total*

Environment. 511676–683.

Hebbali, A. (2018) *olsrr: Tools for Building OLS Regression Models*. [Online]

Henry, R., Schang, C., Coutts, S., Kolotelo, P., Prosser, T., Crosbie, N., Grant, T., Cottam, D., O'Brien, P., Deletic, A. & McCarthy, D. (2016) Into the deep: Evaluation of SourceTracker for assessment of faecal contamination of coastal waters. *Water Research*. 93242–253.

Hou, Y., Zhang, H., Miranda, L. & Lin, S. (2010) Serious overestimation in quantitative pcr by circular (supercoiled) plasmid standard: Microalgal pcnaas the model gene. *PLoS ONE*. 5 (3), .

Hughes, B., Beale, D.J., Dennis, P.G., Cook, S. & Ahmed, W. (2017) Cross-comparison of human wastewater-associated molecular markers in relation to fecal indicator bacteria and enteric viruses in recreational beach waters. *Applied and Environmental Microbiology*. 83 (8), .

Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W. & Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*. 11119.

Iceton, G. & Davenport, R.J. (2018) *Next Generation Sequencing for the Water Industry*. [Online].

Ishii, S., Hansen, D.L., Hicks, R.E. & Sadowsky, M.J. (2007) Beach sand and sediments are temporal sinks and sources of Escherichia coli in lake superior. *Environmental Science and Technology*. 41 (7), 2203–2209.

Ishii, S. & Sadowsky, M.J. (2008) Escherichia coli in the environment: Implications for water quality and human health. *Microbes and Environments*. 23 (2), 101–108.

Iyengar, V., Albaugh, G.P., Lohani, A. & Nair, P.P. (1991) Human stools as a source of viable colonic epithelial cells. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*. 5 (13), 2856–2859.

Jalliffier-Verne, I., Leconte, R., Huaringa-Alvarez, U., Heniche, M., Madoux-Humery, A.S., Autixier, L., Galarneau, M., Servais, P., Prévost, M. & Dorner, S. (2016) Modelling the impacts of global change on concentrations of Escherichia coli in an urban river. *Advances in Water Resources*.

Jofre, J., Blanch, A.R., Lucena, F. & Muniesa, M. (2014) Bacteriophages infecting Bacteroides as a marker for microbial source tracking. *Water Research*. 551–11.

Johnston, C., Ufnar, J.A., Griffith, J.F., Gooch, J.A. & Stewart, J.R. (2010) A real-time qPCR assay for the detection of the nifH gene of Methanobrevibacter smithii, a potential indicator of sewage pollution. *Journal of Applied Microbiology*. 109 (6), 1946–1956.

Kataržytė, M., Mėžinė, J., Vaičiūtė, D., Liaugaudaitė, S., Mukauskaitė, K., Umgiesser, G. & Schernewski, G. (2018) Fecal contamination in shallow temperate estuarine lagoon: Source of the pollution and environmental factors. *Marine Pollution Bulletin*. 133762–772.

Katoh, K., Misawa, K., Kuma, K. & Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*. 30 (14), 3059–3066.

Kay, D., Fleisher, J.M., Salmon, R.L., Jones, F., Wyer, M.D., Godfree, A.F., Zelenauch-Jacquotte, Z. & Shore, R. (1994) Predicting likelihood of gastroenteritis from sea bathing: Results from randomised exposure.

Lancet. 344 (8927), 905–909.

Kay, D., Jones, F., Wyer, M.D., Fleisher, J.M., Salmon, R.L., Godfree, A.F., Zelenauch-Jacquotte, A. & Shore, R. (1994a) Predicting likelihood of gastroenteritis from sea bathing: results from randomised exposure. *The Lancet*. 344 (8927), 905–909.

Kay, D., Jones, F., Wyer, M.D., Fleisher, J.M., Salmon, R.L., Godfree, A.F., Zelenauch-Jacquotte, A. & Shore, R. (1994b) Predicting likelihood of gastroenteritis from sea bathing: results from randomised exposure. *The Lancet*. 344 (8927), 905–909.

Kildare, B.J., Leutenegger, C.M., McSwain, B.S., Bambic, D.G., Rajal, V.B. & Wuertz, S. (2007) 16S rRNA-based assays for quantitative detection of universal, human-, cow-, and dog-specific fecal Bacteroidales: A Bayesian approach. *Water Research*. 41 (16), 3701–3715.

King, S., Exley, J., Winpenny, E., Alves, L., Henham, M.-L. & Larkin, J. (2015) The Health Risks of Bathing in Recreational Waters: A Rapid Evidence Assessment of Water Quality and Gastrointestinal Illness. *Rand Health Quarterly*. 4 (4), 5.

Kinzelman, J.L. & McLellan, S.L. (2009) Success of science-based best management practices in reducing swimming bans—a case study from Racine, Wisconsin, USA. *Aquatic Ecosystem Health & Management*. 12 (2), 187–196.

Kitajima, M., Iker, B.C., Pepper, I.L. & Gerba, C.P. (2014) Relative abundance and treatment reduction of viruses during wastewater treatment processes--identification of potential viral indicators. *The Science of the total environment*. 488–489290–296.

Knights, D., Kuczynski, J., Charlson, E.S., Zaneveld, J., Mozer, M.C., Collman, R.G., Bushman, F.D., Knight, R. & Kelley, S.T. (2011) Bayesian community-wide culture-independent microbial source tracking. *Nature Methods*. 8 (9), 761–765.

Knights, D., Kuczynski, J., Charlson, E.S., Zaneveld, J., Mozer, M.C., Collman, R.G., Bushman, F.D., Knight, R. & Kelley, S.T. (2011) Bayesian community-wide culture-independent microbial source tracking. *Nat Meth*. 8 (9), 761–763.

Koeuth, T., Versalovic, J. & Lupski, J.R. (1995) Differential subsequence conservation of interspersed repetitive Streptococcus pneumoniae BOX elements in diverse bacteria. *Genome research*. 5 (4), 408–418.

Komsta, L. (2011) *Outliers: Tests for outliers*. [Online] [online]. Available from: https://cran.r-project.org/package=outliers (Accessed 26 February 2018).

Korajkic, A., McMinn, B.R., Shanks, O.C., Sivaganesan, M., Fout, G.S. & Ashbolt, N.J. (2014) Biotic interactions and sunlight affect persistence of fecal indicator bacteria and microbial source tracking genetic markers in the upper mississippi river. *Applied and Environmental Microbiology*. 80 (13), 3952–3961.

Kralik, P. & Ricchi, M. (2017) A Basic Guide to Real Time PCR in Microbial Diagnostics: Definitions, Parameters, and Everything. *Frontiers in Microbiology*. 8108.

Kumar, P.S., Brooker, M.R., Dowd, S.E. & Camerlengo, T. (2011) Target region selection is a critical determinant of community fingerprints generated by 16S Pyrosequencing. *PLoS ONE*. 6 (6), .

Lamparelli, C.C., Pogreba-Brown, K., Verhougstraete, M., Sato, M.I.Z., de Castro Bruni, A., Wade, T.J. & Eisenberg, J.N.S. (2015) Are fecal indicator bacteria appropriate measures of recreational water risks in the tropics: A cohort study of beach goers in Brazil? *Water Research*. 8759–68.

Lancefield, R.C. (1933) A SEROLOGICAL DIFFERENTIATION OF HUMAN AND OTHER GROUPS OF HEMOLYTIC STREPTOCOCCI. *The Journal of Experimental Medicine*. 57 (4), 571–595.

Lauber, C.L., Zhou, N., Gordon, J.I., Knight, R. & Fierer, N. (2010) Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. *FEMS Microbiology Letters*. 307 (1), 80–86.

Leecaster, M.K. & Weisberg, S.B. (2001) Effect of sampling frequency on shoreline microbiology assessments. *Marine pollution bulletin*. 42 (11), 1150–1154.

Leonard, A.F.C., Zhang, L., Balfour, A.J., Garside, R., Hawkey, P.M., Murray, A.K., Ukoumunne, O.C. & Gaze, W.H. (2018) Exposure to and colonisation by antibiotic-resistant E. coli in UK coastal water users: Environmental surveillance, exposure assessment, and epidemiological study (Beach Bum Survey). *Environment International*. 114326–333.

Lévesque, B., Brousseau, P., Bernier, F., Dewailly, É. & Joly, J. (2000) Study of the bacterial content of ring-billed gull droppings in relation to recreational water quality. *Water Research*. 34 (4), 1089–1096.

Li, J.-R. (2014) Microbial Source Tracking: A Tool for Identifying Sources of Microbial Contamination in the Food Chain AU - Fu, Ling-Lin.

Critical Reviews in Food Science and Nutrition. 54 (6), 699–707.

Lim, F.Y., Ong, S.L. & Hu, J. (2017) Recent advances in the use of chemical markers for tracing wastewater contamination in aquatic environment: A review. *Water (Switzerland)*. 9 (2), .

Lukjancenko, O., Wassenaar, T.M. & Ussery, D.W. (2010) Comparison of 61 Sequenced Escherichia coli Genomes. *Microbial Ecology*. 60 (4), 708–720.

Lundy, L. & Wade, R. (2013) *A Critical Review of Urban Diffuse Pollution Control: Methodologies to Identify Sources, Pathways and Mitigation Measures with Multiple Benefits*.

Luo, C., Walk, S.T., Gordon, D.M., Feldgarden, M., Tiedje, J.M. & Konstantinidis, K.T. (2011) Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species. *Proceedings of the National Academy of Sciences of the United States of America*. 108 (17), 7200–7205.

Marion, J.W., Lee, C., Lee, C.S., Wang, Q., Lemeshow, S., Buckley, T.J., Saif, L.J. & Lee, J. (2014) Integrating bacterial and viral water quality assessment to predict swimming-associated illness at a freshwater beach: A cohort study. *PLoS ONE*. 9 (11), .

Marion, J.W., Lee, J., Lemeshow, S. & Buckley, T.J. (2010) Association of gastrointestinal illness and recreational water exposure at an inland U.S. beach. *Water Research*. 44 (16), 4796–4804.

Martellini, A., Payment, P. & Villemur, R. (2005) Use of eukaryotic mitochondrial DNA to differentiate human, bovine, porcine and ovine sources in fecally contaminated surface water. *Water Research*. 39 (4), 541–548.

Marti, R., Gannon, V.P.J., Jokinen, C., Lanthier, M., Lapen, D.R., Neumann, N.F., Ruecker, N.J., Scott, A., Wilkes, G., Zhang, Y. & Topp, E. (2013) Quantitative multi-year elucidation of fecal sources of waterborne pathogen contamination in the South Nation River basin using Bacteroidales microbial source tracking markers. *Water Research*. 47 (7), 2315–2324.

Mattioli, M.C., Sassoubre, L.M., Russell, T.L. & Boehm, A.B. (2016) Decay of sewage-sourced microbial source tracking markers and fecal indicator bacteria in marine waters. *Water Research*.

May, L., Place, C., O'Malley, M. & Spears, B. (2015a) The impact of phosphorus inputs from small discharges on designated freshwater sites. Natural England Commissioned Reports

May, L., Place, C., O'Malley, M. & Spears, B. (2015b) *The impact of phosphorus inputs from small discharges on designated freshwater sites*.

Mayer, R.E., Reischer, G.H., Ixenmaier, S.K., Derx, J., Blaschke, A.P., Ebdon, J.E., Linke, R., Egle, L., Ahmed, W., Blanch, A.R., Byamukama, D., Savill, M., Mushi, D., Cristóbal, H.A., Edge, T.A., Schade, M.A., Aslan, A., Brooks, Y.M., Sommer, R., *et al*. (2018) Global Distribution of Human-Associated Fecal Genetic Markers in Reference Samples from Six Continents. *Environmental Science & Technology*. 52 (9), 5076–5084.

McCarthy, D.T., Jovanovic, D., Lintern, A., Teakle, I., Barnes, M., Deletic, A., Coleman, R., Rooney, G., Prosser, T., Coutts, S., Hipsey, M.R., Bruce, L.C. & Henry, R. (2017) Source tracking using microbial community fingerprints: Method comparison with hydrodynamic modelling. *Water research*. 109253–265.

McLellan, S.L. & Eren, A.M. (2014) Discovering new indicators of fecal pollution. *Trends in Microbiology*.

McLellan, S.L., Newton, R.J., Vandewalle, J.L., Shanks, O.C., Huse, S.M., Eren, A.M. & Sogin, M.L. (2013) Sewage reflects the distribution of human faecal lachnospiraceae. *Environmental Microbiology*. 15 (8), 2213–2227.

McMurdie, P.J. & Holmes, S. (2013) Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE*. 8 (4), .

McQuaig, S., Griffith, J. & Harwood, V.J. (2012) Association of fecal indicator bacteria with human viruses and microbial source tracking markers at coastal beaches impacted by nonpoint source pollution. *Applied and Environmental Microbiology*. 78 (18), 6423–6432.

McQuaig, S.M., Scott, T.M., Lukasik, J.O., Paul, J.H. & Harwood, V.J. (2009) Quantification of human polyomaviruses JC virus and BK Virus by TaqMan quantitative PCR and comparison to other water quality indicators in water and fecal samples. *Applied and Environmental Microbiology*. 75 (11), 3379–3388.

Meays, C.L., Broersma, K., Nordin, R. & Mazumder, A. (2004) Source tracking fecal bacteria in water: A critical review of current methods. *Journal of Environmental Management*. 73 (1), 71–79.

Meerburg, B.G., Koene, M.G.J. & Kleijn, D. (2011) Escherichia coli concentrations in feces of geese, coots, and gulls residing on recreational water in The Netherlands. *Vector borne and zoonotic diseases*

(Larchmont, N.Y.). 11 (6), 601–603.

Met-office (2012) *Met Office Integrated Data Archive System (MIDAS) Land and Marine Surface Stations Data (1853-current)*. [Online] [online]. Available from: http://catalogue.ceda.ac.uk/uuid/220a65615218d5c9cc9e4785a3234bd0 (Accessed 23 February 2019).

Mohapatra, B.R., Broersma, K. & Mazumder, A. (2007) Comparison of five rep-PCR genomic fingerprinting methods for differentiation of fecal Escherichia coli from humans, poultry and wild birds. *FEMS Microbiology Letters*. 277 (1), 98–106.

Mohapatra, B.R. & Mazumder, A. (2008) Comparative efficacy of five different rep-PCR methods to discriminate Escherichia coli populations in aquatic environments. Water Science and Technology 58 p.537–547.

Mrozik, W., Vinitnantharat, S., Thongsamer, T., Pansuk, N., Thayanukul, P., Acharya, K., Hazlerigg, C., Robson, A.., Davenport, R.. & Werner, D. (2019) The Food-Water Quality Nexux in Periurban Aquacultures Located Downstream of Bangkok, Thailand. *Science of the Total Environment, Volume 695.*

Muers, M. (2011) Technology: Getting Moore from DNA sequencing. *Nat Rev Genet*. 12 (9), 586–587.

Muller, T., Ulrich, A., Ott, E.M. & Muller, M. (2001) Identification of plant-associated enterococci. *Journal of applied microbiology*. 91 (2), 268–278.

Murray, B.E. (1990) The life and times of the Enterococcus. *Clinical microbiology reviews*. 3 (1), 46–65.

Napier, M.D., Haugland, R., Poole, C., Dufour, A.P., Stewart, J.R., Weber, D.J., Varma, M., Lavender, J.S. & Wade, T.J. (2017) Exposure to human-associated fecal indicators and self-reported illness among swimmers at recreational beaches: A cohort study. *Environmental Health: A Global Access Science Source*. 16 (1), .

National Laboratory Service (2018) *National Laboratory Service: Water Analysis – Microbiological Testing*. [Online] [online]. Available from: http://natlabs.co.uk/our-services/water-analysis/microbiological-testing/ (Accessed 17 December 2018).

Naziri, Z., Derakhshandeh, A., Firouzi, R., Motamedifar, M. & Shojaee Tabrizi, A. (2016) DNA fingerprinting approaches to trace Escherichia coli sharing between dogs and owners. *Journal of Applied Microbiology*. 120 (2), 460–468.

Neave, M., Luter, H., Padovan, A., Townsend, S., Schobben, X. & Gibb, K. (2014) Multiple approaches to microbial source tracking in tropical northern Australia. *MicrobiologyOpen*. 3 (6), 860–874.

Newton, R.J., Bootsma, M.J., Morrison, H.G., Sogin, M.L. & McLellan, S.L. (2013) A Microbial Signature Approach to Identify Fecal Pollution in the Waters Off an Urbanized Coast of Lake Michigan. *Microbial Ecology*. 65 (4), 1011–1023.

Newton, R.J., VandeWalle, J.L., Borchardt, M.A., Gorelick, M.H. & McLellan, S.L. (2011) Lachnospiraceae and bacteroidales alternative fecal indicators reveal chronic human sewage contamination in an Urban harbor. *Applied and Environmental Microbiology*. 77 (19), 6972–6981.

Nguyen, K.H., Senay, C., Young, S., Nayak, B., Lobos, A., Conrad, J. & Harwood, V.J. (2018) Determination of wild animal sources of fecal indicator bacteria by microbial source tracking (MST) influences regulatory decisions. *Water Research*. 144424–434.

Noble, R.T., Blackwood, A.D., Griffith, J.F., McGee, C.D. & Weisberg, S.B. (2010) Comparison of Rapid Quantitative PCR-Based and Conventional Culture-Based Methods for Enumeration of &lt;em&gt;Enterococcus&lt;/em&gt; spp. and &lt;em&gt;Escherichia coli&lt;/em&gt; in Recreational Waters. *Applied and Environmental Microbiology*. 76 (22), 7437 LP-7443.

Noble, R.T. & Weisberg, S.B. (2005) A review of technologies for rapid detection of bacteria in recreational waters. *Journal of water and health*. 3 (4), 381–392.

Nödler, K., Tsakiri, M., Aloupi, M., Gatidou, G., Stasinakis, A.S. & Licha, T. (2016) Evaluation of polar organic micropollutants as indicators for wastewater-related coastal water quality impairment. *Environmental Pollution*. 211282–290.

Nshimyimana, J.P., Cruz, M.C., Thompson, R.J. & Wuertz, S. (2017) Bacteroidales markers for microbial source tracking in Southeast Asia. *Water Research*. 118239–248.

NU-OMICS (2018) *NU-OMICS: Illumina sequencing*. [Online] [online]. Available from: https://www.northumbria.ac.uk/business-services/engage-with-us/research/nu-omics/illumina/ (Accessed 6 January 2019).

NWL (2018) *Northumbrian Water: Our Future Vision*. [Online] [online]. Available from: https://www.nwl.co.uk/your-home/our-future-vision.aspx (Accessed 17 December 2018).

O’Hara, R.B. & Kotze, D.J. (2010) Do not log-transform count data. *Methods in Ecology and Evolution*. 1 (2), 118–122.

Obiri-Danso, K. & Jones, K. (2000) Intertidal sediments as reservoirs for hippurate negative campylobacters, salmonellae and faecal indicators in three EU recognised bathing waters in North West England. *Water Research*. 34 (2), 519–527.

Odagiri, M., Schriewer, A., Hanley, K., Wuertz, S., Misra, P.R., Panigrahi, P. & Jenkins, M.W. (2015) Validation of Bacteroidales quantitative PCR assays targeting human and animal fecal contamination in the public and domestic domains in India. *Science of The Total Environment*. 502462–470.

Office, N.A. (2010) *Tackling Diffuse Pollution*. [Online] [online]. Available from: https://www.nao.org.uk/wp-content/uploads/2010/07/1011188.pdf.

Oh, S., Buddenborg, S., Yoder-Himes, D.R., Tiedje, J.M. & Konstantinidis, K.T. (2012) Genomic Diversity of Escherichia Isolates from Diverse Habitats. *PLoS ONE*. 7 (10), .

Ohkuma, M., Noda, S. & Kudo, T. (1999) Phylogenetic Diversity of Nitrogen Fixation Genes in the Symbiotic Microbial Community in the Gut of Diverse Termites. *Applied and Environmental Microbiology*. 65 (11), 4926 LP-4934.

Oksanen, J., Blanchet, G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.., O’Hara, R.., Simpson, G.., Solymos, P., Henry, M., Stevens, H., Szoecs, E. & Wagner, H. (2018) *Vegan: Comunity Ecology Package*.

Oladeinde, A., Bohrmann, T., Wong, K., Purucker, Bradshaw, K., Brown, R., Snyder, B. & Molina, M. (2014) Decay of fecal indicator bacterial populations and bovine-associated source-tracking markers in freshly deposited cow pats. *Applied and Environmental Microbiology*. 80 (1), 110–118.

Oliver, D.M., Bird, C., Burd, E. & Wyman, M. (2016) Quantitative PCR profiling of Escherichia coli in livestock feces reveals increased population resilience relative to culturable counts under temperature extremes. *Environmental Science and Technology*. 50 (17), 9497–9505.

Oliver, D.M., Hanley, N.D., van Niekerk, M., Kay, D., Heathwaite, A.L., Rabinovici, S.J.M., Kinzelman, J.L., Fleming, L.E., Porter, J., Shaikh, S., Fish, R., Chilton, S., Hewitt, J., Connolly, E., Cummins, A., Glenk, K., McPhail, C., McRory, E., McVittie, A., *et al*. (2016) Molecular tools for bathing water assessment in Europe: Balancing social science research with a rapidly developing environmental science evidence-base. *Ambio*. 45 (1), 52–62.

Oliver, D.M., van Niekerk, M., Kay, D., Heathwaite, A.L., Porter, J., Fleming, L.E., Kinzelman, J.L., Connolly, E., Cummins, A., McPhail, C., Rahman, A., Thairs, T., de Roda Husman, A.M., Hanley, N.D., Dunhill, I., Globevnik, L., Harwood, V.J., Hodgson, C.J., Lees, D.N., *et al*. (2014) Opportunities and limitations of molecular methods for quantifying microbial compliance parameters in EU bathing waters. *Environment International*. 64124–128.

Oliver, J.D. (2010) Recent findings on the viable but nonculturable state in pathogenic bacteria. *FEMS microbiology reviews*. 34 (4), 415–425.

Ostrolenk, M., Kramer, N. & Cleverdon, R.C. (1947) Comparative Studies of Enterococci and Escherichia coli as Indices of Pollution. *Journal of bacteriology*. 53 (2), 197–203.

Ott, E.M., Muller, T., Muller, M., Franz, C.M., Ulrich, A., Gabel, M. & Seyfarth, W. (2001) Population dynamics and antagonistic potential of enterococci colonizing the phyllosphere of grasses. *Journal of applied microbiology*. 91 (1), 54–66.

Owen, G.J., Perks, M.T., Benskin, C.M.H., Wilkinson, M.E., Jonczyk, J. & Quinn, P.F. (2012) Monitoring agricultural diffuse pollution through a dense monitoring network in the River Eden Demonstration Test Catchment, Cumbria, UK. *Area*. 44 (4), 443–453.

Parkkali, S., Joosten, R., Fanoy, E., Pijnacker, R., Van Beek, J., Brandwagt, D. & Van Pelt, W. (2017) Outbreak of diarrhoea among participants of a triathlon and a duathlon on 12 July 2015 in Utrecht, the Netherlands. *Epidemiology and Infection*. 145 (10), 2176–2184.

Parveen, S., Hodge, N.C., Stall, R.E., Farrah, S.R. & Tamplin, M.L. (2001) Phenotypic and genotypic characterization of human and nonhuman Escherichia coli. *Water research*. 35 (2), 379–386.

Parveen, S., Murphree, R.L., Edmiston, L., Kaspar, C.W., Portier, K.M. & Tamplin, M.L. (1997) Association of multiple-antibiotic-resistance profiles with point and nonpoint sources of Escherichia coli in Apalachicola Bay. *Applied and environmental microbiology*. 63 (7), 2607–2612.

Payan, A., Ebdon, J., Taylor, H., Gantzer, C., Ottoson, J., Papageorgiou, G.T., Blanch, A.R., Lucena, F., Jofre, J. & Muniesa, M. (2005) Method for isolation of Bacteroides bacteriophage host strains suitable for tracking sources of fecal pollution in water. *Applied and Environmental*

Microbiology. 71 (9), 5659–5662.

Penakalapati, G., Swarthout, J., Delahoy, M.J., McAliley, L., Wodnik, B., Levy, K. & Freeman, M.C. (2017) Exposure to Animal Feces and Human Health: A Systematic Review and Proposed Research Priorities. *Environmental Science and Technology*. 51 (20), 11537–11552.

Perkins, T.L., Clements, K., Baas, J.H., Jago, C.F., Jones, D.L., Malham, S.K. & McDonald, J.E. (2014) Sediment composition influences spatial variation in the abundance of human pathogen indicator bacteria within an estuarine environment. *PloS one*. 9 (11), e112951.

Phillips, P., Twigger-Ross, C., Cotton, I., Gianferrara, E., Orr, P., Cherchi, F., Wyles, K., Boschoff, J. & Haydon, P. (2018) *The Value of Bathing Waters and the Influence of Bathing Water Quality: Final Research Report*.

Pinner, V. (2014) *Seaton sluice Bathing Water Catchment Investigation, 2014*.

Price, M.N., Dehal, P.S. & Arkin, A.P. (2010) FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE*. 5 (3), .

Priestley, S. (2015) *Water Framework Directive: Achieving Good Status of Water Bodies*.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J. & Glöckner, F.O. (2007) SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*. 35 (21), 7188–7196.

Prüss-Ustün, A., Bartram, J., Clasen, T., Colford, J.M., Cumming, O., Curtis, V., Bonjour, S., Dangour, A.D., De France, J., Fewtrell, L., Freeman, M.C., Gordon, B., Hunter, P.R., Johnston, R.B., Mathers, C., Mäusezahl, D., Medlicott, K., Neira, M., Stocks, M., *et al*. (2014) Burden of disease from inadequate water, sanitation and hygiene in low- and middle-income settings: A retrospective analysis of data from 145 countries. *Tropical Medicine and International Health*. 19 (8), 894–905.

Prüss, A. (1998) Review of epidemiological studies on health effects from exposure to recreational water. *International Journal of Epidemiology*. 27 (1), 1–9.

Prüß, B.M., Besemann, C., Denton, A. & Wolfe, A.J. (2006) A complex transcription network controls the early stages of biofilm development by Escherichia coli. *Journal of Bacteriology*. 188 (11), 3731–3739.

Purnell, S.E., Ebdon, J.E. & Taylor, H.D. (2011) Bacteriophage Lysis of Enterococcus Host Strains: A Tool for Microbial Source Tracking? *Environmental Science & Technology*. 45 (24), 10699–10705.

Purnell, S.E., Ebdon, J.E., Wilkins, H. & Taylor, H.D. (2018) Human-specific phages infecting Enterococcus host strain MW47: are they reliable microbial source tracking markers? *Journal of Applied Microbiology*. 124 (5), 1274–1282.

Quince, C., Lanzen, A., Davenport, R.J. & Turnbaugh, P.J. (2011) Removing Noise From Pyrosequenced Amplicons. *BMC Bioinformatics*. 12 (1), 38.

Raith, M.R., Ebentier, D.L., Cao, Y., Griffith, J.F. & Weisberg, S.B. (2014) Factors affecting the relationship between quantitative polymerase chain reaction (qPCR) and culture-based enumeration of Enterococcus in environmental waters. *Journal of applied microbiology*. 116 (3), 737–746.

Rasko, D.A., Myers, G.S.A. & Ravel, J. (2005) Visualization of comparative genomic analyses by BLAST score ratio. *BMC Bioinformatics*. 62.

Rasko, D.A., Rosovitz, M.J., Myers, G.S.A., Mongodin, E.F., Fricke, W.F., Gajer, P., Crabtree, J., Sebaihia, M., Thomson, N.R., Chaudhuri, R., Henderson, I.R., Sperandio, V. & Ravel, J. (2008) The pangenome structure of Escherichia coli: Comparative genomic analysis of E. coli commensal and pathogenic isolates. *Journal of Bacteriology*. 190 (20), 6881–6893.

Reischer, G.H., Haider, J.M., Sommer, R., Stadler, H., Keiblinger, K.M., Hornek, R., Zerobin, W., Mach, R.L. & Farnleitner, A.H. (2008) Quantitative microbial faecal source tracking with sampling guided by hydrological catchment dynamics. *Environmental Microbiology*. 10 (10), 2598–2608.

Reischer, G.H., Kasper, D.C., Steinborn, R., Mach, R.L. & Farnleitner, A.H. (2006) Quantitative PCR method for sensitive detection of ruminant fecal pollution in freshwater and evaluation of this method in alpine karstic regions. *Applied and Environmental Microbiology*. 72 (8), 5610–5614.

Revitt, D.M. & Ellis, J.B. (2016) Urban surface water pollution problems arising from misconnections. *Science of the Total Environment*. 551–552163–174.

Rhodes, K. walsh and V. (2016) *Position Statement: Using DNA-based methods for environmental monitoring and decision making*.

Ridley, C.M., Jamieson, R.C., Truelstrup Hansen, L., Yost, C.K. & Bezanson, G.S. (2014) Baseline and storm event monitoring of Bacteroidales marker concentrations and enteric pathogen presence in a rural Canadian watershed. *Water Research*. 60278–288.

Roguet Adélaïdeand Eren, A.M. and N.R.J. and M.S.L. (2018) Fecal source identification using random forest. *Microbiome*. 6 (1), 185.

Rosario, K., Symonds, E.M., Sinigalliano, C., Stewart, J. & Breitbart, M. (2009) Pepper mild mottle virus as an indicator of fecal pollution. *Applied and Environmental Microbiology*. 75 (22), 7261–7267.

Rowland, C.S., Morton, R.D., Carrasco, L., McShane, G., O'Neil, A.W. & Wood, C.M. (2017) *Land Cover Map 2015 (25m raster, GB)*. [Online]

Royal Haskoning (2007) *Cost-Effectiveness of Measures: Analysis of Measures to Reduce Non-Agricultural Diffuse Pollution*.

Sabat, G., Rose, P., Hickey, W.J. & Harkin, J.M. (2000) Selective and sensitive method for PCR amplification of Escherichia coli 16S rRNA genes in soil. *Applied and Environmental Microbiology*. 66 (2), 844–849.

Sahl, J.W., Gregory Caporaso, J., Rasko, D.A. & Keim, P. (2014) The large-scale blast score ratio (LS-BSR) pipeline: A method to rapidly compare genetic content between bacterial genomes. *PeerJ*. 2014 (1), .

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A. & Arnheim, N. (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science (New York, N.Y.)*. 230 (4732), 1350–1354.

Salvidge, R. (2016) England's Waters to Remain Illegally Polluted Beyond 2021. *theENDSreport*.

Samarajeewa, A.D., Hammad, A., Masson, L., Khan, I.U.H., Scroggins, R.

& Beaudette, L.A. (2015) Comparative assessment of next-generation sequencing, denaturing gradient gel electrophoresis, clonal restriction fragment length polymorphism and cloning-sequencing as methods for characterizing commercial microbial consortia. *Journal of Microbiological Methods*. 108103–111.

Sassoubre, L.M., Yamahara, K.M. & Boehm, A.B. (2015) Temporal stability of the microbial community in sewage-polluted seawater exposed to natural sunlight cycles and marine microbiota. *Applied and Environmental Microbiology*. 81 (6), 2107–2116.

Sauvé, S., Aboulfadl, K., Dorner, S., Payment, P., Deschamps, G. & Prévost, M. (2012) Fecal coliforms, caffeine and carbamazepine in stormwater collection systems in a large urban area. *Chemosphere*. 86 (2), 118–123.

Sawilowsky, S.. (2009) New Effect Size Rules of Thumb. *Journal of Modern Applied Statistical Methods*. 8 (2), 597–599.

Schang, C., Henry, R., Kolotelo, P.A., Prosser, T., Crosbie, N., Grant, T., Cottam, D., O'Brien, P., Coutts, S., Deletic, A. & McCarthy, D.T. (2016) Evaluation of techniques for measuring microbial hazards in bathing waters: A comparative study. *PLoS ONE*. 11 (5), .

Schill, W.B. & Mathes, M. V (2008) Real-Time PCR Detection and Quantification of Nine Potential Sources of Fecal Contamination by Analysis of Mitochondrial Cytochrome b Targets. *Environmental Science & Technology*. 42 (14), 5229–5234.

Schleifer, K.H. & Kilpper-Balz, R. (1984) Transfer of Streptococcus faecalis and Streptococcus faecium to the genus Enterococcus nom. rev. as Enterococcus faecalis comb. nov. and Enterococcus faecium comb. nov. *International Journal of Systematic Bacteriology*. 34 (1), 31–34.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., Van Horn, D.J. & Weber, C.F. (2009) Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology*. 75 (23), 7537 LP-7541.

Schriewer, A., Goodwin, K.D., Sinigalliano, C.D., Cox, A.M., Wanless, D., Bartkowiak, J., Ebentier, D.L., Hanley, K.T., Ervin, J., Deering, L.A., Shanks, O.C., Peed, L.A., Meijer, W.G., Griffith, J.F., SantoDomingo, J., Jay, J.A., Holden, P.A. & Wuertz, S. (2013) Performance evaluation of canine-associated Bacteroidales assays in a multi-laboratory comparison study. *Water Research*. 47 (18), 6909–6920.

Scott, T.M., Rose, J.B., Jenkins, T.M., Farrah, S.R. & Lukasik, J. (2002) Microbial source tracking: Current methodology and future directions. *Applied and Environmental Microbiology*. 68 (12), 5796–5803.

Shah, A.H., Abdelzaher, A.M., Phillips, M., Hernandez, R., Solo-Gabriele, H.M., Kish, J., Scorzetti, G., Fell, J.W., Diaz, M.R., Scott, T.M., Lukasik, J., Harwood, V.J., McQuaig, S., Sinigalliano, C.D., Gidley, M.L., Wanless, D., Ager, A., Lui, J., Stewart, J.R., *et al*. (2011) Indicator microbes correlate with pathogenic bacteria, yeasts and helminthes in sand at a subtropical recreational beach site. *Journal of Applied Microbiology*. 110 (6), 1571–1583.

Shanks, O.C., Kelty, C.A., Archibeque, S., Jenkins, M., Newton, R.J., McLellan, S.L., Huse, S.M. & Sogin, M.L. (2011) Community Structures of Fecal Bacteria in Cattle from Different Animal Feeding Operations. *Applied and Environmental Microbiology*. 77 (9), 2992 LP-

3001.

Shokralla, S., Gibson, J.F., Nikbakht, H., Janzen, D.H., Hallwachs, W. & Hajibabaei, M. (2014) Next-generation DNA barcoding: using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. *Molecular ecology resources*. 14 (5), 892–901.

Shuval, H. (2003) Estimating the global burden of thalassogenic diseases: human infectious diseases caused by wastewater pollution of the marine environment. *Journal of water and health*. 1 (2), 53–64.

Simpson, J.M., Santo Domingo, J.W. & Reasoner, D.J. (2002) Microbial source tracking: State of the science. *Environmental Science and Technology*. 36 (24), 5279–5288.

Sinton, L.W., Hall, C.H., Lynch, P.A. & Davies-Colley, R.J. (2002) Sunlight inactivation of fecal indicator bacteria and bacteriophages from waste stabilization pond effluent in fresh and saline waters. *Applied and Environmental Microbiology*. 68 (3), 1122–1131.

Soller, J., Bartrand, T., Ravenscroft, J., Molina, M., Whelan, G., Schoen, M. & Ashbolt, N. (2015) Estimated human health risks from recreational exposures to stormwater runoff containing animal faecal material. *Environmental Modelling and Software*. 7221–32.

Soller, J.A., Schoen, M.E., Bartrand, T., Ravenscroft, J.E. & Ashbolt, N.J. (2010) Estimated human health risks from exposure to recreational waters impacted by human and non-human sources of faecal contamination. *Water Research*. 44 (16), 4674–4691.

Solo-Gabriele, H.M., Wolfert, M.A., Desmarais, T.R. & Palmer, C.J. (2000) Sources of Escherichia coli in a coastal subtropical environment.

Applied and Environmental Microbiology. 66 (1), 230–237.

Spearman, C. (1904) The proof and measurement of association between two things. *The American Journal of Psychology*. 15 (1), 72–101.

Stachler, E., Akyon, B., De Carvalho, N.A., Ference, C. & Bibby, K. (2018) Correlation of crAssphage qPCR Markers with Culturable and Molecular Indicators of Human Fecal Pollution in an Impacted Urban Watershed. *Environmental Science and Technology*. 52 (13), 7505–7512.

Stachler, E. & Bibby, K. (2014) Metagenomic Evaluation of the Highly Abundant Human Gut Bacteriophage CrAssphage for Source Tracking of Human Fecal Pollution. *Environmental Science and Technology Letters*. 1 (10), 405–409.

Stachler, E., Kelty, C., Sivaganesan, M., Li, X., Bibby, K. & Shanks, O.C. (2017) Quantitative CrAssphage PCR Assays for Human Fecal Pollution Measurement. *Environmental Science and Technology*. 51 (16), 9146–9154.

Staley, C., Kaiser, T., Gidley, M.L., Enochs, I.C., Jones, P.R., Goodwin, K.D., Sinigalliano, C.D., Sadowsky, M.J. & Chun, C.L. (2017) Differential impacts of land-based sources of pollution on the microbiota of southeast Florida coral reefs. *Applied and Environmental Microbiology*. 83 (10), .

Staley, C., Kaiser, T., Lobos, A., Ahmed, W., Harwood, V.J., Brown, C.M. & Sadowsky, M.J. (2018) Application of SourceTracker for Accurate Identification of Fecal Pollution in Recreational Freshwater: A Double-Blinded Study. *Environmental Science & Technology*.

Staley, Z.R., Robinson, C. & Edge, T.A. (2016) Comparison of the occurrence and survival of fecal indicator bacteria in recreational sand between urban beach, playground and sandbox settings in Toronto, Ontario. *Science of the Total Environment*. 541520–527.

Stea, E.C., Truelstrup Hansen, L., Jamieson, R.C. & Yost, C.K. (2015) Fecal contamination in the surface waters of a rural and an urban-source watershed. *Journal of Environmental Quality*. 44 (5), 1556–1567.

Stewart, J.R., Boehm, A.B., Dubinsky, E.A., Fong, T.T., Goodwin, K.D., Griffith, J.F., Noble, R.T., Shanks, O.C., Vijayavel, K. & Weisberg, S.B. (2013) Recommendations following a multi-laboratory comparison of microbial source tracking methods. *Water Research*. 47 (18), 6829–6838.

Stoeckel, D.M. & Harwood, V.J. (2007) Performance, design, and analysis in microbial source tracking studies. *Applied and Environmental Microbiology*. 73 (8), 2405–2415.

Stoeckel, D.M., Mathes, M. V, Hyer, K.E., Hagedorn, C., Kator, H., Lukasik, J., O'Brien, T.L., Fenger, T.W., Samadpour, M., Strickler, K.M. & Wiggins, B.A. (2004) Comparison of Seven Protocols To Identify Fecal Contamination Sources Using Escherichia coli. *Environmental Science & Technology*. 38 (22), 6109–6117.

Straughan, E.R. (2012) Touched by water: The body in scuba diving. *Emotion, Space and Society*. 5 (1), 19–26.

Sun, H., He, X., Ye, L., Zhang, X.-X., Wu, B. & Ren, H. (2017) Diversity, abundance, and possible sources of fecal bacteria in the Yangtze River. *Applied Microbiology and Biotechnology*. 101 (5), 2143–2152.

Surfers, A.S. (2009) *Campaign Success For Surfers Against Sewage, Great News For Guernsey's Water Users*. [Online] [online]. Available from: https://www.sas.org.uk/news/campaigns/campaign-success-for-surfers-against-sewage-great-news-for-guernseys-water-users/.

Survey, O. (2017) *Terrain 5 DTM [ASC geospatial data]*. [Online] [online]. Available from: http://digimap.edina.ac.uk/ (Accessed 16 February 2017).

Symonds, E.M., Nguyen, K.H., Harwood, V.J. & Breitbart, M. (2018) Pepper mild mottle virus: A plant pathogen with a greater purpose in (waste)water treatment development and public health management. *Water Research*. 1441–12.

Symonds, E.M., Sinigalliano, C., Gidley, M., Ahmed, W., McQuaig-Ulrich, S.M. & Breitbart, M. (2016) Faecal pollution along the southeastern coast of Florida and insight into the use of pepper mild mottle virus as an indicator. *Journal of applied microbiology*. 121 (5), 1469–1481.

Tan, B., Ng, C., Nshimyimana, J.P., Loh, L.L., Gin, K.Y.-H. & Thompson, J.R. (2015) Next-generation sequencing (NGS) for assessment of microbial water quality: Current progress, challenges, and future opportunities. *Frontiers in Microbiology*. 6 (SEP), .

Tartera, C., Lucena, F. & Jofre, J. (1989) Human origin of Bacteroides fragilis bacteriophages present in the environment. *Applied and Environmental Microbiology*. 55 (10), 2696–2701.

Team, R.D.C. (2017) *R: A Language and Environment for Statistical Computing*. [Online] [online]. Available from: http://www.r-project.org/.

Team, R.S. (2015) *RStudio: Integrated Development for R. RStudio, Inc.*

[Online] [online]. Available from: http://www.rstudio.com/.

Tedjo, D.I., Jonkers, D.M.A.E., Savelkoul, P.H., Masclee, A.A., Best, N. V, Pierik, M.J. & Penders, J. (2015) The effect of sampling and storage on the fecal microbiota composition in healthy and diseased subjects. *PLoS ONE*. 10 (5), .

Tymensen, L.D. (2016) CRISPR1 analysis of naturalized surface water and fecal Escherichia coli suggests common origin. *MicrobiologyOpen*. 5 (3), 527–533.

Tymensen, L.D., Pyrdok, F., Coles, D., Koning, W., McAllister, T.A., Jokinen, C.C., Dowd, S.E. & Neumann, N.F. (2015a) Comparative accessory gene fingerprinting of surface water Escherichia coli reveals genetically diverse naturalized population. *Journal of Applied Microbiology*. 119 (1), 263–277.

Tymensen, L.D., Pyrdok, F., Coles, D., Koning, W., McAllister, T.A., Jokinen, C.C., Dowd, S.E. & Neumann, N.F. (2015b) Comparative accessory gene fingerprinting of surface water Escherichia coli reveals genetically diverse naturalized population. *Journal of Applied Microbiology*. 119 (1), 263–277.

Ufnar, J.A., Wang, S.Y., Christiansen, J.M., Yampara-Iquise, H., Carson, C.A. & Ellender, R.D. (2006) Detection of the nifH gene of Methanobrevibacter smithii: A potential tool to identify sewage pollution in recreational waters. *Journal of Applied Microbiology*. 101 (1), 44–52.

UKAS (2018) *UKAS: Testing laboratory larger scope*. [Online] [online]. Available from: https://www.ukas.com/services/accreditation-services/apply-for-accreditation/what-are-the-costs-of-accreditation/customer-type-testing-laboratory-larger-scope/ (Accessed 17 December 2018).

UNDP United Nations Development programme (2018) *Sustainable Development Goal 6 Targets*. [Online] [online]. Available from: http://www.undp.org/content/undp/en/home/sustainable-development-goals/goal-6-clean-water-and-sanitation/targets/ (Accessed 21 November 2018).

Unno, T., Staley, C., Brown, C.M., Han, D., Sadowsky, M.J. & Hur, H.-G. (2018) Fecal pollution: new trends and challenges in microbial source tracking using next-generation sequencing. *Environmental microbiology*. 20 (9), 3132–3140.

USEPA (2012) *Method 1611: Enterococci in Water by TaqMan® Quantitative Polymerase Chain Reaction (qPCR) Assay*.

USEPA (2005) *Microbial Source Tracking Guide Document*.

USEPA (2004) *Report to Congress: Impacts and Control of CSOs and SSOs*.

USEPA (2015) *REVIEW OF COLIPHAGES AS POSSIBLE INDICATORS OF FECAL CONTAMINATION FOR AMBIENT WATER QUALITY*.

Venables, W.. & Ripley, B.D. (2002) *Modern Applied Statistics with S*. Fourth Edi. New York: Springer.

Verhougstraete, M.P., Byappanahalli, M.N., Rose, J.B. & Whitman, R.L. (2010) Cladophora in the Great Lakes: Impacts on beach water quality and human health. *Water Science and Technology*. 62 (1), 68–76.

Versalovic, J., Schneider, M., De Bruijn, F.J. & Lupski, J.R. (1994) Genomic fingerprinting of bacteria using repetitive sequence-based polymerase chain reaction. *Methods in Molecular and Cellular Biology*. 5 (1), 25–40.

Vierheilig, J., Savio, D., Farnleitner, A.H., Reischer, G.H., Ley, R.E., Mach, R.L., Farnleitner, A.H. & Reischer, G.H. (2015) Potential applications of next generation DNA sequencing of 16S rRNA gene amplicons in microbial water quality monitoring. *Water Science and Technology*. 72 (11), 1962–1972.

Vignaroli, C., Di Sante, L., Magi, G., Luna, G.M., Di Cesare, A., Pasquaroli, S., Facinelli, B. & Biavasco, F. (2015) Adhesion of marine cryptic Escherichia isolates to human intestinal epithelial cells. *ISME Journal*. 9 (2), 508–515.

Villemur, R., Imbeau, M., Vuong, M.N., Masson, L. & Payment, P. (2015) An environmental survey of surface waters using mitochondrial DNA from human, bovine and porcine origin as fecal source tracking markers. *Water Research*. 69 (0), 143–153.

Vogel, L.J., O’Carroll, D.M., Edge, T.A. & Robinson, C.E. (2016) Release of Escherichia coli from Foreshore Sand and Pore Water during Intensified Wave Conditions at a Recreational Beach. *Environmental Science and Technology*. 50 (11), 5676–5684.

Voulvoulis, N., Arpon, K.D. & Giakoumis, T. (2017) The EU Water Framework Directive: From great expectations to problems with implementation. *Science of the Total Environment*. 575358–366.

Wade, T.J., Pai, N., Eisenberg, J.N.S. & Colford Jr, J.M. (2003a) Do U.S. Environmental Protection Agency water quality guidelines for recreational waters prevent gastrointestinal illness? A systematic review and meta-analysis. *Environmental Health Perspectives*. 111 (8), 1102–

1109.

Wade, T.J., Pai, N., Eisenberg, J.N.S. & Colford Jr, J.M. (2003b) Do U.S. Environmental Protection Agency water quality guidelines for recreational waters prevent gastrointestinal illness? A systematic review and meta-analysis. *Environmental Health Perspectives*. 111 (8), 1102–1109.

Walk, S.T., Alm, E.W., Gordon, D.M., Ram, J.L., Toranzos, G.A., Tiedje, J.M. & Whittam, T.S. (2009) Cryptic lineages of the genus Escherichia. *Applied and Environmental Microbiology*. 75 (20), 6534–6544.

Walker, J.W., van Duivenboden, R. & Neale, M.W. (2015) A tiered approach for the identification of faecal pollution sources on an Auckland urban beach. *New Zealand Journal of Marine and Freshwater Research*.

Walters, S.P., Yamahara, K.M. & Boehm, A.B. (2009) Persistence of nucleic acid markers of health-relevant organisms in seawater microcosms: Implications for their use in assessing risk in recreational waters. *Water Research*. 43 (19), 4929–4939.

Wang, D., Farnleitner, A.H., Field, K.G., Green, H.C., Shanks, O.C. & Boehm, A.B. (2013) Enterococcus and escherichia coli fecal source apportionment with microbial source tracking genetic markers - Is it feasible? *Water Research*. 47 (18), 6849–6861.

Wangkahad, B., Mongkolsuk, S. & Sirikanchana, K. (2017) Integrated Multivariate Analysis with Nondetects for the Development of Human Sewage Source-Tracking Tools Using Bacteriophages of Enterococcus faecalis. *Environmental Science & Technology*. 51 (4), 2235–2245.

Wanjugi, P., Sivaganesan, M., Korajkic, A., Kelty, C.A., McMinn, B., Ulrich, R., Harwood, V.J. & Shanks, O.C. (2016) Differential decomposition of bacterial and viral fecal indicators in common human pollution types. *Water Research*. 105591–601.

Warish, A., Triplett, C., Gomi, R., Gyawali, P., Hodgers, L. & Toze, S. (2015) Assessment of Genetic Markers for Tracking the Sources of Human Wastewater Associated Escherichia coli in Environmental Waters. *Environmental Science and Technology*. 49 (15), 9341–9346.

Waso, M., Khan, S. & Khan, W. (2018) Development and small-scale validation of a novel pigeon-associated mitochondrial DNA source tracking marker for the detection of fecal contamination in harvested rainwater. *Science of the Total Environment*. 61599–106.

Weidhaas, J., Mantha, S., Hair, E., Nayak, B. & Harwood, V.J. (2015) Evidence for extraintestinal growth of Bacteroidales originating from poultry litter. *Applied and Environmental Microbiology*. 81 (1), 196–202.

Whitehead, P.G., Leckie, H., Rankinen, K., Butterfield, D., Futter, M.N. & Bussi, G. (2016) An INCA model for pathogens in rivers and catchments: Model structure, sensitivity analysis and application to the River Thames catchment, UK. *The Science of the total environment*. 5721601–1610.

Whitman, R.L., Shively, D.A., Pawlik, H., Nevers, M.B. & Byappanahalli, M.N. (2003) Occurrence of Escherichia coli and enterococci in Cladophora (Chlorophyta) in nearshore water and beach sand of Lake Michigan. *Applied and Environmental Microbiology*. 69 (8), 4714–4719.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

Wiggins, B.A., Cash, P.W., Creamer, W.S., Dart, S.E., Garcia, P.P., Gerecke, T.M., Han, J., Henry, B.L., Hoover, K.B., Johnson, E.L., Jones, K.C., McCarthy, J.G., McDonough, J.A., Mercer, S.A., Noto, M.J., Park, H., Phillips, M.S., Purner, S.M., Smith, B.M., *et al*. (2003) Use of Antibiotic Resistance Analysis for Representativeness Testing of Multiwatershed Libraries. *Applied and Environmental Microbiology*. 69 (6), 3399–3405.

Wilcoxon, F. (1945) Individual Comparisons by Ranking Methods. *Biometrics Bulletin*. 1 (6), 80–83.

WILK, M.B. & SHAPIRO, S.S. (1965) An analysis of variance test for normality (complete samples)†. *Biometrika*. 52 (3–4), 591–611.

Wilson, H.J., Khokhar, F., Enoch, D.A., Brown, N.M., Ahluwalia, J., Dougan, G. & Török, M.E. (2018) Point-prevalence survey of carbapenemase-producing Enterobacteriaceae and vancomycin-resistant enterococci in adult inpatients in a university teaching hospital in the UK. *Journal of Hospital Infection*. 100 (1), 35–39.

Woese, C.R., Fox, G.E., Zablen, L., Uchida, T., Bonen, L., Pechman, K., Lewis, B.J. & Stahl, D. (1975) Conservation of primary structure in 16S ribosomal RNA. *Nature*. 254 (5495), 83–86.

Wood, F.., Heathwaite, A.L. & Haygarth, P.. (2005) Evaluating diffuse and point phosphorus contributions to river transfers at different scales in the Taw catchment, Devon, UK. *Journal of Hydrology*. 304 (1–4), 118–138.

Wright, M.E., Solo-Gabriele, H.M., Elmir, S. & Fleming, L.E. (2009)

Microbial load from animal feces at a recreational beach. *Marine pollution bulletin*. 58 (11), 1649–1656.

Wuertz, S., Wang, D. & Reischer, Georg H. Farnleitner, A.H. (2011) 'Library-Independent Bacterial Source Tracking Methods', in C Hagedorm, A R Blanch, & V J Harwood (eds.) *Microbial Source Tracking, Applications, and Case Studies*. 1st edition [Online]. New York: Springer. p.

WWF (2017) *Rivers on the Edge – an assessment of the impact of sewage pollution*.

Yamahara, K.M., Walters, S.P. & Boehm, A.B. (2009) Growth of Enterococci in Unaltered, Unseeded Beach Sands Subjected to Tidal Wetting. *Applied and Environmental Microbiology*. 75 (6), 1517 LP-1524.

Yampara-Iquise, H., Zheng, G., Jones, J.E. & Carson, C.A. (2008) Use of a Bacteroides thetaiotaomicron-specific α-1-6, mannanase quantitative PCR to detect human faecal pollution in water. *Journal of Applied Microbiology*. 105 (5), 1686–1693.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S. & Madden, T.L. (2012) Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*. 13134.

Zhi, S., Banting, G., Li, Q., Edge, T.A., Topp, E., Sokurenko, M., Scott, C., Braithwaite, S., Ruecker, N.J., Yasui, Y., McAllister, T., Chui, L. & Neumann, N.F. (2016a) Evidence of naturalized stress-tolerant strains of Escherichia coli in municipal wastewater treatment plants. *Applied and Environmental Microbiology*. 82 (18), 5505–5518.

Zhi, S., Banting, G., Li, Q., Edge, T.A., Topp, E., Sokurenko, M., Scott, C., Braithwaite, S., Ruecker, N.J., Yasui, Y., McAllister, T., Chui, L. & Neumann, N.F. (2016b) Evidence of naturalized stress-tolerant strains of Escherichia coli in municipal wastewater treatment plants. *Applied and Environmental Microbiology*. 82 (18), 5505–5518.

Zhi, S., Li, Q., Yasui, Y., Banting, G., Edge, T.A., Topp, E., McAllister, T.A. & Neumann, N.F. (2016) An evaluation of logic regression-based biomarker discovery across multiple intergenic regions for predicting host specificity in Escherichia coli. *Molecular Phylogenetics and Evolution*. 103133–142.

Zhi, S., Li, Q., Yasui, Y., Edge, T., Topp, E. & Neumann, N.F. (2015) Assessing host-specificity of Escherichia coli using a supervised learning logic-regression-based analysis of single nucleotide polymorphisms in intergenic regions. *Molecular Phylogenetics and Evolution*. 9272–81.

Ziebuhr, W., Krimmer, V., Rachid, S., Lößner, I., Götz, F. & Hacker, J. (1999) A novel mechanism of phase variation of virulence in Staphylococcus epidermidis: Evidence for control of the polysaccharide intercellular adhesin synthesis by alternating insertion and excision of the insertion sequence element IS256. *Molecular Microbiology*. 32 (2), 345–356.

Appendix

Appendix A - Methods

Appendix A.1 – SPAdes commands biomarkers SI3

#Trim fastq files with Trimmomatic

```
java -jar ~/trimmomatic-0.36.jar PE  1.fastq.gz 2.fastq.gz  R1-p.fq R1-u.fq R2-p.fq R2-u.fq ILLUMINACLIP:/Trimmomatic-0.36/adapters/TruSeq2-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:150

#Combine unpaired sequences into a single fastq file.

cat R1-u.fq R2-u.fq > unpaired.fq

#Assemble genomes with SPAdes

spades.py --cov-cutoff auto -m 70 -t 10 --careful -k 21,31,55,71,91,101,127 -1 R1-p.fq -2 R2-p.fq -s unpaired.fq -o SPADES_assembled_cov_cutoff
```

Appendix A.2 – Qiime script used for runs using Qiime 1.9.1 and SourceTracker

Step 1 – Script preparation, file conversion

```
> mkdir -p PATH:/

> cp " PATH:/Sequence_File.fastq" " PATH:/Sequence_File.fq"

> convert_fastaqual_fastq.py -f " PATH:/Sequence_File.fq" -c fastq_to_fastaqual \

-o " PATH:/fna_quals"
```

Step 2 – Assigning sample ID’s to reads

```
> split_libraries.py -m " PATH:/Mapping_File.txt" \ -f " PATH:/Sequence_File.fna" \

-q “PATH/fna_quals/Morland.qual"  -s 20 -l 100 -M 100 -d -b golay_12 \

-o "PATH/split_out"
```

Step 3 – OTU picking

```
> pick_open_reference_otus.py -r

"PATH/SILVA/release_119/Silva119_release/rep_set/97/Silva_119_rep_set97.fna" \

-i "PATH/split_out/seqs.fna" \

-p "PATH/parameters.txt" -o "PATH/otus" \ -a -m uclust --suppress_align_and_tree
```

Step 4 – Filtering and chimera removal

```
> parallel_align_seqs_pynast.py -i " PATH:rep_set.fna" \

-o " PATH: pynast_aligned_seqs" –t " PATH:/core _Silva119_alignment.fna" –T
```

```
filter_alignment.py -o " \ PATH:/pynast_aligned_seqs"  non     -i " PATH:/ rep_set_aligned.fasta"

> filter_otus_from_otu_table.py -o

" PATH:/otu_table_mc2_w_tax_no_pynast_failures.biom" -i " PATH:/otu_table_mc2_w_tax.biom" \ -e
“\PATH:/rep_set_failures.fasta"

> parallel_identify_chimeric_seqs.py -i " PATH:/rep_set_aligned.fasta" \

-a " PATH:/SILVA/Silva119_release/core_alignment/core_Silva119_alignment.fna" -m ChimeraSlayer -O
15 -T -o " PATH: chimeric_seqs.txt"

> filter_fasta.py -f " PATH:/rep_set_aligned.fasta" -o " PATH:/non_chimeric_rep_set_aligned.fasta" \

-s " PATH:/chimeric_seqs.txt"  -n

> filter_otus_from_otu_table.py –I " PATH:/ otu_table_mc2_w_tax_no_pynast_failures.biom" \

-e " PATH:/chimeric_seqs.txt" -o " PATH:/otu_table_non_chimeric.biom"
```

Step 5 – Rebuilding the tree

```
> make_phylogeny.py -i " PATH: /non_chimeric_rep_set_aligned.fasta" -o "PATH: /non_chimeric.tre"
```

Step 6 - SourceTracker commands

Convert otu_table from .biom to .txt

```
biom convert -i out_ otu_table_non_chimeric.biom -o final_otu_table.txt --to-tsv
```

Run sourcetracker with default settings

```
R PATH:/sourcetracker_for_QIIME.r --slave --vanilla --args -i final_otu_table.txt -m mapping_file.txt -o
Output.
```

Appendix B – Morland

Appendix B.1 – E.coli isolate and faecal samples

Table B.1.1.Area and date of collection of E.coli isolates used in this study

Host	Sample type	Area	Month/ Year	Number used in study

Human	Sewage treatment works 1	County Durham	08/2015	5
	Sewage treatment works 3	County Durham	08/2015	5
	Sewage treatment works 4	County Durham	01/2017	5
	Sewage treatment works 4	County Durham	12/2016	10
	Septic Tanks	Cumbria	05/2016	5
Chicken	Free range individual faeces	County Durham North	04/2015	8
	Free range individual faeces	Northumberland	08/2015	4
	Free range individual faeces	South Northumberland	04/2016	8
Cow	Beef cow Slurry	Cumbria	05/2016	5
	Beef cow Individual faeces	Newcastle	12/2016	9
	Beef cow Individual faeces	County Durham	08/2015	6
Horse	Individual faeces	North Northumberland	04/2015	8
	Individual faeces	County Durham	05/2015	8
	Individual faeces	South Northumberland	12/2016	4
Pig	Individual faeces	North Northumberland	04/2015	12
	Individual faeces	South Northumberland	08/2015	5
	Individual faeces	South Northumberland	12/2016	3
Sheep	Individual faeces	Cumbria	05/2016	3
	Individual faeces	South Northumberland	04/2015	10
	Individual faeces	North Northumberland	12/2016	7
Dog	Individual faeces	South Northumberland	04/2015	10
	Individual faeces	Individual dog owners from across the North East.	12/2016	10

Table B.1.2. Faecal Samples Used in this study

Host	Sample type	Location	Month/ Year	Number of samples
Human	Sewage treatment works 1	County Durham	08/2015	1
	Sewage treatment works 2	County Durham	08/2015	1
	Sewage treatment works 3	County Durham	08/2015	1
	Sewage treatment works 3	County Durham	12/2016	1

	Sewage treatment works 4	Newcastle	07/2016	1
	Sewage treatment works 4	Newcastle	12/2016	1
	Sewage treatment works 5	Northumberland	12/2016	1
	Septic Tank 1	Cumbria	05/2016	1
	Septic Tank 2	Cumbria	06/2016	1
Chicken	Free range chickens	County Durham	04/2015	3
	Free range chickens	North Northumberland	08/2015	3
	Free range chickens	South Northumberland	04/2016	4
Cow	Slurry	Cumbria	15/2016	5
	Individual faeces	Newcastle	12/2016	4
	Individual faeces	County Durham	08/2015	1
Horse	Individual faeces	North Northumberland	04/2015	4
	Individual faeces	South Northumberland	12/2016	4
	Individual faeces	County Durham	05/2015	1
	Individual faeces	County Durham	08/2015	1
Pig	Individual faeces	North Northumberland	04/2015	3
	Individual faeces	South Northumberland	08/2015	4
	Individual faeces	South Northumberland	12/2016	3
Sheep	Individual faeces	North Northumberland	04/2015	4
	Individual faeces	South Northumberland	12/2016	5
	Individual faeces	Cumbria	05/2016	1
Dog	Individual faeces	South Northumberland	04/2015	2
	Individual faeces	Individual dog owners from across the North East.	12/2016	8

Appendix B2 – Morland catchment report

Out of interest, the case study is reported here, giving more background to the catchment and further analysis of results. This is modified from the MSc Thesis of Oliver Crudge to include additional analysis and figures.

1. Background

Demonstration Test Catchments (DTC) is a research project jointly funded by the Department for Environment, Farming and Rural Affairs (DEFRA), the Environment Agency (EA) and the Welsh Assembly Government with the remit of investigating means to balance the need for intensive farming practices and the associated increase in diffuse pollution (Owen *et al*., 2012). The Eden DTC is one of three test catchments in England studied as part of the project. The Eden DTC is located within the Solway-Tweed river basin and contains three focus catchments at Pow, Dacre and Newby. This research focused on the Newby catchment, which contains a mitigation sub-catchment and a control sub-catchment shown above the sampling points Dedra Banks and Sleagill Village (Figure B2.1). The catchment contains two small villages: the catchment outlet sits just below Newby Village with approximately 60 properties; Sleagill village, with 32 properties, sits at the control sub catchment outlet. Approximately 45 additional properties lay scattered around the catchment. The recipient water body of the catchment is the Newby Beck: between 2009 and 2015 the ecological status of the beck has declined from Good to Moderate (Environment Agency 2016).

The DTC project has implemented a number of managerial, structural and vegetative measures within the mitigation catchment, including: agricultural waste management plans, runoff attenuation ponds, ‘dirty’ and ‘clean’ water separation systems, improved agricultural silage and slurry storage, as well as fencing and tree planting along the banks of the water course. These aim to reduce the amount of diffuse pollutants such as suspended solids and attached nutrients from entering the fluvial network. Sub-stations and a weather station have been installed at the mitigation and control sub catchment outlets: at Dedra Banks and Sleagill Village (Figure B.2.1) to measure turbidity and water levels; at the catchment outlet nutrient data, dissolved oxygen content and chlorophyll levels are also measured. It is thought that none of the dwellings in the catchment are connected to a sewerage system and therefore rely on septic tanks. There is a question over the contribution and relative proportion of various faecal sources to the fluvial network. One possible answer is to use microbial source tracking methods to identify and proportion the sources of faecal pollution within the catchment. In the sub-catchment above Sleagill Village there are considerably more dwellings than above Dedra Banks, therefore there may be a greater potential for human faecal pollution contribution, whereas the mitigation catchment above Dedra Banks may have reduced non-human faecal contamination, resulting in a higher proportion of human faecal pollution.

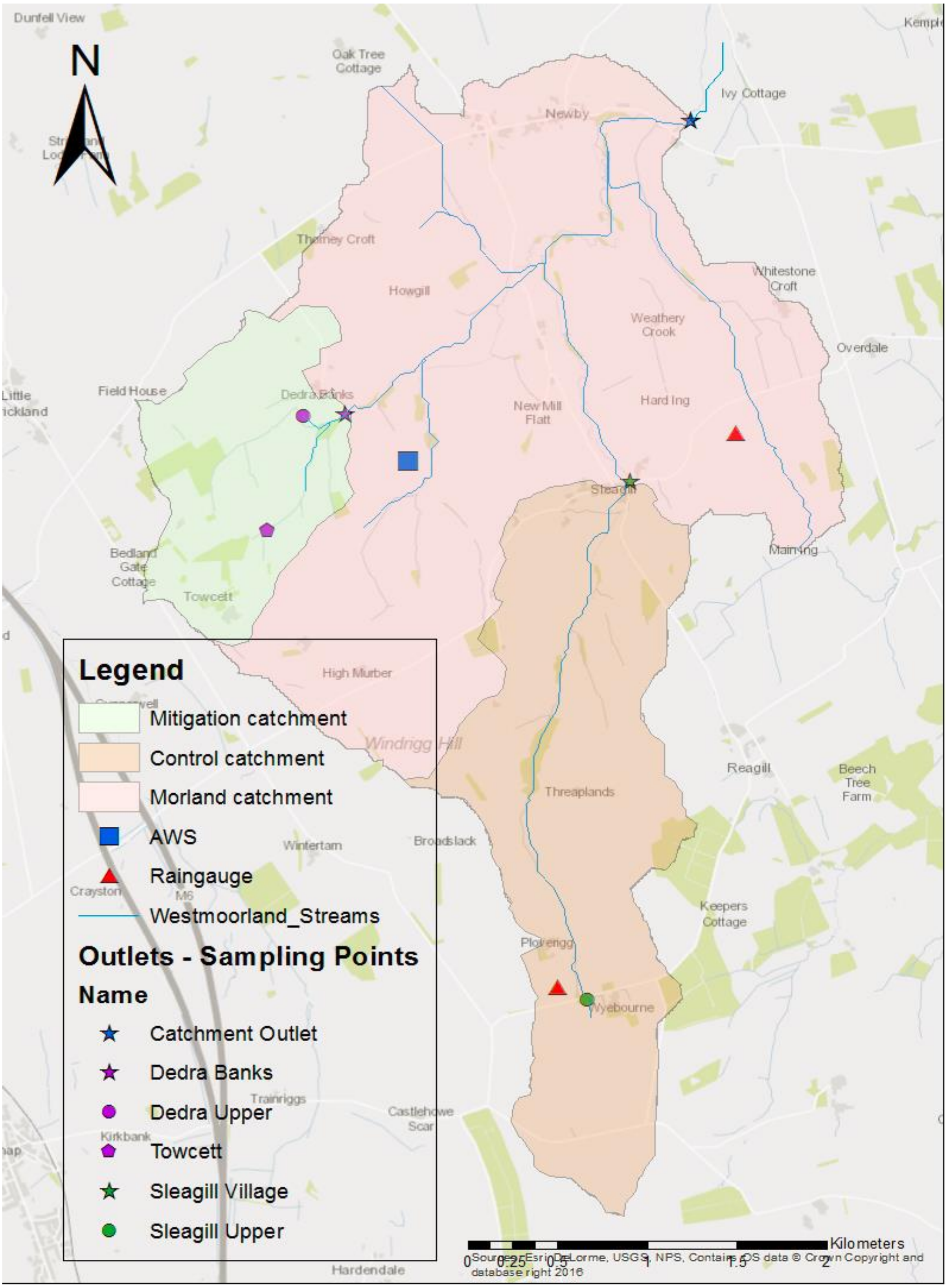


Figure B.2.1. Newby Catchment Outline

2. Current research in the Eden DTC

The availability of high frequency data on a number of water quality parameters and meteorological conditions in the Eden DTC has allowed research into diffuse pollution dynamics within the catchment. Perks *et al.* (2015) assessed the dominant mechanisms for the delivery of phosphorus and suspended solids into the fluvial network over a period of one year. In general, both pollutants had a fast hydrological response to storm events, however analysis of hysteresis curves suggested individual pathways for each. The transport of suspended solids is thought to be dominated by overland flow especially following heavy rainfall, subsurface flow is a lesser factor but is thought to play a larger role at low discharge levels and in the lower areas of the catchment. At low flows, a large proportion of flow within the mitigation sub catchment is thought to take underground pathways, avoiding detection by the substation at Dedra Banks (Figure B2.1). Perks *et al.* (2015) suggested the hysteresis patterns of phosphorus concentration were attributed to a dominant soil water pathway, although have previously suggested that similar dynamics could indicate influence by sources such as septic tanks.

Snell *et al.* (2014) used the high frequency data to assess dynamics of benthic diatoms against antecedent discharge and nutrient conditions over 2 years. An increasing correlation between ecological parameters (trophic diatomic index) and discharge, was found, as the antecedent period increased towards a maximum correlation at 18 days (Snell *et al.* 2014). Concluding that benthic conditions in the catchment rely on meteorological events occurring over the 12 preceding weeks, rather than days. Both Snell *et al* (2014) and Perks *et al* (2015) praised the availability of high frequency data, noting that low frequency sampling can often miss out on significant spikes in pollutant concentrations in catchments with a ‘flashy’ response to precipitation.

3. **Objective of the study**

- To determine if human sources may contribute to the overall pollution and *E.coli* concentration in the Newby catchment.
- To identify other sources of contribution and their relevance compared to human sources.
- To direct future catchment management and research activities in the catchment.

4. Results and discussion

4.1 *E.coli*

Figure B2.2 shows the concentration of total *E. coli* at each site on each sampling visit. The site at Towcett displayed surprisingly large concentrations of *E. coli*. Towcett sits within the DTC mitigation catchment and although the mitigation measures are designed primarily to reduce diffuse sediment and nutrient pollution, it would be expected that these mitigation measures would also reduce faecal contamination.

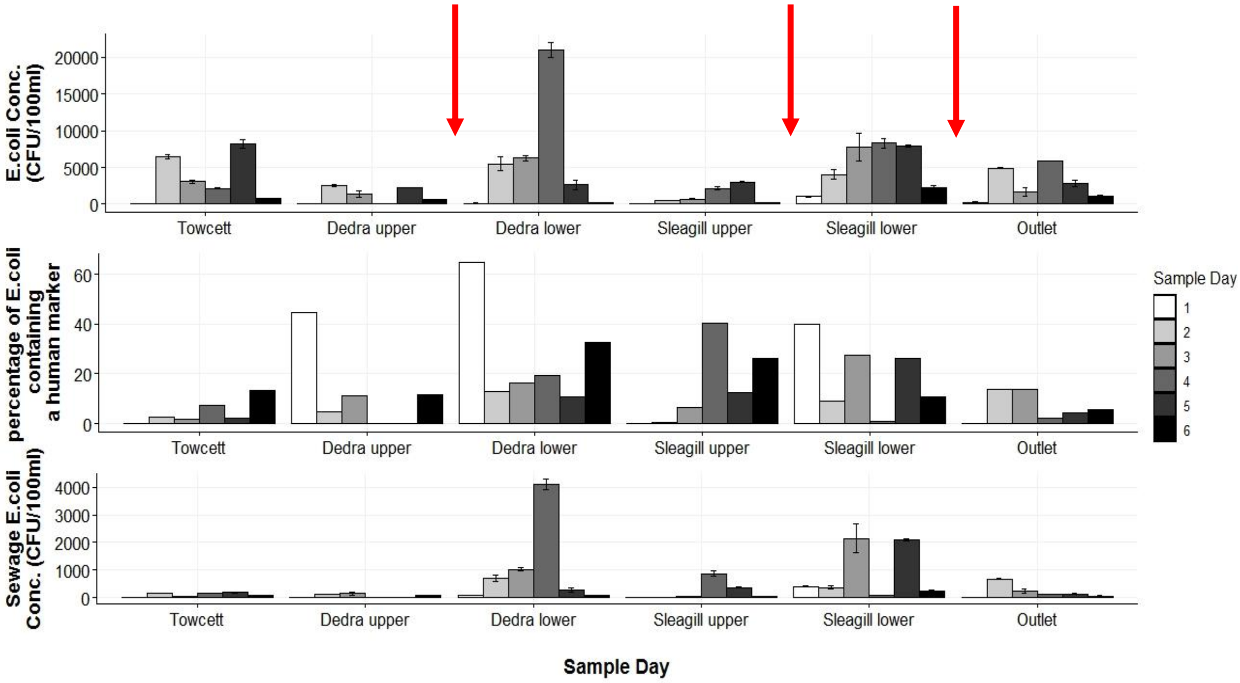


Figure B.2.2. E.coli concentrations at each sample point on sample days 1-6 (Top). Percentage of *E.coli* estimated to be of human origin (Middle). Estimated concentration of E.coli from sewage (Bottom). Arrows denote a farm or settlement along the river system. Error bars show standard error.

Dedra Upper and Sleagill Upper both had low concentrations although below both settlements (Dedra Lower and Sleagill Lower, respectively) there was typically an increase in *E.coli* concentration (Figure B.2.2). Figure B.2.2 also shows the estimated percentage and concentration of *E.coli* from sewage.

Table. B.2.1. Number of samples at each sampling location PCR positive for each E.coli marker.

	H8	H12	H14	H24
Outlet	0	1	3	4
Dedra lower	1	0	5	6
Dedra upper	0	0	2	3
Towcett	0	0	0	5
Sleagill lower	1	1	2	6
Sleagill upper	1	1	2	4
Total	3	3	14	28
Percentage	8.3	8.3	38.9	77.8

Table B.2.1 shows that the H24 marker was the most commonly detected at all sampling points in the catchment, followed by H14. H8 and H12 were rarely detected. The concentration of *E.coli* from sewage generally increases below Dedra and Sleagill as expected, and decreases at the catchment outlet. Interestingly on the fourth sample day (Day 4, Figure B.2.2), the proportion of *E.coli* coming from sewage at Sleagill decreased between Sleagill upper and lower; it is not clear whether prior rainfall or an event at the farm caused the likely increase from agricultural sources on this day.

The total and sewage derived *E.coli* concentrations generally decreased between Sleagill and Dedra lower and the catchment outlet (Figure B2.2). This reduction could be due to die-off of *E.coli*, dilution or a combination of both. Water is added to the river network between Dedra, Sleagill and the catchment Outlet (Figure B.2.3) and farming continues down to the catchment outlet with the largest density of dwellings at Newby sitting above the outlet sampling point, the die off hypothesis therefore seems unlikely.

As the mitigation catchment contains considerably less dwellings than the control catchment, a hypothesis based on uniform septic tank quality would suggest Sleagill Village would have the highest percentage human contribution. A number of reasons exist as to why this may not be true. The catchment largely consists of limestone bedrock; underground flow pathways could therefore allow a large proportion of flow to bypass the sampling site. Sources of overland flow close to the sampling site would therefore have a larger relative contribution, than sources further away, as they will have less chance of entering the underground pathways. Close to the Dedra Banks sampling site a septic tank with obvious signs of over flowing was evident and a compacted earth track provided a surface pathway directly to the sampling site. At certain times of high rainfall or high flows into the tank, this could contribute large amounts of human sourced faecal pollution at this point. Whilst Figure B.2.2 shows general increases in total and human derived *E.coli* concentrations below each settlement, the loading of *E.coli* is much larger at Sleagill compared to Dedra due to the differential flows. This supports the hypothesis that Sleagill contributes a greater amount of *E.coli* than Dedra, and should be an objective for mitigation of pollution in the catchment. Flow data was obtained from the DTC sub-stations at Sleagill village and Dedra Banks as well as at the catchment outlet. The different flows also impact the pollution levels downstream at the Catchment outlet differently. Dedra Banks and Sleagill village were shown to be similarly polluted, however, the flow at Dedra Banks is much lower than at Sleagill village, so the relative impact of the pollution from this site is less. This suggests that pollution above Dedra Banks is of little concern when considering the catchment as a whole, however further analysis of the discharge data is revealing. The average discharge at Dedra Banks and

Sleagill village, during the hours spent sampling accounts for 1.8% and 23% of the discharge at the catchment outlet respectively. Based on the surface area of the two sub catchments this would be expected to be 17% and 30% respectively. The mismatch is much greater at Dedra Banks than Sleagill village and gives evidence that underground flow pathways could be bypassing the sampling sites. The *E. coli* concentrations within the sub surface flow are unknown, thus the contribution of *E. coli* released in the sub catchment above Dedra Banks, to the whole catchment, is difficult to determine, as such a large proportion of flow appeared to follow this pathway.

4.2 Community analysis and other sources of pollution

The community analysis, reported human pollution more often than the human *E.coli* markers. This was not unexpected since community analyses tend towards false positives whereas biomarkers, with their lower sensitivity tend towards false negative results. No significant correlation between the percentage of *E.coli* estimated as human and the proportion of human microbial communities was found, two observations are noteworthy. The decrease in the abundance of human *E.coli* at the outlet (Figure B.2.2) and increase in the human bacterial community is noteworthy. This is likely to be due to the differential die-off rates between *E.coli* and other members of the human faecal community as well as the more rapid die-off of culturable *E.coli* compared with the bacterial DNA examined in the community analysis (Warish *et al*., 2015). Both *E.coli* biomarkers and community analysis indicate a general increase in faecal pollution and slight increase in Human pollution following the Dedra and Sleagill settlements.

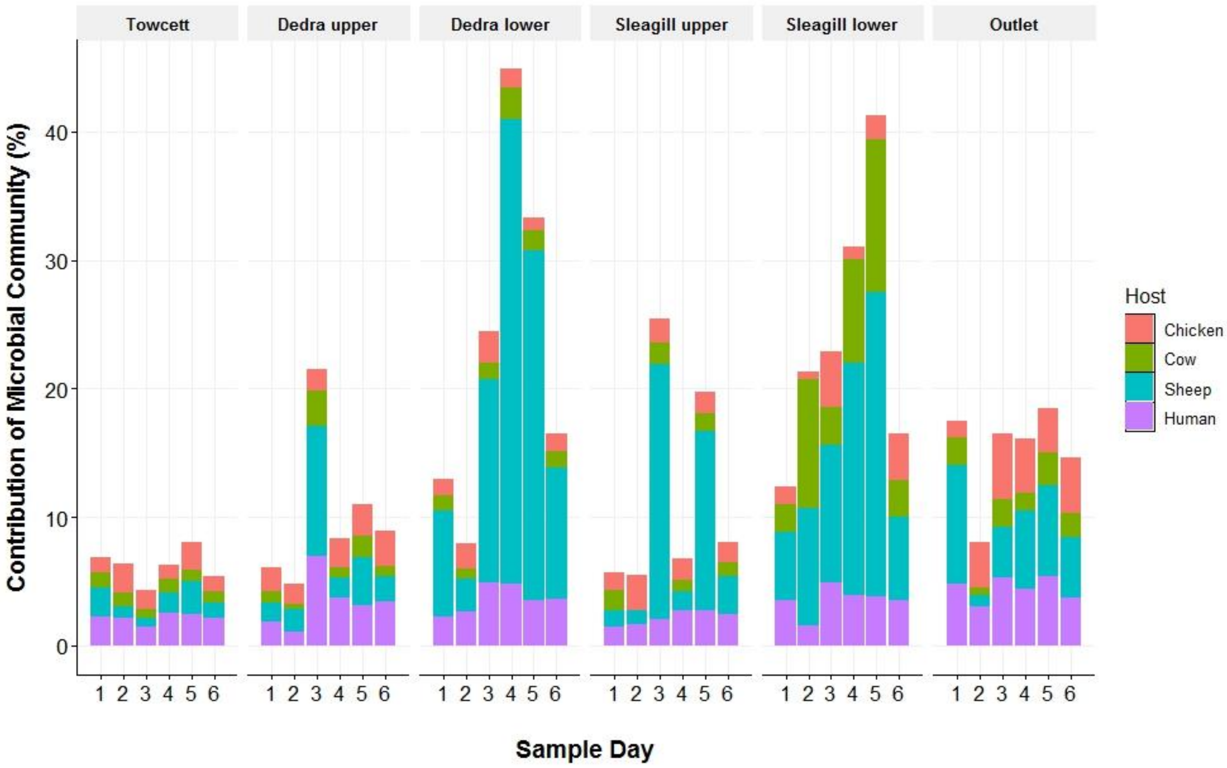


Figure B.2.3 Predicted contribution of source microbial communities to microbial communities in each sample using community analysis.

It is difficult to determine the exact contribution of *E.coli* from sheep or other sources using the community analysis, however, we can make inferences using the human *E.coli* biomarker analysis (Figure B.2.2). Interestingly, at the catchment outlet, the human contribution appears to increase, whereas the *E.coli* markers decrease, which is likely cause by the dilution and differential decay rates as described above, but may also de due to the septic tanks in the immediate vicinity working well and removing a high proportion of the *E.coli* whilst the surviving bacteria leach into the water course.

Following the settlements, sheep appear to be the most abundant source of faecal pollution, particularly at Dedra banks. This is supported by the *E.coli* biomarkers which indicate sewage is typically responsible for 13-50% of the *E.coli* concentrations. Cow faeces is also prevalent in the catchment, although generally less abundant than sheep which could be attributed to many of the cows being housed in sheds. The ubiquitous nature of chicken faeces in the catchment is particularly interesting and was unexpected.

This could be due to the use of fertilizer containing chicken faeces, although it may also be due to cross-reactivity between chicken faeces and other sources. This requires further research if the community analysis technique is going to be used to make investment and management decisions.

4.3 Reducing FIO concentrations in the catchment

Both Dedra and Sleagill farm settlements were highlighted by *E.coli* biomarker and community analyses to be contributors to the total *E.coli* concentrations in the catchment (Figure 2). Due to the differential flow of the fluvial network Sleagill, contributing an average of 35% of the *E.coli* loading to the catchment outlet, is likely to contribute a significantly greater mass of *E.coli* to the catchment than Dedra (3.5%).

A combination of sheep management and septic tank improvements is likely to reduce FIO concentrations at Dedra as these sources appear to be the major contributors. Concentrating effort at Sleagill is likely to result in greater reductions in *E.coli* concentration further down the catchment. Human sources generally contribute less than 25%, of the total *E.coli* concentration, although human contributions of up to 60% were observed (Figure B.2.2). The community analysis suggests that sheep and cow faeces were the major contributors following each farm. At Sleagill, both sheep and cattle management are likely to be required, although septic tank improvements will have a positive impact on FIO concentrations. At both Sleagill and Dedra, we suspect that livestock management is likely to reduce peak FIO concentrations whilst septic tank improvements, through education or regulatory maintenance, are likely to reduce the base concentrations of *E.coli*.

5. Conclusion

This study revealed that human sources do contribute to faecal pollution in the catchment. A bacterial community analysis revealed that human, sheep and cattle faecal sources were, as expected, common in the Newby catchment. The ubiquity of chicken faeces was unexpected and could be a result of the use of fertilizer containing chicken faeces.

The relative contribution of these sources increased, as expected, after Dedra Banks and Sleagill. Enumeration and analysis of *E.coli* and human *E.coli* biomarkers suggest that human sources were responsible for between 1 - 40% and 12 - 60% of the *E.coli* entering the water course following Sleagill and Dedra, respectively.

Discharge data was combined with the estimated concentrations of *E. coli* to show how large proportions of a pollutant do not always equal a large impact on the water quality downstream in the network. Sleagill is likely to contribute a significantly greater mass of *E.coli* to the catchment outlet, although we cannot dismiss the possibility that underground flows are bypassing the Dedra sampling point.

Future catchment management activities should focus on Sleagill and to a lesser extent, Dedra. At Sleagill, management of sheep and cattle in relation to the local water course is likely to have a positive impact. Improvements to septic systems throughout the catchment are likely to reduce FIO concentrations in the catchment.

Appendix B.3 – SourceTracker outputs for Morland SI3

Table 9.3 Sourcetracker outputs for the morland catchment

			Run 1							Run 2							Run 3							Run 4							Run 5						
Location	**Date**	**Day**	**Chicken**	**Cow**	**Horse**	**Sewage**	**Septage**	**Sheep**	**Unknown**	**Chicken**	**Cow**	**Horse**	**Sewage**	**Septage**	**Sheep**	**Unknown**	**Chicken**	**Cow**	**Horse**	**Sewage**	**Septage**	**Sheep**	**Unknown**	**Chicken**	**Cow**	**Horse**	**Sewage**	**Septage**	**Sheep**	**Unknown**	**Chicken**	**Cow**	**Horse**	**Sewage**	**Septage**	**Sheep**	**Unknown**
Dedra lower	13/05/2016		0.0078	0.0127	0.0005	0.0151	0.0049	0.0881	0.8709	0.0144	0.0129	0.0011	0.0214	0.005	0.089	0.8562	0.0206	0.0119	0.0009	0.0185	0.0048	0.0713	0.872	0.0103	0.0118	0.0007	0.0185	0.003	0.0846	0.8711	0.0107	0.0096	0.0008	0.0171	0.0038	0.0799	0.8781
Dedra lower	19/05/2016	2	0.0203	0.0057	0.0006	0.0148	0.0043	0.0259	0.9284	0.0224	0.0052	0.0006	0.0261	0.0089	0.0285	0.9083	0.0184	0.007	0.0012	0.0117	0.0157	0.0281	0.9179	0.0227	0.0116	0.0006	0.0157	0.014	0.0178	0.9176	0.0186	0.0092	0.0007	0.0159	0.0063	0.0247	0.9246
Dedra lower	20/06/2016	3	0.0241	0.0101	0.0007	0.0434	0.0054	0.1745	0.7418	0.0175	0.0146	0.001	0.0456	0.0084	0.1787	0.7342	0.0375	0.0145	0.0009	0.0413	0.007	0.134	0.7648	0.013	0.0077	0.0011	0.0308	0.008	0.1635	0.7759	0.0328	0.0149	0.0022	0.0493	0.0043	0.1441	0.7524
Dedra lower	29/06/2016	5	0.0157	0.0195	0.0013	0.0338	0.0119	0.3016	0.6162	0.01	0.0108	0.001	0.0303	0.0089	0.2811	0.6579	0.0088	0.0202	0.0014	0.0248	0.0073	0.2187	0.7188	0.0057	0.0121	0.0011	0.0262	0.0067	0.2873	0.6609	0.009	0.0166	0.0024	0.0213	0.0065	0.2703	0.6739
Dedra lower	03/07/2016	6	0.0153	0.0101	0.0013	0.0192	0.0085	0.1072	0.8384	0.0122	0.0102	0.0014	0.0233	0.009	0.0962	0.8477	0.0102	0.0175	0.0009	0.0238	0.0163	0.1123	0.819	0.0141	0.0055	0.0006	0.0256	0.0144	0.111	0.8288	0.0204	0.021	0.0015	0.035	0.0085	0.0801	0.8335
Dedra upper	13/05/2016	1	0.0213	0.0108	0.0015	0.0229	0.0027	0.0143	0.9265	0.015	0.005	0.0006	0.011	0.0023	0.014	0.9521	0.0185	0.0056	0.0015	0.0112	0.0037	0.0115	0.948	0.0279	0.0112	0.0012	0.0136	0.0046	0.0147	0.9268	0.0111	0.009	0.001	0.0161	0.0063	0.0196	0.9369
Dedra upper	19/05/2016	2	0.0103	0.0039	0.0011	0.0067	0.0018	0.0202	0.956	0.019	0.0039	0.001	0.0071	0.0035	0.0165	0.949	0.0271	0.0049	0.0008	0.0062	0.0075	0.0178	0.9357	0.0082	0.003	0.0009	0.0044	0.0025	0.0164	0.9646	0.0138	0.0036	0.0018	0.0108	0.0039	0.0152	0.9509
Dedra upper	20/06/2016	3	0.0127	0.0395	0.002	0.0639	0.0101	0.093	0.7788	0.0228	0.023	0.0008	0.0593	0.0078	0.1141	0.7722	0.0179	0.0304	0.0023	0.0374	0.0124	0.0939	0.8057	0.0157	0.0237	0.0023	0.048	0.0131	0.1183	0.7789	0.0115	0.0208	0.0016	0.0834	0.0112	0.0908	0.7807
Dedra upper	27/06/2016	4	0.0303	0.0113	0.0019	0.0371	0.0023	0.0182	0.8989	0.03	0.0081	0.0008	0.0385	0.0026	0.0209	0.8991	0.0259	0.0063	0.0009	0.0298	0.0044	0.0115	0.9212	0.0124	0.0078	0.0009	0.0349	0.0049	0.0124	0.9267	0.0146	0.0059	0.0017	0.0302	0.0026	0.0127	0.9323
Dedra upper	29/06/2016	5	0.0254	0.0158	0.0007	0.0293	0.0051	0.036	0.8877	0.0301	0.0141	0.0008	0.0342	0.0043	0.0399	0.8766	0.0252	0.0207	0.0006	0.0266	0.0059	0.0353	0.8857	0.0163	0.0133	0.0015	0.0144	0.0085	0.0409	0.9051	0.0286	0.0178	0.0002	0.0226	0.0068	0.034	0.89
Dedra upper	03/07/2016	6	0.0269	0.0096	0.001	0.0353	0.0035	0.0175	0.9062	0.0354	0.0095	0.0008	0.0289	0.0033	0.03	0.8921	0.0246	0.0055	0.0006	0.0221	0.0043	0.0177	0.9252	0.0224	0.0083	0.0005	0.0209	0.006	0.0175	0.9244	0.0257	0.0085	0.0014	0.0402	0.0058	0.0169	0.9015
Outlet	13/05/2016	1	0.0196	0.0154	0.0007	0.0224	0.0118	0.0971	0.833	0.0102	0.0187	0.0026	0.041	0.0073	0.1151	0.8051	0.0112	0.019	0.001	0.0362	0.0215	0.0749	0.8362	0.0094	0.0221	0.0016	0.0306	0.0139	0.0876	0.8348	0.0159	0.0334	0.0014	0.0394	0.0147	0.0871	0.8081
Outlet	19/05/2016	2	0.0503	0.0045	0.0007	0.0209	0.002	0.0066	0.915	0.0249	0.0059	0.0008	0.0467	0.0026	0.0126	0.9065	0.0408	0.0046	0.0005	0.0185	0.0031	0.0089	0.9236	0.0328	0.0082	0.0011	0.0246	0.0028	0.0089	0.9216	0.0301	0.0062	0.0006	0.0313	0.0011	0.0056	0.9251
Outlet	20/06/2016	3	0.0721	0.0297	0.0005	0.0437	0.0053	0.043	0.8057	0.0406	0.0227	0.0005	0.0442	0.0067	0.0384	0.8469	0.0538	0.0171	0.0009	0.0318	0.0105	0.047	0.8389	0.0287	0.0177	0.0002	0.0419	0.0105	0.0366	0.8644	0.0613	0.0228	0.0013	0.0603	0.0082	0.033	0.8131
Outlet	27/06/2016	4	0.0373	0.0175	0.0013	0.0236	0.0124	0.0589	0.849	0.0425	0.0128	0.0011	0.0432	0.0114	0.0498	0.8392	0.044	0.0159	0.0006	0.0237	0.0197	0.0656	0.8305	0.035	0.0065	0.0015	0.0235	0.0151	0.0649	0.8535	0.0527	0.0145	0.0013	0.0306	0.0158	0.0692	0.8159
Outlet	29/06/2016	5	0.0307	0.0232	0.0006	0.043	0.0283	0.0745	0.7997	0.0317	0.0321	0.0006	0.0346	0.0162	0.088	0.7968	0.0448	0.0212	0.0014	0.0216	0.0256	0.0698	0.8156	0.0235	0.0168	0.0015	0.0342	0.0206	0.0673	0.8361	0.0425	0.0346	0.0016	0.0366	0.0092	0.0536	0.8219
Outlet	03/07/2016	6	0.0714	0.0217	0.0005	0.0223	0.0062	0.0476	0.8303	0.0262	0.0141	0.0005	0.0292	0.0041	0.0529	0.873	0.0617	0.019	0.0002	0.0294	0.0082	0.0344	0.8471	0.0177	0.0179	0.0004	0.0365	0.0134	0.0555	0.8586	0.0393	0.0178	0.0007	0.0312	0.0083	0.0435	0.8592
Sleagill lower	13/05/2016	1	0.0101	0.0285	0.0009	0.0217	0.0114	0.0585	0.8689	0.0143	0.0206	0.0023	0.0365	0.0078	0.0443	0.8742	0.022	0.0133	0.0016	0.0232	0.0153	0.0501	0.8745	0.01	0.0199	0.0017	0.0194	0.0135	0.0607	0.8748	0.0116	0.0262	0.0021	0.0208	0.005	0.0524	0.8819
Sleagill lower	19/05/2016	2	0.0049	0.1028	0.0063	0.0127	0.0065	0.0982	0.7686	0.0083	0.0998	0.0058	0.0082	0.0059	0.0968	0.7752	0.0079	0.0952	0.0096	0.0119	0.0029	0.0761	0.7964	0.0043	0.1093	0.006	0.0093	0.0038	0.0941	0.7732	0.005	0.0974	0.0046	0.0117	0.006	0.0891	0.7862
Sleagill lower	20/06/2016	3	0.0553	0.0221	0.0007	0.0516	0.0076	0.1091	0.7536	0.0307	0.0293	0.0005	0.0381	0.01	0.096	0.7954	0.0512	0.036	0.001	0.0229	0.0167	0.1005	0.7717	0.0384	0.0262	0.0005	0.0312	0.0138	0.12	0.7699	0.041	0.0348	0.0009	0.0391	0.0128	0.1116	0.7598
Sleagill lower	27/06/2016	4	0.0076	0.0783	0.002	0.0284	0.0095	0.178	0.6962	0.0095	0.0834	0.0025	0.0247	0.0106	0.204	0.6653	0.0146	0.0746	0.0025	0.0275	0.016	0.1593	0.7055	0.0125	0.0988	0.0042	0.0181	0.0147	0.1752	0.6765	0.0066	0.0659	0.0024	0.03	0.0187	0.1852	0.6912
Sleagill lower	03/07/2016	6	0.0341	0.0238	0.0015	0.0224	0.0088	0.0787	0.8307	0.0338	0.029	0.001	0.0311	0.0067	0.0477	0.8507	0.0429	0.0288	0.0014	0.0136	0.0102	0.0706	0.8325	0.0347	0.0172	0.0005	0.0244	0.0164	0.0645	0.8423	0.0353	0.0419	0.0017	0.0334	0.0096	0.065	0.8131
Sleagill upper	19/05/2016	2	0.0308	0.0048	0.0004	0.0184	0.0025	0.0112	0.9319	0.0267	0.0048	0.0009	0.0135	0.0024	0.0113	0.9404	0.0287	0.0042	0.0014	0.0103	0.0021	0.0112	0.9421	0.0292	0.0019	0.0016	0.0114	0.0027	0.0066	0.9466	0.0216	0.0061	0.0013	0.018	0.0023	0.0121	0.9386
Sleagill upper	13/05/2016	1	0.0141	0.0101	0.0009	0.0078	0.0033	0.0124	0.9514	0.0171	0.014	0.0009	0.0172	0.0033	0.0148	0.9327	0.0223	0.0128	0.0009	0.0084	0.0046	0.0065	0.9445	0.0077	0.0114	0.0018	0.0112	0.0074	0.0148	0.9457	0.01	0.0067	0.0007	0.01	0.0021	0.0135	0.957
Sleagill upper	20/06/2016	3	0.0223	0.0185	0.0012	0.021	0.0078	0.2607	0.6685	0.0218	0.0172	0.0015	0.0097	0.0035	0.1808	0.7655	0.0204	0.0132	0.0014	0.0153	0.0052	0.1204	0.8241	0.0206	0.0205	0.0015	0.0173	0.0052	0.2169	0.718	0.0083	0.0173	0.0025	0.0138	0.0044	0.2126	0.7411
Sleagill upper	29/06/2016	5	0.0166	0.0073	0.0011	0.0135	0.0046	0.175	0.7819	0.0254	0.0123	0.0019	0.025	0.0127	0.1325	0.7902	0.0211	0.0197	0.0019	0.0244	0.0057	0.1001	0.8271	0.0098	0.0137	0.001	0.0201	0.0068	0.1566	0.792	0.0104	0.0167	0.0013	0.016	0.0065	0.1353	0.8138
Sleagill upper	27/06/2016	4	0.0144	0.0075	0.0006	0.0324	0.0031	0.0166	0.9254	0.0172	0.0103	0.0021	0.0187	0.0028	0.0133	0.9356	0.0302	0.009	0.001	0.0314	0.0038	0.0086	0.916	0.0113	0.0083	0.0014	0.0123	0.003	0.0221	0.9416	0.0095	0.0128	0.0004	0.0278	0.0024	0.0105	0.9366
Sleagill upper	03/07/2016	6	0.0176	0.0106	0.0004	0.0226	0.0022	0.0274	0.9192	0.013	0.0114	0.0014	0.0206	0.0044	0.0319	0.9173	0.0147	0.0087	0.0008	0.0185	0.0035	0.0266	0.9272	0.016	0.0095	0.0012	0.0112	0.0074	0.0266	0.9281	0.0201	0.0136	0.0009	0.027	0.0072	0.0324	0.8988
Towcett	13/05/2016	1	0.0114	0.0115	0.0006	0.02	0.0011	0.0191	0.9363	0.0165	0.0043	0.0013	0.0258	0.0015	0.0178	0.9328	0.0111	0.0217	0.0015	0.0154	0.0012	0.0252	0.9239	0.0068	0.009	0.0011	0.0187	0.0027	0.0226	0.9391	0.012	0.0139	0.0012	0.0237	0.0014	0.0292	0.9186
Towcett	19/05/2016	2	0.0282	0.0091	0.0011	0.0201	0.0047	0.0115	0.9253	0.0146	0.0061	0.0008	0.0151	0.0081	0.0102	0.9451	0.0283	0.0191	0.0014	0.0084	0.0075	0.0084	0.9269	0.0209	0.0068	0.001	0.0142	0.0032	0.0103	0.9436	0.0187	0.0144	0.0007	0.0205	0.0037	0.0055	0.9365
Towcett	20/06/2016	3	0.0106	0.0057	0.0007	0.0153	0.0013	0.0034	0.963	0.0158	0.007	0.0018	0.0094	0.0015	0.0091	0.9554	0.0204	0.0072	0.0009	0.0111	0.0018	0.0039	0.9547	0.0116	0.0047	0.0008	0.008	0.0021	0.0117	0.9611	0.0145	0.0092	0.0014	0.0204	0.0025	0.0083	0.9437
Towcett	27/06/2016	4	0.0099	0.0064	0.0013	0.0236	0.0018	0.0184	0.9386	0.0163	0.0077	0.0005	0.0144	0.0024	0.0155	0.9432	0.0081	0.0199	0.0005	0.0275	0.0021	0.0127	0.9292	0.0148	0.0066	0.0012	0.0234	0.0023	0.0148	0.9369	0.0092	0.0133	0.0024	0.0253	0.0037	0.0164	0.9297
Towcett	29/06/2016	5	0.0255	0.0093	0.0011	0.0207	0.0022	0.0285	0.9127	0.0235	0.0068	0.001	0.0282	0.0011	0.0189	0.9205	0.0263	0.0114	0.001	0.0214	0.0016	0.0234	0.9149	0.0139	0.0057	0.0009	0.021	0.0016	0.0317	0.9252	0.0211	0.0087	0.002	0.0231	0.0024	0.0265	0.9162
Towcett	03/07/2016	6	0.0113	0.0103	0.0022	0.0181	0.002	0.0127	0.9434	0.0081	0.0093	0.0006	0.0111	0.0048	0.013	0.9531	0.0164	0.0083	0.0004	0.0279	0.001	0.0057	0.9403	0.0127	0.007	0.0008	0.0143	0.0042	0.0182	0.9428	0.0108	0.0089	0.0014	0.0216	0.0031	0.0079	0.9463
Dedra lower	27/06/2016	4	0.0151	0.0269	0.002	0.0346	0.0147	0.3786	0.5281	0.015	0.0263	0.0018	0.0472	0.0106	0.3562	0.5429	0.0104	0.0232	0.0009	0.0369	0.0082	0.3951	0.5253	0.0152	0.0349	0.0006	0.0296	0.0058	0.3491	0.5648	0.0169	0.011	0.0007	0.0461	0.007	0.3302	0.5881
Sleagill lower	29/06/2016	5	0.013	0.1397	0.0024	0.0138	0.0232	0.2089	0.599	0.0322	0.1082	0.0025	0.0123	0.019	0.2261	0.5997	0.0191	0.1058	0.0021	0.0197	0.0204	0.2397	0.5932	0.0176	0.1228	0.0026	0.0144	0.0254	0.2608	0.5564	0.014	0.1178	0.0029	0.0204	0.0213	0.249	0.5746

Appendix B.4 – PCR and qPCR comparison of the individual markers H14 and H24

These two most abundant markers, H14 and H24 appeared in 6 and 7 out of the 10 duplicate samples tested and, 14 and 28 out of the 36 samples used in the study, respectively. Figure B.4.1 shows a comparison of the proportion of isolates containing each of the H14 and H24 markers with a single outlier removed. This outlier was removed because it was the only sample where no common markers were found on both duplicate plates. The maximum difference between duplicate plates for individual markers was 8%.

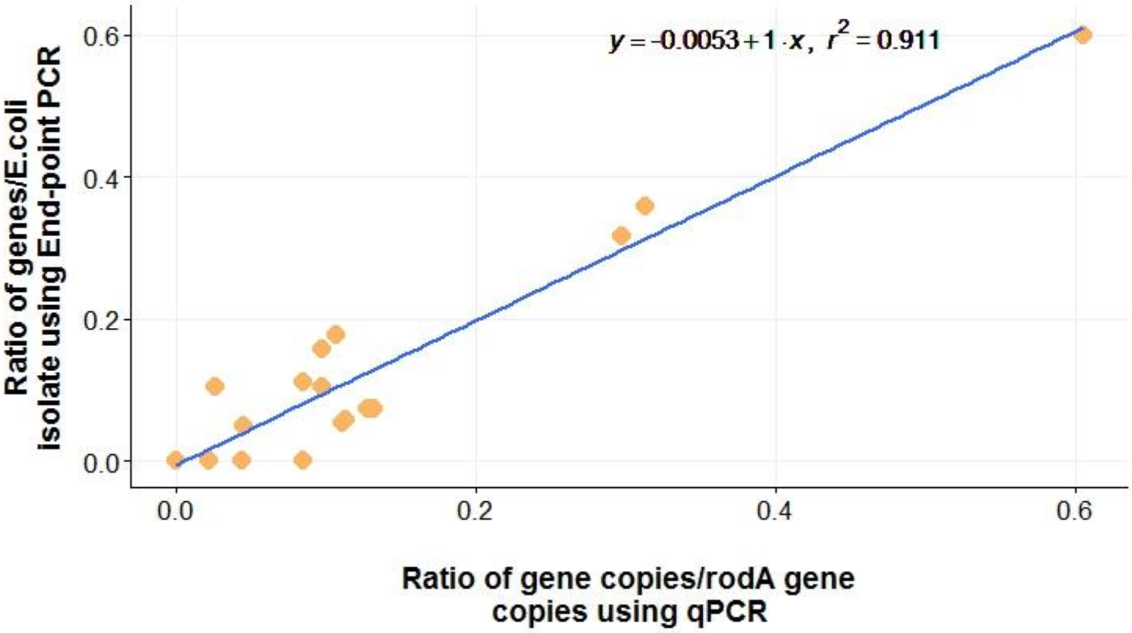


Figure B.4.9.1. Comparison of the ratios of H14 or H24 genes to total E.coli determined through culture and point PCR or qPCR.

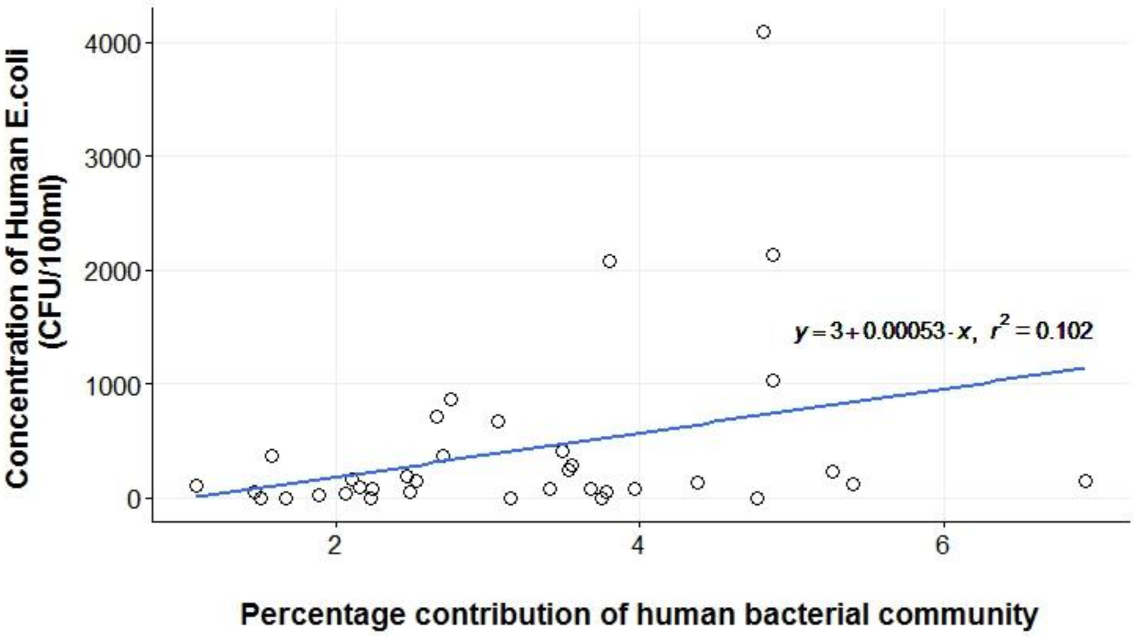


*Figure B.4.3. Relationship between the percentage contributions of human E.coli to the predicted contribution of sewage. - **Pearson's Correlation coefficient = 0.32 P=0.0577***

Appendix C - Biomarkers

Appendix C.1 – Database of genomes

Table C.1.1 Database of E.coli genomes used in this study

Name	Organisms Name (NCBI)	Host/ Environment	Sequence completeness	Genbank Ascensions
H1	Escherichia coli DH1	Human	Complete Genome	GCA_000023365.1
H2	Escherichia coli KO11FL	Human	Complete Genome	GCA_000147855.3
H3	Escherichia coli O145:H28 str. RM13514	Human	Complete Genome	GCA_000520035.1
H4	Escherichia coli O145:H28 str. RM13516	Human	Complete Genome	GCA_000520055.1

H5	Escherichia coli ST2747	Human	Complete Genome	GCA_000599665.1
H6	Escherichia coli MS 198-1	Human	Scaffold	GCA_000164195.1
H7	Escherichia coli MS 84-1	Human	Scaffold	GCA_000164215.1
H8	Escherichia coli MS 115-1	Human	Scaffold	GCA_000164235.1
H9	Escherichia coli MS 182-1	Human	Scaffold	GCA_000164255.1
H10	Escherichia coli MS 146-1	Human	Scaffold	GCA_000164275.1
H11	Escherichia coli MS 45-1	Human	Scaffold	GCA_000164295.1
H12	Escherichia coli MS 69-1	Human	Scaffold	GCA_000164315.1
H13	Escherichia coli MS 187-1	Human	Scaffold	GCA_000164335.1
H14	Escherichia coli O104:H4 str. ON2010	Human	Scaffold	GCA_000258635.1
H15	Escherichia coli LCT-EC106	Human	Scaffold	GCA_000259695.1
H16	Escherichia coli 95JB1	Human	Scaffold	GCA_000478705.1
H17	Escherichia coli 2362-75	Human	Contig	GCA_000183005.2
H18	Escherichia coli 3431	Human	Contig	GCA_000184765.2
H19	Escherichia coli E128010	Human	Contig	GCA_000188775.2
H20	Escherichia coli RN587/1	Human	Contig	GCA_000188875.2

H21	Escherichia coli NCCP15647	Human	Contig	GCA_000259385.1
H22	Escherichia coli NCCP15658	Human	Contig	GCA_000260475.1
H23	Escherichia coli 541-15	Human	Contig	GCA_000264115.1
H24	Escherichia coli 576-1	Human	Contig	GCA_000264135.1
H25	Escherichia coli 75	Human	Contig	GCA_000264155.1
H26	Escherichia coli HM605	Human	Contig	GCA_000285375.1
H27	Escherichia coli 541-1	Human	Contig	GCA_000264215.1
H28	Escherichia coli O104:H4 str. E112/10	Human	Contig	GCA_000350005.1
H29	Escherichia coli ONT:H33 str. C48/93	Human	Contig	GCA_000350025.2
H30	Escherichia coli TOP382-1	Human	Contig	GCA_000350005.1
H31	Escherichia coli TOP382-2	Human	Contig	GCA_000397265.1
H32	Escherichia coli TOP382-3	Human	Contig	GCA_000397285.1
H33	Escherichia coli TOP550-2	Human	Contig	GCA_000397445.1
H34	Escherichia coli TOP550-3	Human	Contig	GCA_000397465.1
H35	Escherichia coli TOP550-4	Human	Contig	GCA_000397485.1

H36	Escherichia coli TOP2396-1	Human	Contig	GCA_000397525.1
H37	Escherichia coli TOP2396-2	Human	Contig	GCA_000397545.1
H38	Escherichia coli TOP2396-3	Human	Contig	GCA_000397565.1
H39	Escherichia coli TOP2522-1	Human	Contig	GCA_000397605.1
H40	Escherichia coli TOP2662-1	Human	Contig	GCA_000397645.1
H41	Escherichia coli TOP2662-2	Human	Contig	GCA_000397665.1
H42	Escherichia coli TOP2662-3	Human	Contig	GCA_000397685.1
H43	Escherichia coli TOP2662-4	Human	Contig	GCA_000397705.1
H44	Escherichia coli C639_08	Human	Contig	GCA_000410655.2
H45	Escherichia coli C844_97	Human	Contig	GCA_000410675.2
H46	Escherichia coli O127:H6	Human	Contig	GCA_000442065.2
H47	Escherichia coli O127:H6	Human	Contig	GCA_000442085.2
H48	Escherichia coli C12_92	Human	Contig	GCA_000446075.2
H49	Escherichia coli C1244_91	Human	Contig	GCA_000446095.2
H50	Escherichia coli C1214_90	Human	Contig	GCA_000446115.2
H51	Escherichia coli C154_11	Human	Contig	GCA_000446135.2

H52	Escherichia coli C155_11	Human	Contig	GCA_000446155.2
H53	Escherichia coli C157_11	Human	Contig	GCA_000446175.2
H54	Escherichia coli C161_11	Human	Contig	GCA_000446195.2
H55	Escherichia coli C213_10	Human	Contig	GCA_000446225.2
H56	Escherichia coli OK1114	Human	Contig	GCA_000446245.1
H57	Escherichia coli 2-005-03_S4_C3	Human	Contig	GCA_000627075.1
H58	Escherichia coli 1-182-04_S4_C3	Human	Contig	GCA_000627095.1
H59	Escherichia coli 1-176-05_S4_C3	Human	Contig	GCA_000627135.1
H60	Escherichia coli 1-250-04_S4_C1	Human	Contig	GCA_000627155.1
H61	Escherichia coli 1-182-04_S4_C1	Human	Contig	GCA_000627175.1
H62	Escherichia coli STEC O174:H2 str. 02-04446	Human	Contig	GCA_000647455.1
H63	Escherichia coli STEC O174:H8 str. 02-07607	Human	Contig	GCA_000647495.1
H65	Escherichia coli 1-176-05_S4_C2	Human	Contig	GCA_000687005.1
H66	Escherichia coli 4541-1	Human	Contig	GCA_000699265.1
H67	Escherichia coli 4552-1	Human	Contig	GCA_000699285.1

H68	Escherichia coli 10810	Human	Contig	GCA_000699305.1
H69	Escherichia coli 7996-1	Human	Contig	GCA_000699365.1
H70	Escherichia coli 11117	Human	Contig	GCA_000699385.1
H71	Escherichia coli 3-267-03_S3_C1	Human	Contig	GCA_000700105.1
H72	Escherichia coli 3-105-05_S1_C1	Human	Contig	GCA_000700125.1
H73	Escherichia coli 3-105-05_S4_C2	Human	Contig	GCA_000700145.1
H74	Escherichia coli 2-011-08_S1_C3	Human	Contig	GCA_000700165.1
H75	Escherichia coli 2-052-05_S4_C3	Human	Contig	GCA_000703445.1
H76	Escherichia coli 8-415-05_S4_C1	Human	Contig	GCA_000711455.1
H77	Escherichia coli 2-316-03_S1_C1	Human	Contig	GCA_000711475.1
H78	Escherichia coli 2-460-02_S1_C3	Human	Contig	GCA_000711485.1
H79	Escherichia coli 3-020-07_S3_C2	Human	Contig	GCA_000711515.1
H80	Escherichia coli 6-319-05_S4_C2	Human	Contig	GCA_000713095.1
H81	Escherichia coli 6-537-08_S1_C1	Human	Contig	GCA_000713105.1
H82	Escherichia coli 6-319-05_S4_C3	Human	Contig	GCA_000713115.1
H83	Escherichia coli 6-175-07_S1_C3	Human	Contig	GCA_000713135.1

H84	Escherichia coli 6-175-07_S4_C3	Human	Contig	GCA_000713175.1
H85	Escherichia coli CS03	Human	Contig	GCA_000740625.1
H86	Escherichia coli TOP293-2	Human	Contig	GCA_000397345.1
H87	Escherichia coli TOP293-3	Human	Contig	GCA_000397365.1
H88	Escherichia coli TOP293-4	Human	Contig	GCA_000397385.1
H89	Escherichia coli TOP498	Human	Contig	GCA_000397405.1
H90	Escherichia coli O104:H21 str. CFSAN002237	Human	Contig	GCA_000464915.1
H91	Escherichia coli O104:H21 str. CFSAN002236	Human	Contig	GCA_000464955.1
H92	Escherichia Coli K009	Human	Assembled	DRX016668
H93	Escherichia Coli K008	Human	Assembled	DRX016667
H94	Escherichia Coli K007	Human	Assembled	DRX016666
H95	Escherichia Coli K006	Human	Assembled	DRX016665
H96	Escherichia Coli K005	Human	Assembled	DRX016664
H97	Escherichia Coli K004	Human	Assembled	DRX016663
H98	Escherichia Coli K003	Human	Assembled	DRX016662

H99	Escherichia Coli K002	Human	Assembled	DRX016661
H100	Escherichia Coli K001	Human	Assembled	DRX016660
H101	Hu2-2	Human	Assembled	
H102	DH7	Human	Assembled	
H103	DH8	Human	Assembled	
H104	KB4	Human	Assembled	
Ch1	Escherichia coli O08	Chicken	Contig	GCA_000340235.1
Ch2	Escherichia coli S17	Broiler chick	Contig	GCA_000340255.1
Ch3	Escherichia coli SEPT362	Laying Hen	Contig	GCA_000340275.1
Ch4	Escherichia coli APEC IMT5155	Chicken	Complete Genome	GCA_000813165.1
Ch5	Escherichia coli strain KCh005	Chicken	Contig	DRR018455
Ch6	Escherichia coli strain KCh004	Chicken	Contig	DRR018454
Ch7	Escherichia coli strain KCh003	Chicken	Contig	DRR018453
Ch8	Escherichia coli strain KCh002	Chicken	Contig	DRR018452
Ch9	Escherichia coli strain KCh001	Chicken	Contig	DRR018451
Ch10	Escherichia coli AD30	Chicken	Contig	GCA_000304255.1
Ch11	Escherichia coli AD30	Chicken	Contig	GCA_001244915.1
Ch12	Escherichia coli	Chicken Faeces	Contig	GCA_001268185.1

Ch13	Escherichia coli	Chicken Faeces	Contig	GCA_001268205.1
Ch14	Escherichia coli	Chicken Faeces	Contig	GCA_001268225.1
Ch15	Escherichia coli	Chicken Faeces	Contig	GCA_001268425.1
Ch16	Escherichia coli	Chicken Faeces	Contig	GCA_001268885.1
Ch17	Escherichia coli	Chicken Faeces	Contig	GCA_001268965.1
Ch18	Escherichia coli	Chicken Faeces	Contig	GCA_001268985.1
Ch19	Escherichia coli	Chicken Faeces	Contig	GCA_001269085.1
Ch20	Escherichia coli	Chicken Faeces	Contig	GCA_001269105.1
Ch21	Escherichia coli	Chicken Faeces	Contig	GCA_001269285.1
Ch22	Escherichia coli	Chicken Faeces	scaffold	GCA_001268925.1
Ch23	Escherichia coli	Chicken Faeces	scaffold	GCA_001269065.1
Ch24	Escherichia coli	Chicken Faeces	Complete Genome	GCA_001660565.1
Ch25	Escherichia coli APEC O2	Chicken Faeces	Contig	GCA_001620375.1
Ch26	E.coli Strain: EC2_7	Chicken Intestine	Contig	GCA_000812385.1
Ch27	Escherichia coli 38.16 (E. coli)	Chicken Caecum	Contig	GCA_000503295.1
Ch28	Escherichia coli 38.52 (E. coli)	Chicken Caecum	Contig	GCA_000503675.1

Ch29	Escherichia coli 38.29 (E. coli)	Chicken Caecum	Contig	GCA_000503355.1
Ch30	CK4-2	Chicken Faeces	Assembled	
Ch31	CK6-2	Chicken Faeces	Assembled	
Ch32	CK8-2	Chicken Faeces	Assembled	
Hor1	Escherichia coli 3.2608	Horse	Contig	GCA_000215205.2
Hor2	H1-3	Horse	Assembled	
Hor3	H2-3	Horse	Assembled	
Hor4	H1-1	Horse	Assembled	
Hor5	Escherichia coli MOD1-EC5143	Horse	Assembled	GCA_002516765.1
Hor6	Escherichia coli MOD1-EC6554	Horse	Assembled	GCA_002511685.1
Hor7	Escherichia coli MOD1-EC6535	Horse	Assembled	GCA_002513055.1
Hor8	Escherichia coli MOD1-EC6533	Horse	Assembled	GCA_002513075.1
Hor9	Escherichia coli MOD1-EC6495	Horse	Assembled	GCA_002513315.1
Hor10	Escherichia coli MOD1-EC6491	Horse	Assembled	GCA_002512015.1
Hor11	Escherichia coli MOD1-EC6489	Horse	Assembled	GCA_002513395.1
Hor12	Escherichia coli MOD1-EC5108	Horse	Assembled	GCA_002231925.1
Hor13	Escherichia coli MOD1-EC5107	Horse	Assembled	GCA_002232375.1
Hor14	Escherichia coli MOD1-EC6487	Horse	Assembled	GCA_002512035.1

Hor15	Escherichia coli MOD1-EC6486	Horse	Assembled	GCA_002513415.1
Hor16	Escherichia coli MOD1-EC6420	Horse	Assembled	GCA_002510855.1
S1	CHS3-3	Sheep Faeces	Contig	
S2	E.coli O157	Sheep Faeces	Contig	SAMEA3635213
S6	E.coli O157	Sheep Faeces	Contig	SAMEA3635214
S7	E.coli O157	Sheep Faeces	Contig	SAMEA3635215
S8	E.coli O157	Sheep Faeces	Contig	SAMEA3635216
S9	E.coli O157	Sheep Faeces	Contig	SAMEA3635217
S10	E.coli O157	Sheep Faeces	Contig	SAMEA3635218
S11	S5-3	Sheep Faeces	Assembled	
S12	CH3-2	Sheep Faeces	Assembled	
S13	Eshcerichia coli FHI38	Sheep Faeces	Scaffold	GCA_000753035.1
S14	Eshcerichia coli FHI39	Sheep Faeces	Scaffold	GCA_000752875.1
S15	Eshcerichia coli FHI37	Sheep Faeces	Scaffold	GCA_000752815.1
D1	Eshcerichia coli IMT31352	Dog Faeces	Contig	GCA_001282235.1
D2	Escherichia coli KD1	Dog Faeces	Contig	GCA_000264095.1
D3	Escherichia coli KD2	Dog Faeces	Contig	GCA_000264195.1
D4	Eshcerichia Coli Strain: IMT31359	Dog Faeces	Contig	GCA_001282195.1
D5	Escherichia Coli Strain: IMT31351	Dog Faeces	Contig	GCA_001282155.1
D6	Escherichia Coli Strain: IMT31487	Dog Faeces	Contig	GCA_001282345.1
D7	D1-2	Dog Faeces	Assembled	
D8	D3-2	Dog Faeces	Assembled	

D9	Eshcerichia coli MOD1-EC6946	Dog Faeces	Contig	GCA_002232275.1
D10	Eshcerichia coli MOD1-EC5069	Dog Faeces	Contig	GCA_002232275.1
D11	Eshcerichia coli MOD1-EC5083	Dog Faeces	Contig	GCA_002232865.1
B1	Escherichia coli O157:H7 str. SS17	Cattle	Complete Genome	GCA_000730345.1
B2	Escherichia coli AA86	Cow	Scaffold	GCA_000211395.2
B3	Escherichia coli EC4196	Cattle	Scaffold	GCA_000267445.2
B4	Escherichia coli EC4203	Cattle	Scaffold	GCA_000267465.2
B5	Escherichia coli FRIK1996	Cattle	Scaffold	GCA_000267505.2
B6	Escherichia coli FRIK1985	Cattle	Scaffold	GCA_000267525.2
B7	Escherichia coli 93-001	Cattle	Scaffold	GCA_000267945.2
B8	Escherichia coli FRIK1990	Cattle	Scaffold	GCA_000267965.2
B9	Escherichia coli O157:H7 str. FRIK966	Bovine	Contig	GCA_000175735.1
B10	Escherichia coli O157:H7 str. FRIK2000	Bovine	Contig	GCA_000175755.1
B11	Escherichia coli 1.2741	Cow	Contig	GCA_000194175.2
B12	Escherichia coli 97.0246	Cow	Contig	GCA_000194215.2

B13	Escherichia coli 97.0264	Cow	Contig	GCA_000194295.2
B14	Escherichia coli 4.0522	Cow	Contig	GCA_000194335.2
B15	Escherichia coli 99.0741	Cow	Contig	GCA_000194435.2
B16	Escherichia coli 900105 (10e)	Calf	Contig	GCA_000194725.2
B17	Escherichia coli 5.0588	Cow	Contig	GCA_000215145.2
B18	Escherichia coli 3.3884	Cow	Contig	GCA_000215285.2
B19	Escherichia coli O111:H11 str. CVM9534	Cow	Contig	GCA_000263935.1
B20	Escherichia coli O111:H11 str. CVM9545	Cow	Contig	GCA_000263955.1
B21	Escherichia coli O111:H8 str. CVM9570	Cow	Contig	GCA_000263975.1
B22	Escherichia coli O26:H11 str. CVM9942	Cow	Contig	GCA_000264015.1
B23	Escherichia coli O26:H11 str. CVM10026	Cow	Contig	GCA_000264035.1
B24	Escherichia coli O111:H8 str. CVM9634	Cow	Contig	GCA_000276765.1
B25	Escherichia coli O26:H11 str. CVM10030	Cow	Contig	GCA_000276845.1

B26	Escherichia coli O111:H11 str. CVM9553	Cow	Contig	GCA_000276925.1
B27	Escherichia coli O26:H11 str. CVM10021	Cow	Contig	GCA_000276945.1
B28	Escherichia coli O111:H11 str. CFSAN001630	Cow	Contig	GCA_000313425.1
B29	Escherichia coli C842_97	Cattle	Contig	GCA_000447025.2
B30	Escherichia coli ECC-Z	Bovine	Contig	GCA_000498235.1
B31	Escherichia coli LAU-EC2	Bovine	Contig	GCA_000498795.2
B32	Escherichia coli KC001	Cattle	Assembled	DRR018443
B33	Escherichia coli KC002	Cattle	Assembled	DRR018444
B34	C8-3		Assembled	
B35	C9-3		Assembled	
B36	C5-3		Assembled	
P1	Escherichia coli UMNK88	Pig	Complete Genome	GCA_000212715.2
P2	Escherichia coli UMNF18	Pig	Complete Genome	GCA_000220005.2
P3	Escherichia coli 9.1649	Pig	Contig	GCA_000194475.2
P4	Escherichia coli 2.3916	Pig	Contig	GCA_000194535.2
P5	Escherichia coli B41	Pig	Contig	GCA_000194705.2

P6	Escherichia coli AI27	Pig	Contig	GCA_000259135.1
P7	Escherichia coli O26:H11 str. CVM9952	Pig	Contig	GCA_000276885.1
P8	Escherichia coli IMT8073	Pig	Contig	GCA_000414155.2
P9	Escherichia coli C900_01	Pig	Contig	GCA_000447085.2
P10	Escherichia coli E455	Pig	Contig	GCA_000647795.2
P11	Escherichia coli 77302533	Pig	Contig	GCA_000754845.1
P12	Escherichia coli 77300132	Pig	Contig	GCA_000754855.1
P13	Escherichia coli 77300095	Pig	Contig	GCA_000754865.1
P14	Escherichia coli KP001	Pig	Assembled	
P15	Escherichia coli KP002	Pig	Assembled	
P16	Escherichia coli KP003	Pig	Assembled	
P17	Escherichia coli KP004	Pig	Assembled	
P18	Escherichia coli KP005	Pig	Assembled	
P19	Escherichia coli KP006	Pig	Assembled	
P20	Escherichia coli (E. coli) Strain: W25K	Pig	Contig	GCA_000696835.1

P21	Escherichia coli (E. coli) Strain: 912	Pig	Scaffold	GCA_000806195.1
P22	Escherichia coli PCN033 (Ex PEC)	Pig	Complete Genome	GCA_000219515.3
P23	Escherichia coli PCN061 (E. coli)	Pig	Complete Genome	GCA_001029125.1
P24	Escherichia coli FCP1 (E. coli)	Pig	Contig	GCA_000511565.1
P25	Escherichia coli FBP1 (E. coli)	Pig	Contig	GCA_000511525.1
P26	Escherichia coli FAP 2	Pig	Contig	GCA_000511505.1
P27	Escherichia coli FAP1 (E. coli)	Pig	Contig	GCA_000511485.1
P28	P2-2	Pig	Assembled	
P29	P3-3	Pig	Assembled	
P30	P5-1	Pig	Assembled	
G1	G1-1	Laridae - Sea gull	Assembled	
G2	G2-2	Laridae - Sea gull	Assembled	
G3	Eshcerichia coli MOD1-EC5497	Laridae - Sea gull	Contig	GCA_002229795.1
G4	Eshcerichia coli MOD1-EC5496	Laridae - Sea gull	Contig	GCA_002229775.1
G5	Eshcerichia coli MOD1-EC5495	Laridae - Sea gull	Assembled	SRX1991313
G6	Eshcerichia coli MOD1-EC5492	Laridae - Sea gull	Contig	GCA_002229875.1

Env1	Escherichia sp. TW15838 (enterobacteria)	Environmental Clade I	Contig	GCA_000208485.2
Env2	Escherichia sp. TW09231 (enterobacteria)	Environmental Clade III	Contig	GCA_000208465.2
Env3	Escherichia sp. TW09276 (enterobacteria)	Environmental Clade III	Contig	GCA_000208445.2
Env4	Escherichia sp. TW11588 (enterobacteria)	Environmental Clade IV	Contig	GCA_000208585.2
Env5	Escherichia sp. TW14182 (enterobacteria)	Environmental Clade IV	Contig	GCA_000208525.2
Env6	Escherichia sp. TW09308 (enterobacteria)	Environmental Clade V	Contig	GCA_000208565.2
O1	Escherichia coli 48	Deer	Contig	GCA_000736735.1
O2	Escherichia coli strain:TW18710 (STEC)	Deer	Scaffold	GCA_000969495.1
O3	Escherichia coli strain:117 (STEC)	Deer	Contig	GCA_001902685.1 (
O4	Escherichia coli 1.2264	Goat	Contig	GCA_000194415.2
O5	Escherichia coli CUMT8	Mouse	Contig	GCA_000264235.1
O6	Escherichia coli MP1	Mouse	Contig	GCA_000576655.1
O7	Escherichia coli SWW33	Mouse	Scaffold	GCA_000364305.1

O8	Escherichia coli K02	Mouse	Scaffold	GCA_000607285.1
O9	Escherichia coli 4.0967	Rabbit	Contig	GCA_000194495.2
O10	Escherichia coli C527_94	Rabbit	Contig	GCA_000446625.2
O11	Escherichia coli APEC O1	Turkey	Complete Genome	GCA_000014845.1

Appendix D - Chapter 5 Evaluating the Effect of Library Composition on Community-based MST

Appendix D.1 – R code for mixing simulated microbial communities.

```
####********************Start of R-code************************####

#Load up the required packages

library(phyloseq)

library(plyr)


#Pick an arbitrary number for the seed

set.seed(100)

#set directory

setwd("…")


####**************Load data into phyloseq object****************####

otu_table <- read.csv2("…", sep = ",", row.names = 1)

otu_table <- as.matrix(otu_table)
```

```
#read taxonomy

taxonomy <- read.csv2("…", sep = ",", row.names = 1)

taxonomy <- as.matrix(taxonomy)


#Read metadata

meta <-  read.table("…", sep = "\t", header = TRUE, stringsAsFactors = TRUE)

rownames(meta) <- meta$sample_id


#read in tree

tree <- read_tree("…")

#get into phyloseq

OTU <-  otu_table(otu_table, taxa_are_rows = TRUE)

TAX <-  tax_table(taxonomy)

META <- sample_data(meta)


#Combine in phyloseq object

all_data <- phyloseq(OTU, TAX, META, tree)


#change rank names – if necessary

colnames(tax_table(all_data)) <- c("kingdom", "phylum", "class", "order", "family",
"genus", "species")

rank_names(all_data)


#Remove unnecessary data

rm(META)
rm(meta)
rm(otu_table)
rm(tree)
rm(taxonomy)
rm(OTU)
rm(TAX)

#Filter what samples you require for mixing and to include in the faecal taxon library
Faecal_samples <- subset_samples(all_data,
                              Sample.type == "Faecal")
Faecal_samples <- prune_taxa(taxa_sums(Faecal_samples) > 0, Faecal_samples)

#Display the sample data
sample_data(Faecal_samples)

#Construct OTU table to rebuild the Phyloseq object later
otu_df <- as.data.frame(otu_table(Faecal_samples))
head(otu_df)

#Define sample composition either manually here, or supply in a CSV file. Columns should be the sample names you want to sample from, and rows are the mixtures.
#This is if you want to supply composition manually. String in “” is the sample_id.
#s1=data.frame("cp"=0.1, "dy"=0.9)
#s2=data.frame("dy"=0.2, "eh"=0.1, "cp"=0.7)
#s3=data.frame("eh"=0.05, "cp"=0.1, "dy"=0.85)
```

```
#s4=data.frame("cp"=0.05, "dy"=0.95)

#simlist <- list("sim1"=s1,"sim2"=s2,"sim3"=s3, "sim4"=s4)

#mixture_df <- ldply(simlist, .id = NULL)

#mixture_df[is.na(mixture_df)] <- 0

#rownames(mixture_df) <- names(simlist)

#mixture_df


#Import sample composition using csv file.

mixture_df <- read.csv("… ",row.names = 1)


#Define number of reads to sample

#numreads=mean(sample_sums(Faecal_samples))

#min(sample_sums(Faecal_samples))

numreads = 50000

####**************************Start the sampling****************************####


#Subset the phyloseq object by source, and then convert to relative abundances

relabunds <- list()

for (m in colnames(mixture_df)){

  relabunds[[m]] <- as.data.frame(t(otu_table(transform_sample_counts(subset_samples(Faecal_samples, sample_id==m), function(x) x/sum(x)))))

}
```

```
#Join that all together into a data.frame (for sampling)
relative_abundance_OTUs <- t(do.call( rbind, relabunds))

#Define a "genericish" function to do the actual sampling
runSimulation <- function(fullmat, mixture_vector){
#Extract the columns we want
submat <- fullmat[,names(mixture_vector)]

#Multiply each row by the corresponding composition value, and then add them together
sampling_probs <- apply(submat, 1, function(x) sum(x[names(mixture_vector)] *
mixture_vector))

#Sample (with replacement) from a list of 1..n OTU indices, weighted by the sampling
probability vector
tvec = sample(seq_along(rownames(fullmat)), numreads, replace = T, prob =
sampling_probs)
#Build up a results vector, full of 0
result_vec <- rep(0, length(sampling_probs))

#And substitute counts for the OTUs we've "detected"
sim_result <- table(tvec )
result_vec[as.numeric(names(sim_result))] <- sim_result
result_vec
}
```

```
#Run the simulation(s)

sim_results <- apply(mixture_df, 1, function(x)
runSimulation(relative_abundance_OTUs, x))


#### Turn it all back into a phyloseq object ####


#Fixup the sample data so it includes the new simulated samples, otherwise Phyloseq
ignores them :-(

newsample_df <- as.data.frame(sample_data(Faecal_samples), stringsAsFactors = F)

newsample_df$Simulated <- F

newsample_df[rownames(mixture_df),"Simulated"] <- T

newsample_df$sample_id <- rownames(newsample_df)

newsample_df


#Combine it all back into a phyloseq object

everything_ps <- phyloseq(otu_table(cbind(otu_df, sim_results), taxa_are_rows = T),
tax_table(Faecal_samples), sample_data(newsample_df))

everything_ps



####*******************SENSE CHECK THE RESULTS
*********************####


rel_everything_ps <- transform_sample_counts(everything_ps, function(x) x/sum(x))

o <- ordinate(rel_everything_ps, "NMDS", "bray")
```

```
plot_ordination(rel_everything_ps, o, label = "sample_id", color="Simulated")


####***********************Export OTU Table for Sourcetracker*****************####


# Extract abundance matrix from the phyloseq object NB you need otus as rows
mixed_OTU1 <- as(otu_table(everything_ps), "matrix")


# Coerce to a data frame
mixed_OTUdf = as.data.frame(mixed_OTU1)
mixed_OTUdf <- cbind("#OTU ID" = rownames(mixed_OTUdf), mixed_OTUdf)
rownames(mixed_OTUdf) <- NULL
View(head(mixed_OTUdf))


#export OTU table for sourcetracker
write.table(mixed_OTUdf, file=’ NAME OTU TABLE', quote=FALSE, sep='\t',row.names = F)
#export mapping file for sourcetracker
newsample_df$Simulated <- NULL
write.table(newsample_df, file='NAME MAPPING FILE’, sep='\t', row.names = F)
```

Appendix D.2 – Results of sewage and sea water communities with and without a background source included in the FTL.

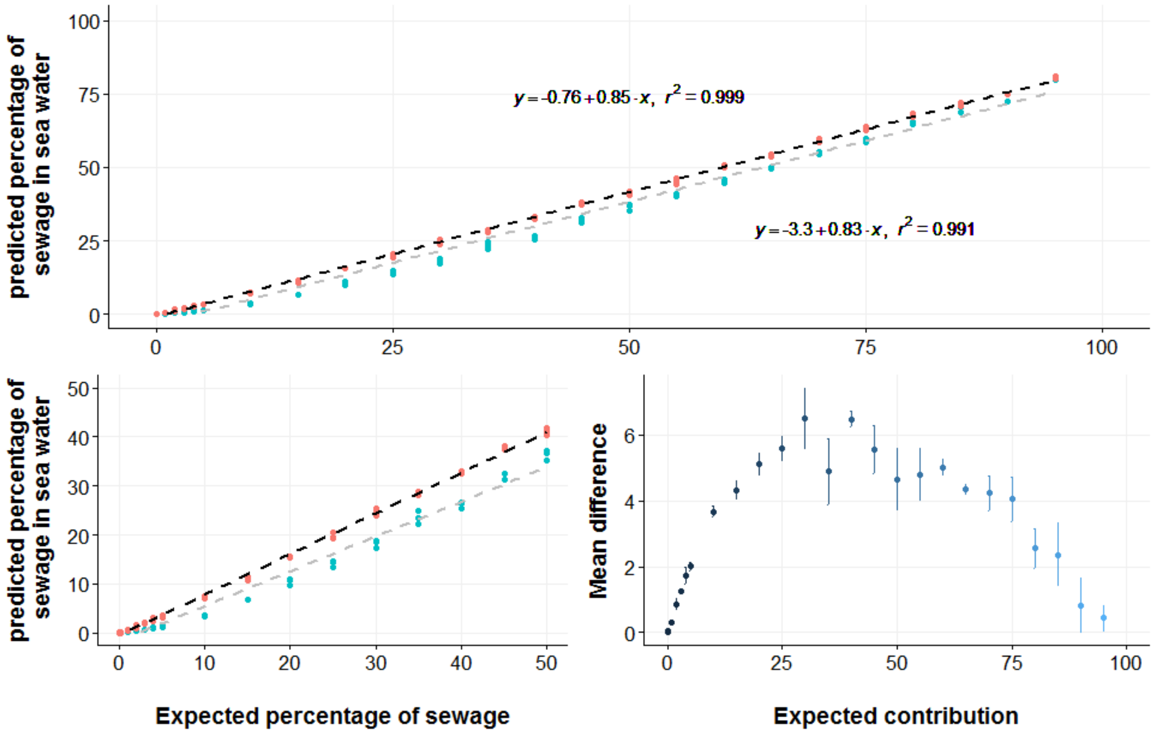


Figure D.2.1 SourceTracker predictions for the detection of sewage in sea water, when no background source was used, or when sea water was included as a background source in the FTL.

Appendix E – Seaton Sluice Catchment Case Study

Appendix E.1 - Environment Agency Bathing Water Sampling Regime 2016

Table E.1.1 Environment Agency Bathing Water Sampling Regime 2016

Day	Date	Sample Number	Sampled?	Used for qPCR?	Used for Sequencing?
Wednesday	04-May-16	Pre-Season	Yes	Yes	Yes
Tuesday	17-May-16	2	Yes	Yes	Yes
Monday	23-May-16	3	Yes	No	No
Thursday	02-Jun-16	4	Yes	No	No
Wednesday	08-Jun-16	5	Yes	No	No
Tuesday	14-Jun-16	6	Yes	No	No
Wednesday	22-Jun-16	7	Yes	No	No
Monday	27-Jun-16	8	Yes	No	No
Saturday	09-Jul-16	9	No	No	No
Thursday	14-Jul-16	10	Yes	Yes	No
Tuesday	19-Jul-16	11	Yes	No	No
Wednesday	27-Jul-16	12	Yes	Yes	Yes
Tuesday	04-Aug-16	13	Yes	Yes	No
Wednesday	10-Aug-16	14	Yes	Yes	No
Tuesday	16-Aug-16	15	No	No	No
Tuesday	23-Aug-16	16	Yes	No	No
Saturday	03-Sep-16	17	Yes	Yes	No
Thursday	08-Sep-16	18	Yes	Yes	No
Wednesday	14-Sep-16	19	Yes	Yes	No
Tuesday	20-Sep-16	20	Yes	Yes	Yes

Appendix E.2 E.coli concentrations across sampling days

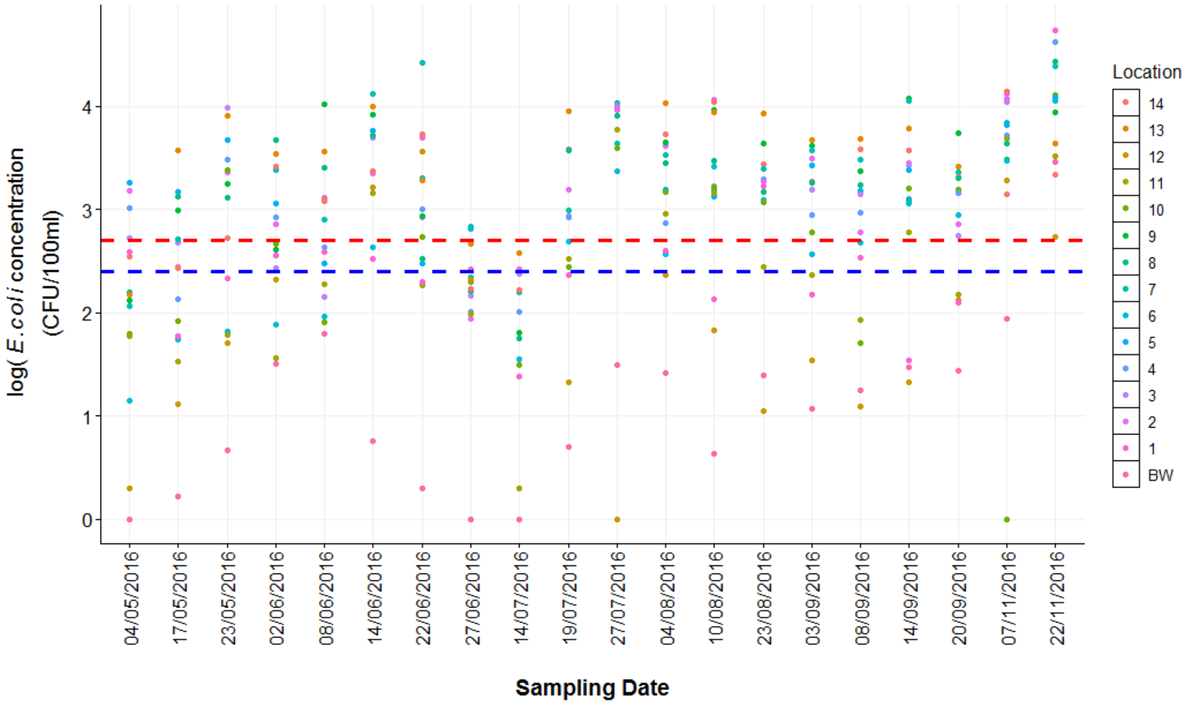


Figure E.2.1. Culturable E.coli shown by sampling date and coloured by sample location. Blue and red dotted line show concentrations of E.coli required to achieve excellent and good status, respectively.

Appendix E.3 Rainfall pattern throughout the bathing water season.

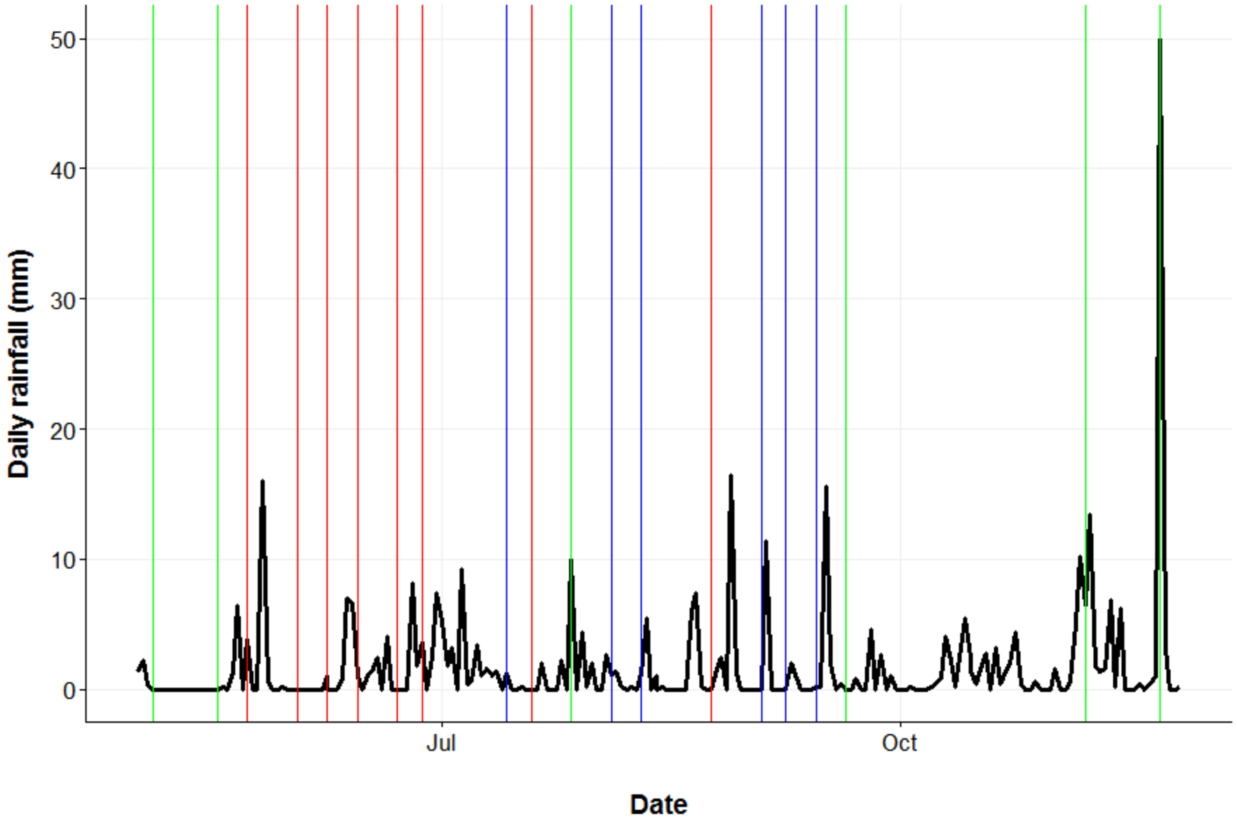


Figure E.3.1. Rainfall across the bathing water season (2016). Vertical lines represent sampling days used for analysis of culturable E.coli (all), enumeration of genetic markers by qPCR (Blue and green), and community analysis (Green).

Appendix E.4 E.coli and marker concentrations and classifications or each sample under the BWD (2006/7/EC)

Table E.4.1 E.coli and marker concentrations and classifications or each sample under the BWD

Location	Indicator	04/05/2016	17/05/2016	14/07/2016	27/07/2016	04/08/2016	10/08/2016	03/09/2016	08/09/2016	14/09/2016	20/09/2016	07/11/2016	22/11/2016	Location
BW	*E.coli*	0.00	0.00	0.00	30.00	25.00	3.33	10.67	16.67	28.67	26.50	86.50	2850.00	BW
	RodA	0.00	0.00	370.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	125.15	6754.72	
	HF183	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	61.41	4748.84	
	Hu100	0.00	0.00	34.26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	15.31	681.59	
1	*E.coli*	383.33	57.73	23.30	9650.00	400.00	133.33	150.00	343.44	33.33	123.33	13000.00	54666.67	1
	RodA	140.87	102.07	125.56	14783.50	1235.51	2532.75	2355.37	853.17	167.93	335.71	15223.74	86571.97	
	HF183	26.65	0.00	250.97	1103.03	0.00	0.00	154.80	36.92	0.00	0.00	1852.56	49204.73	
	Hu100	0.00	0.00	0.00	234.72	0.00	0.00	26.11	0.00	0.00	54.06	1485.77	2726.65	
2	*E.coli*	1500.00	275.00	260.00	9166.66	4166.66	17000.00	3150.00	606.67	2833.33	720.00	12000.00	ND*	2
	RodA	460.21	632.53	671.41	23110.21	4359.45	20037.90	6248.64	3973.88	4839.05	537.01	27350.92		
	HF183	104.36	0.00	147.15	3311.46	0.00	11534.99	0.00	0.00	0.00	55.73	4836.89		
	Hu100	273.53	0.00	0.00	550.61	49.43	976.95	65.42	0.00	0.00	31.69	382.25		
3	*E.coli*	533.33	470.00	236.66	9650.00	900.00	1700.00	1550.00	1400.00	2700.00	550.00	11000.00	24666.67	3
	RodA	166.50	770.42	758.58	17723.54	1595.77	3359.56	5039.71	2136.01	2975.22	429.38	19855.37	69573.81	
	HF183	52.48	28.88	77.46	699.19	0.00	0.00	0.00	0.00	0.00	73.99	7233.97	27055.53	
	Hu100	419.04	190.89	43.99	259.17	59.55	0.00	0.00	49.16	0.00	0.00	683.83	4038.17	
4	*E.coli*	1020.00	135.00	100.00	10500.00	733.33	1600.00	876.67	920.00	1266.67	1433.33	5200.00	42333.33	4
	RodA	553.05	748.25	441.81	13027.54	2476.89	1645.842	2917.51	2548.29	4237.55	1390.51	7845.17	52561.53	

	HF183	384.74	298.38	476.53	2203.22	78.24	494.13	144.51	118.34	127.98	1104.03	1831.09	45244.11	
	Hu100	182.14	43.06	47.46	201.71	111.59	54.86	95.95624	80.52	0.00	207.80	0.00	21221.17	
5	*E.coli*	1800.00	1466.67	55.50	10750.00	1566.00	1333.33	2700.00	1500.00	2400.00	2000.00	6500.00	12333.33	5
	RodA	597.78	1696.49	473.06	20360.00	1607.83	4976.26	4018.95	1596.06	1510.25	1142.41	12678.38	44122.40	
	HF183	689.64	138.27	599.76	2904.12	105.98	0.00	0.00	0.00	0.00	107.99	1880.88	14557.46	
	Hu100	229.84	106.17	266.65	1570.48	0.00	0.00	117.13	28.91	51.75	47.13	438.53	1082.32	
6	*E.coli*	13.00	54.50	34.00	2333.33	366.67	2600.00	366.66	476.67	1150.00	880.00	2966.67	11333.33	6
	RodA	0.00	124.37	406.08	4395.72	637.61	2761.35	4988.44	2641.33	2002.90	1118.16	7031.86	38342.83	
	HF183	0.00	0.00	59.33	0.00	0.00	920.38	0.00	0.00	0.00	0.00	0.00	11543.96	
	Hu100	0.00	0.00	54.69	0.00	0.00	108.08	26.61	0.00	50.51	0.00	0.00	4815.38	
7	*E.coli*	115.47	510.00	155.00	4400.00	3400.00	1466.67	3733.33	3000.00	11200.00	2300.00	7000.00	24333.33	7
	RodA	202.42	2732.44	1593.73	31081.72	24943.97	2596.80	11181.93	2613.37	76592.52	742.31	17010.38	50990.60	
	HF183	0.00	215.10	2833.90	2481.17	425.65	95.76	1338.04	25.22	7748.68	2239.87	6141.77	16524.89	
	Hu100	0.00	182.66	79.51	379.57	507.13	49.35	480.95	37.10	3558.16	259.99	299.60	4175.27	
8	*E.coli*	156.66	1340.00	55.00	8000.00	2800.00	1966.67	1800.00	1733.33	1233.33	2000.00	4366.67	27000.00	8
	RodA	86.40	6977.53	0.00	26495.01	54987.49	7579.24	14492.99	949.18	1271.52	2710.85	9335.02	57234.72	
	HF183	89.76	1806.23	43.07	596.78	0.00	0.00	555.57	28.56	0.00	190.57	1807.42	19555.46	
	Hu100	0.00	557.08	0.00	229.58	230.06	34.19	134.72	0.00	0.00	119.20	1044.26	5739.80	
9	*E.coli*	130.00	42.42	62.50	10650.00	4466.67	9100.00	4100.00	2333.33	11900.00	5500.00	3033.33	8666.67	9
	RodA	353.49	380.78	1853.85	21797.30	41289.77	2812.13	6096.16	8492.42	22028.02	2855.25	7587.26	23276.40	
	HF183	160.37	165.94	0.00	576.55	24.55	220.37	0.00	0.00	120.29	0.00	0.00	6147.40	
	Hu100	0.00	51.21	0.00	252.99	229.66	38.94	0.00	129.39	135.64	215.33	0.00	0.00	
10	*E.coli*	60.66	82.67	30.00	3966.66	1466.66	1650.00	593.33	50.55	1600.00	1550.00	NR[†]	12666.67	10
	RodA	475.76	184.06	1095.48	3149.06	10594.35	2931.00	3279.16	4355.13	1168.05	1520.32	5710.21	37878.21	
	HF183	88.86	0.00	0.00	110.62	0.00	208.02	0.00	0.00	0.00	715.30	0.00	409.76	
	Hu100	0.00	0.00	0.00	44.16	574.60	65.20	0.00	0.00	0.00	204.54	0.00	499.06	
11	*E.coli*	57.73	33.00	1.00	5900.00	900.00	1433.33	233.33	83.33	600.00	146.67	4766.67	3300.00	11
	RodA	0.00	142.04	0.00	10812.49	3105.89	402.32	590.45	688.70	869.35	0.00	10120.75	2888.12	
	HF183	30.16	0.00	0.00	4751.04	0.00	0.00	0.00	0.00	0.00	0.00	545.93	3841.98	

	Hu100	0.00	0.00	0.00	97.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	474.17	
12	*E.coli*	1.00	12.00	0.00	0.00	233.33	66.67	33.33	11.33	20.00	130.00	1900.00	535.00	12
	RodA	0.00	0.00	471.88	0.00	10475.96	0.00	402.32	515.59	152.35	259.46	3401.41	2301.88	
	HF183	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	397.56	2924.50	
	Hu100	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	139.42	
13	*E.coli*	150.00	3766.00	380.00	10700.00	10750.00	8800.00	4666.67	4850.00	6066.67	2633.33	14000.00	4333.33	13
	RodA	2251.14	2499.84	3284.29	14803.49	7392.54	10885.12	1802.59	276.28	15465.20	11762.57	35714.66	23123.61	
	HF183	5239.52	13251.89	972.84	8554.91	0.00	2730.56	675.48	86.31	3196.83	4593.31	7755.50	8318.99	
	Hu100	51.53	0.00	28.64	144.97	37.15	352.45	31.56	39.11	204.93	81.46	97.55	679.63	
14	*E.coli*	346.67	270.00	166.66	9350.00	5333.33	11000.00	1866.67	3800.00	3733.33	2050.00	1400.00	2166.67	14
	RodA	3565.13	3135.72	542.61	477.87	33105.37	NR**	4101.12	14929.06	7129.01	4075.93	27500.33	24076.94	
	HF183	3475.49	31649.54	983.06	9692.56	1928.13		1053.89	156.95	4087.24	5174.70	1759.59	2234.74	
	Hu100	60.37	62.26	0	129.37	133.04		44.13	0.00	198.30	62.31	0.00	276.71	

Highlighted cells denote concentrations between the limit of detection and quantification.

*ND - Not determined as sample could not be collected safely due to surface water flooding.

†NR - Not reported as all replicate plates were unreadable.

**NR - Not reported as all qPCR assays appeared inhibited and dilution resulted in the gene copies decreasing below the limit of detection.

Table E.4.2 Classifications of each sample according to the Bathing Water Directive (2006/7/EC)

Location	Indicator	04/05/2016	17/05/2016	14/07/2016	27/07/2016	02/08/2016	10/08/2016	03/09/2016	08/09/2016	14/09/2016	20/09/2016	07/11/2016	22/11/2016	Location	
BW	*E.coli*	Excellent	Excellent	Excellent	Excellent	Excellent	Excellent	Excellent	Excellent	Excellent	Excellent	Excellent	Poor	BW	*E.coli*
	RodA	Excellent	Excellent	Good	Excellent	Excellent	Excellent	Excellent	Excellent	Excellent	Excellent	Excellent	Poor		*RodA*
1	*E.coli*	Good	Excellent	Excellent	Poor	Good	Excellent	Excellent	Good	Excellent	Excellent	Poor	Poor	1	*E.coli*
	RodA	Excellent	Excellent	Excellent	Poor	Poor	Poor	Poor	Poor	Excellent	Good	Poor	Poor		*RodA*
2	*E.coli*	Poor	Good	Good	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor		2	*E.coli*
	RodA	Good	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor			*RodA*
3	*E.coli*	Poor	Good	Excellent	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	3	*E.coli*
	RodA	Excellent	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Good	Poor	Poor		*RodA*
4	*E.coli*	Poor	Excellent	Excellent	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	4	*E.coli*
	RodA	Poor	Poor	Good	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor		*RodA*
5	*E.coli*	Poor	Poor	Excellent	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	5	*E.coli*
	RodA	Poor	Poor	Good	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor		*RodA*
6	*E.coli*	Excellent	Excellent	Excellent	Poor	Good	Poor	Good	Good	Poor	Poor	Poor	Poor	6	*E.coli*
	RodA	Excellent	Excellent	Good	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor		*RodA*
7	*E.coli*	Excellent	Poor	Excellent	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	7	*E.coli*
	RodA	Excellent	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor		*RodA*
8	*E.coli*	Excellent	Poor	Excellent	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	8	*E.coli*
	RodA	Excellent	Poor	Excellent	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor		*RodA*
9	*E.coli*	Excellent	Excellent	Excellent	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	9	*E.coli*
	RodA	Good	Good	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor		*RodA*
10	*E.coli*	Excellent	Excellent	Excellent	Poor	Poor	Poor	Poor	Excellent	Poor	Poor		Poor	10	*E.coli*
	RodA	Good	Excellent	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor		*RodA*
11	*E.coli*	Excellent	Excellent	Excellent	Poor	Poor	Poor	Excellent	Excellent	Poor	Excellent	Poor	Poor	11	*E.coli*
	RodA	Excellent	Excellent	Excellent	Poor	Poor	Good	Poor	Poor	Poor	Excellent	Poor	Poor		*RodA*
12	*E.coli*	Excellent	Excellent	Excellent	Excellent	Excellent	Excellent	Excellent	Excellent	Excellent	Excellent	Poor	Poor	12	*E.coli*
	RodA	Excellent	Excellent	Good	Excellent	Poor	Excellent	Good	Poor	Excellent	Good	Poor	Poor		*RodA*
13	*E.coli*	Excellent	Poor	Good	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	13	*E.coli*
	RodA	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Good	Poor	Poor	Poor	Poor		*RodA*
14	*E.coli*	Good	Good	Excellent	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	14	*E.coli*
	RodA	Poor	Poor	Poor	Good	Poor		Poor	Poor	Poor	Poor	Poor	Poor		*RodA*

Appendix E.5 CSO spills up to 12 hours prior to sampling

Table 9E.5.1. CSO spill data up to 12 hours prior to sampling.

PLR	Site name	Spill duration	Spill start time
NZ30747507	05D01SEATONVALLEYCSO07 SewerLevel	00 00:15	22/11/2016 12:15
NZ30747507	05D01SEATONVALLEYCSO07 SewerLevel	00 00:15	22/11/2016 11:00
NZ27766901	02D02CRAMLINGTONCSO03 SewerLevel	00 00:15	22/11/2016 10:15
NZ23724809	05D01BRUNSWICKCRESCSO SewerLevel	00 00:15	22/11/2016 09:00
NZ23724809	05D01BRUNSWICKCRESCSO SewerLevel	00 00:15	22/11/2016 08:30
NZ23724809	05D01BRUNSWICKCRESCSO SewerLevel	00 00:15	22/11/2016 07:30
NZ26735701	05D01DUDLEYSCHOOLCSO SewerLevel	00 01:45	22/11/2016 07:15
NZ26738505	05D01FORDLEYDRIVECSO SewerLevel	00 03:45	22/11/2016 07:00
NZ23724809	05D01BRUNSWICKCRESCSO SewerLevel	00 00:30	22/11/2016 06:30
NZ25738406	05D01SEATONVALLEYCSO23 SewerLevel	00 00:45	22/11/2016 06:30
NZ26735701	05D01DUDLEYSCHOOLCSO SewerLevel	00 00:30	22/11/2016 06:15
NZ23724809	05D01BRUNSWICKCRESCSO SewerLevel	00 00:15	22/11/2016 06:00
NZ28743504	05D01SEATONVALLEYCSO16 SewerLevel	00 00:30	22/11/2016 06:00
NZ31741610	05D01NORTHSIDEPLACECSO SewerLevel	00 00:45	22/11/2016 05:45
NZ25738406	05D01SEATONVALLEYCSO23 SewerLevel	00 00:30	22/11/2016 05:45
NZ27766901	02D02CRAMLINGTONCSO03 SewerLevel	00 01:45	22/11/2016 05:30
NZ20731303	05D01DINNINGTON SewerLevel	00 07:30	22/11/2016 05:30
NZ26735701	05D01DUDLEYSCHOOLCSO SewerLevel	00 00:45	22/11/2016 05:15
NZ30749413	05D01SEATONVALLEYCSO09 SewerLevel	00 02:00	22/11/2016 05:15
NZ28743504	05D01SEATONVALLEYCSO16 SewerLevel	00 00:30	22/11/2016 05:00
NZ27766901	02D02CRAMLINGTONCSO03 SewerLevel	00 00:30	22/11/2016 04:30
NZ30749413	05D01SEATONVALLEYCSO09 SewerLevel	00 00:15	22/11/2016 04:30
NZ26738716	05D01SEATONVALLEYCSO31 SewerLevel	00 04:30	22/11/2016 04:30
NZ27766901	02D02CRAMLINGTONCSO03 SewerLevel	00 00:15	22/11/2016 04:00
NZ33758710	05D01SEATONVALLEYCSO14 SewerLevel	00 00:15	22/11/2016 03:00
NZ26735701	05D01DUDLEYSCHOOLCSO SewerLevel	00 02:15	22/11/2016 02:45
NZ33766703	05D01SEATONVALLEYCSO11 SewerLevel	00 00:15	22/11/2016 02:45
NZ25738406	05D01SEATONVALLEYCSO23 SewerLevel	00 00:30	22/11/2016 02:30
NZ33758710	05D01SEATONVALLEYCSO14 SewerLevel	00 00:15	22/11/2016 02:15
NZ33766703	05D01SEATONVALLEYCSO11 SewerLevel	00 00:15	22/11/2016 02:00
NZ28743504	05D01SEATONVALLEYCSO16 SewerLevel	00 00:15	22/11/2016 02:00
NZ27766901	02D02CRAMLINGTONCSO03 SewerLevel	00 02:00	22/11/2016 01:45
NZ30749413	05D01SEATONVALLEYCSO09 SewerLevel	00 02:45	22/11/2016 01:30
NZ30749413	05D01SEATONVALLEYCSO09 SewerLevel	00 00:15	22/11/2016 01:00
NZ23724809	05D01BRUNSWICKCRESCSO SewerLevel	00 03:00	22/11/2016 00:45
NZ27766901	02D02CRAMLINGTONCSO03 SewerLevel	00 00:30	22/11/2016 00:30
NZ25738406	05D01SEATONVALLEYCSO23 SewerLevel	00 00:30	21/11/2016 23:45
NZ27766901	02D02CRAMLINGTONCSO03 SewerLevel	00 00:45	21/11/2016 23:30

NZ30749413	05D01SEATONVALLEYCSO09 SewerLevel	00 02:00	21/11/2016 22:45
NZ25738406	05D01SEATONVALLEYCSO23 SewerLevel	00 00:30	21/11/2016 22:45
NZ30749413	05D01SEATONVALLEYCSO09 SewerLevel	00 00:15	21/11/2016 22:15
NZ26735701	05D01DUDLEYSCHOOLCSO SewerLevel	00 04:30	21/11/2016 22:00
NZ30747507	05D01SEATONVALLEYCSO07 SewerLevel	00 00:15	07/11/2016 12:45
NZ30777106	05D01SEATONVALLEYCSO30 SewerLevel	00 00:45	07/11/2016 12:45
NZ31770604	05D01SEATONVALLEYCSO04 SewerLevel	00 08:00	07/11/2016 10:30
NZ30747507	05D01SEATONVALLEYCSO07 SewerLevel	00 00:15	07/11/2016 10:30
NZ30777106	05D01SEATONVALLEYCSO30 SewerLevel	00 00:30	07/11/2016 10:30
NZ23739410	05D01SEATONVALLEYCSO27 SewerLevel	00 01:00	07/11/2016 10:15
NZ33758710	05D01SEATONVALLEYCSO14 SewerLevel	00 00:15	07/11/2016 09:45
NZ23739410	05D01SEATONVALLEYCSO27 SewerLevel	00 01:30	10/08/2016 23:30
NZ27791208	02D02EASTHARTFORDCSO SewerLevel	00 01:00	27/07/2016 12:30
NZ30747507	05D01SEATONVALLEYCSO07 SewerLevel	00 00:15	27/07/2016 04:30
NZ30777106	05D01SEATONVALLEYCSO30 SewerLevel	00 00:30	27/07/2016 04:30
NZ23739410	05D01SEATONVALLEYCSO27 SewerLevel	00 01:30	27/07/2016 03:45
NZ23739410	05D01SEATONVALLEYCSO27 SewerLevel	00 00:45	27/07/2016 02:15
NZ30747507	05D01SEATONVALLEYCSO07 SewerLevel	00 00:15	19/07/2016 10:15
NZ29742907	05D01SEATONVALLEYCSO17 SewerLevel	00 00:30	22/05/2016 21:45

Appendix E.6 Hu100 concentrations in CSO impacted samples and non-CSO impacted samples.

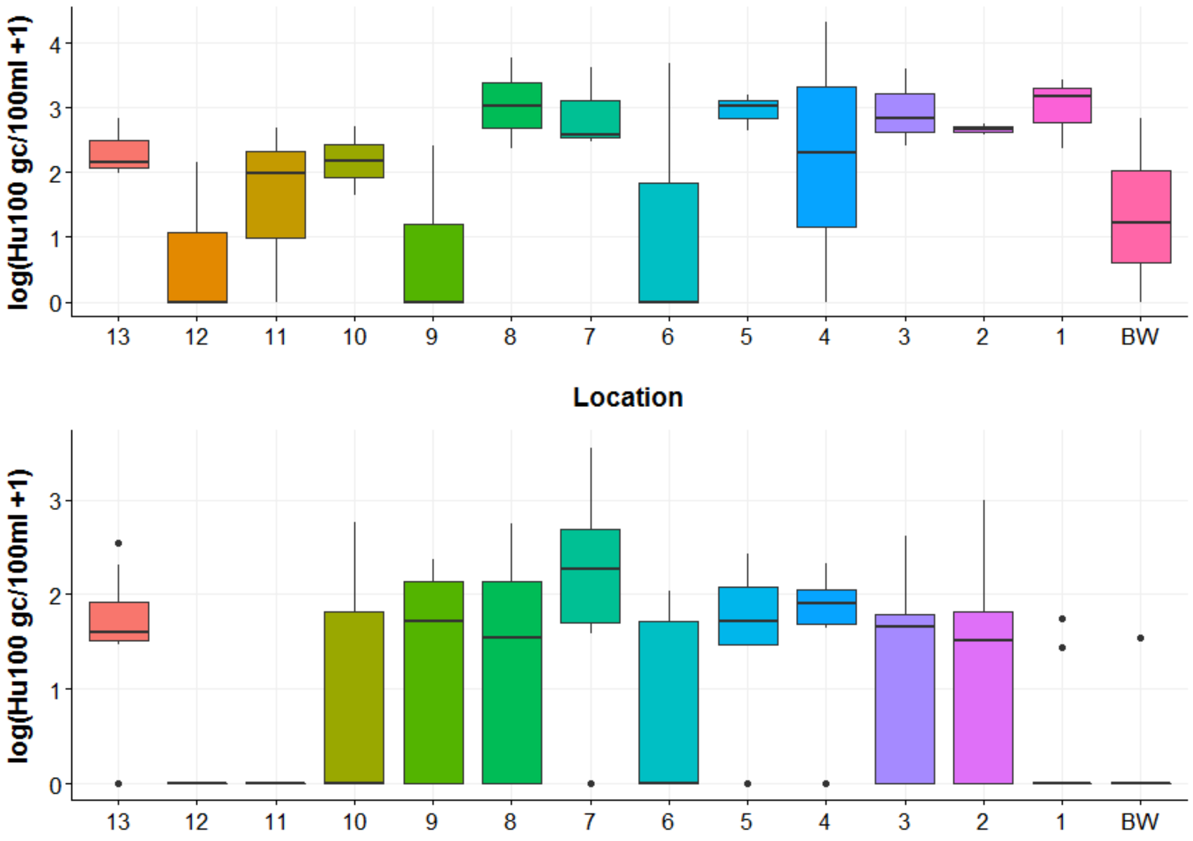


Figure E.6.1. HU100 concentrations for CSO impacted (top) and non-CSO impacted (bottom) samples.