Next Generation Sequencing for the Water Industry

Gregg Iceton

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

September 2016

School of Civil Engineering and Geosciences

Abstract

The wastewater industry uses biotechnology to ensure that the discharge of sewage does not have deleterious effects on the environment, yet knowledge of the underlying microbiology is poor. This leads to over engineered and inefficient processes which occasionally and unexpectedly fail. Similarly the impact of sewage on the microbiology of receiving waters is unclear. Recent developments in DNA sequencing have enabled its use where cost was prohibitive. I investigated two applications of Next Generation Sequencing (NGS); activated sludge process monitoring for nitrification, foaming and bulking, and microbial source tracking of faecal contamination in bathing waters.

Samples from 32 activated sludge plants (ASPs) were collected and analysed. Cell specific ammonia oxidation rates were calculated using the equation $CSAOR = \frac{(A \times M \times 10^6) \times r}{AOB \times MLSS \times V}$ where A = grams of ammonia oxidised, M = the number of moles of ammonia in a gram, r = correction factor of 0.9 due to some ammonia removal by adsorption and assimilation (Daims, Ramsing, et al. 2001), MLSS = mixed liquor suspended solids in mg/L and V = the volume of the aeration basin in litres. The CSAOR in nitrifying plants ranged from one to ten mmol / cell / hour, in agreement with other CSAOR studies using alternative techniques.

Biological foaming in ASPs occurs when the abundance of filamentous bacteria with hydrophobic surface membranes becomes excessive, though the exact abundance threshold above which foaming occurs has not yet been established. The relative abundance of bacteria associated with foaming was measured for all ASPs which were then categorised as non-foaming, occasionally-foaming or currently-foaming based on operator assessment. There was a significant difference in the abundance of foaming bacteria between non-foaming and occasionally foaming plants (ANOVA $p < 0.001$), with all non-foaming plants having less than 1% relative abundance of foaming bacteria. These results demonstrate that NGS could be a useful ASP process monitoring tool.

A bathing water catchment was sampled throughout a bathing season, including a storm event. Partial least squares analysis showed there was a significant correlation

between faecal indicator bacteria and the cumulative apportioned fraction of sources (using Bayesian statistics) in the bathing water community ($p < 0.001$, $r^2 = 87\%$). Faecal host marker analysis detected human contamination upstream of any wastewater network inputs, illustrating the impact of diffuse human pollution. Whole community analysis apportioned the bathing water microbial community to point and diffuse sources, and found that whilst human sources were dominant during storm conditions, in dry weather the primary source of faecal contamination was variable and in some cases could not be attributed to known faecal sources.

Acknowledgements

I thank my supervisor Dr Russell Davenport and the wider Environmental team. In particular Donna, Sarah, Dave and Fiona who taught me so much and looked on with only gentle disdain as I made my many mistakes. Thanks to the postdocs, especially Paola Meynet, keeper of all things useless who calls me “cowboy”. I am grateful to all of the admin staff who tolerated my terrible memory and poor grasp of the various perplexing forms. Thanks also to Tania, Ewan and Paul for setting up STREAM so well. It’s been a fantastic programme and I am eternally grateful. Highest honours, however, are reserved for the inimitable Justine Easten. Her unending sense of dedication and genuine concern for our wellbeing and success marks her out as a star like no other. Without her I have no doubt that the STREAM programme at Newcastle would be far less successful and we students would certainly be worse off. I fear she fails to recognise her brilliance and worse still she doesn’t often get the recognition she deserves. So here’s a heartfelt thank you for everything you’ve done for me.

To the staff at NWL – Andrew who let me in in the first place, Chris and the rest of the innovation team, as well as the various operators and team leaders that put up with me on site. My thanks to the sampling team who were instrumental in the success of the bathing water work, especially Ian and Chris. I will always remember the unique “humour” from the chod particle posse which helped me feel welcome. Bernie whose bear trap mind knew everything and remains my idea of a true “feminist”. Derek who both protected and pushed me throughout. Lastly but most importantly, Rheanne for putting up with my OCD nature – sorry for all that ethanol you had to filter – and for always being helpful and accommodating. I miss Scientific Support and wish you all the best.

Of course none of this would have been possible without my family. Mam, Dad and Rola, all of whom encourage me to aim high yet accept it with grace when I fall short. Thank you for your guidance and support. But most of all, thanks go to my amazing wife. She quietly destroys every challenge she faces, providing our wonderful little family with its every need. They talk of standing on the shoulders of giants. She truly is a giant amongst men and I can only hope she comes to value herself the way we do. In the end, all you need is faith, trust, and pixie dust.

Table of Contents

Chapter 1. Introduction 1

1.1 Aim and Objectives 2

Chapter 2. Molecular Techniques Overview 4

2.1 Isolation and enrichment of bacterial cells 4

2.2 Polymerase chain reaction 5

2.2.1 Quantitative PCR 6

2.3 Fluorescence In Situ Hybridization 7

2.4 Sequencing 7

2.4.1 The 16S ribosomal RNA gene 7

2.4.2 Sequencing platforms 8

2.5 Bioinformatics 9

2.5.1 Pipeline choice 9

2.5.2 OTU picking 10

2.5.3 Taxonomy assignment 11

2.5.4 Databases 11

2.5.5 Chimeras 12

Chapter 3. Activated Sludge Process Monitoring Literature Review 14

3.1 Introduction 14

3.2 Carbonaceous activated sludge plants 16

3.3 Nitrifying activated sludge plants 17

3.3.1 Ammonia oxidising bacteria 18

3.3.2 Nitrite oxidising bacteria 20

3.4 Foaming 21

3.5 Bulking 24

3.6 Control of foaming and bulking 27

3.7 Removal of micropollutants 28
3.8 Operational utility of molecular methods in activated sludge process control 29
Chapter 4. Microbial Source Tracking Literature Review 30
4.1 Concept 30
4.2 Legislation 31
4.3 Techniques 33
4.3.1 Phenotype analysis 33
4.3.2 Sterols 34
4.3.3 Mitochondria 34
4.3.4 Host associated bacteria 35
4.3.5 Whole community 37
4.3.6 Whole genome 39
4.3.7 Host associated viruses 40
4.3.8 Phage plaques 41
Chapter 5. Methodological development 42
5.1 Sampling 42
5.1.1 Samples taken from the bathing water 49
5.1.2 Other aqueous samples 49
5.1.3 Samples from solid matrices 49
5.1.4 Membrane filtration and sample processing 53
5.2 Primer selection 53
5.3 DNA extraction comparison 56
5.4 PCR, clean up and sequencing protocol 57
5.5 Bioinformatics and selection of a taxonomic database (DB) 58
5.6 Statistical analyses 62
5.6.1 UniFrac phylogenetic distance 62

5.6.2 Beta significance tests 62

5.6.3 Accurate estimation of AOB abundance 62

5.6.4 Comparison of AOB abundance as measured by qPCR, FISH and PGM... 63

5.6.5 Calculation of cell specific ammonia oxidation rate (CSAOR) 63

5.6.6 Correlation statistics 63

5.6.7 Source Tracker 64

5.7 Primer selection 65

5.7.1 How does primer pair selection affect observed microbial community structure? 65

5.7.2 Comparison of the abundance of specific taxa-groups detected by each primer 68

5.7.3 Comparison of the taxonomic resolution observed using each primer... 72

5.7.4 Primer selection: Discussion 76

5.8 DNA extraction and recovery method selection 77

5.8.1 Comparison of DNA extraction and recovery methods 78

5.8.2 Comparison of the abundance of specific taxa-groups across DNA extraction and recovery methods 80

5.8.3 DNA extraction and recovery method: Discussion..... 86

5.9 Assessing the use of NGS for the quantification of process-important taxonomic groups 87

5.9.1 Comparing quantification of AOB by NGS against other quantitative molecular tools..... 87

5.9.2 Evaluating the effect of sampling depth on precision of AOB abundance measurements 90

5.9.3 NGS AOB enumeration: Discussion 91

5.10 Microbial source tracking method development 92

5.10.1 NGS enumeration of faecal indicator bacteria: Results 92

5.11 NGS enumeration of faecal indicator bacteria: Discussion 94

5.11.1 Faecal host marker sensitivity and specificity: Results 96

5.11.2 Faecal host marker sensitivity and specificity: Discussion 98

5.11.3 Change in CSO microbial communities over time and space: Results 99

5.11.4 Change in CSO microbial communities over time and space: Discussion 101

Chapter 6. Activated Sludge Process Monitoring Results 102

6.1 Overview of taxa in activated sludge plants 102

6.2 Nitrification and the abundance of ammonia oxidising bacteria (AOB)..... 103

6.3 The abundance of foam causing bacteria and the incidence of foaming 103

6.4 The abundance of bacteria associated with bulking and the incidence of bulking..... 104

Chapter 7. Microbial source tracking results 107

7.1 Faecal indicator and host marker bacteria 107

7.1.1 Correlation of faecal host markers across the catchment with respect to bathing water faecal indicator bacteria 107

7.2 SourceTracker 108

7.3 Faecal contamination sources 111

Chapter 8. Discussion 115

8.1 Using NGS to diagnose and predict activated sludge operational issues..... 115

8.1.1 Nitrification monitoring by NGS AOB enumeration 115

8.1.2 Prediction of ASP foaming by NGS 116

8.1.3 Prediction of ASP bulking by NGS..... 118

8.2 NGS for microbial source tracking 119

8.2.1 What does the study tell us about the function of the catchment during dry and wet weather 119

8.2.2 The impact of NWL water company assets versus other human and non-human sources 120

Chapter 9. Further Work 122

9.1 Further NGS method optimisation 122

9.2 NGS for ASP control 122

9.3 NGS for microbial source tracking 124

Chapter 10. Business case 127

10.1 NGS monitoring of ASPs 127

10.2 NGS for microbial source tracking 129

Chapter 11. Conclusions 132

Chapter 12. References 134

Chapter 13. Appendix 159

13.1 Protocols 159

13.1.1 Manufacturer's protocol for Ion Torrent instruments 159

13.1.2 Manufacturer's standard DNA extraction protocol for the FastDNA Spin Kit for Soil 159

13.1.3 Ratio of Ampure XP to PCR products used 160

13.1.4 Total cell counts 160

13.2 Additional data from qPCR/FISH/PGM comparison 164

Acronyms and nomenclature

ANOVA	Analysis of variance
AOB	ammonia oxidising bacteria
ASM	activated sludge model
ASP	activated sludge plant
BLAST	basic local alignment search tool
BOD	biological oxygen demand
CB	currently bulking
CF	currently foaming
CFU	colony forming unit
CMC	critical micellar concentration
COD	chemical oxygen demand
CSAOR	cell specific ammonia oxidation rate
CSO	combined sewer overflow
DNA	deoxyribonucleic acid
DO	disolved oxygen
EA	Environment Agency
EBPR	enhanced biological phosphorus removal
EDC	endocrine disrupting compounds
EEC	European Economic Council
EU	European Union
FIB	Faecal indicator bacteria
FISH	fluoresence in situ hybridization
GC	as in the DNA bases guanine and cytosine
MAFFT	multiple alignment using fast fourier transform
Mbp	mega base pair
MST	microbial source tracking
MUSCLE	multiple sequence alignment by log expectation
NB	non bulking
NF	non foaming
NGS	Next generation sequencing
NOB	nitrite oxidising bacteria
NWL	Northumbrian Water limited
OB	occasionally bulking
OF	occasionally foaming
OFWAT	Office of water services, the economic regulator
OTU	operational taxonomic unit
PCoA	principal coordinate analysis
PCR	polymerase chain reaction
PGM	personal genome machine
PLS	partial least squares
PR	public relations
PyNAST	python nearest alignment space termination
QIIME	quantitative insights into microbial ecology
qPCR	quantitative polymerase chain reaction
RAxML	randomized axelerated maximum liklihood
RBCOD	readily biodegradable chemical oxygen demand

RDP	ribosomal database project
RNA	ribonucleic acid
rRNA	ribosomal ribonucleic acid
SILVA	from the latin silva which means forest
STP	sewage treatment plant
UV	ultra violet
UWWTD	urban wastewater treatment directive
V4/V5/V6	16S rRNA gene hypervariable region 4 etc

Table of Figures

Figure 1: Flow diagram showing bioinformatics steps when processing 16S rRNA data. From https://sites.google.com/site/knightslabwiki/qiime-workflow..... 13

Figure 2: Schematic of the activated sludge secondary sewage treatment process. Raw sewage is first settled in a primary clarifier to remove solids before being pumped in to the aeration tank. The activated sludge microbial community then oxidises complex organic matter and in some cases ammonia. The suspended sludge and treated sewage is then pumped to the final settlement tank where the activated sludge settles out and the supernatant is discharged as final effluent. The majority of the settled activated sludge is pumped back to the aeration tank, with the remainder "wasted" out of the system. The wastage rate influences the mean sludge age and is a key control parameter. Image from http://www.staffs.ac.uk/schools/sciences/consultancy/dladmin/zCIWEMWWT/Activity5/Images/bigasp.jpg 16

Figure 3: Samples from several nitrifying activated sludge plants where ammonia oxidising bacteria have been quantified using fluorescence in-situ hybridization. Plant operators classified the plants as stable (no failures in 12 months), unstable (at least one failure in 12 months) or failing (consistently failing) where failure is final effluent ammonia concentration exceeding consent. From Pickering et al (2008)..... 20

Figure 4: Intra-laboratory variability of viral concentration methods in artificial seawater and freshwater. Virus recovery values obtained after spiking sets of ten 10-l samples with HAdV 2, concentrating by: Method 1: electronegative filters of nitrocellulose and glycine 0.05 M pH 9.5 – skimmed milk buffer. Method 2: electronegative filters of nitrocellulose and glycine 0.25 M pH 9.5 – beef extract buffer. Method 3: a column of glass wool and glycine 0.25 M pH 9.5-beef extract buffer. Method 4: Direct organic flocculation with skimmed milk and quantifying the recovery by qPCR according to Bofill-Mas et al. (2006). Taken from Girones et al (2010)..... 40

Figure 5: Map showing location of activated sludge plants sampled 43

Figure 6: Schematic of the Saltburn catchment and NWL assets. Triangle = pumping station. Rectangle = combined sewer overflow (CSO). Large circle = storage tank.

Small circle = outfall. Solid line = sewer. Dashed line = rising main. Adapted from MWH Saltburn bathing water report 2012. 47

Figure 7: Map showing the location of sampling points in the Pit Hills Stell, Skelton Beck and Saltburn Gill catchments, as well as faecal samples from across the North of England. Pins show a selection of conurbations. Red triangles show the sampling points with their respective identification code (see Table 6)..... 50

Figure 8: Flow diagram showing sample processing sequence 58

Figure 9: Two dimensional Procrustes analyses plot showing the weighted and unweighted UniFrac beta diversity distance between three activated sludge samples amplified by polymerase chain reaction. Sample A2 is from Tudhoe Mill activated sludge plant (ASP) which treats domestic sewage. Sample A4 is from Aycliffe ASP which treats heavy industrial and domestic sewage. Sample A10 is from Browney ASP which treats light industrial and domestic sewage. Each sample was PCR amplified using primers targeting the V3, V4, V5 and V4V5 hypervariable regions of the 16S rRNA gene. 66

Figure 10: Weighted UniFrac beta diversity principal coordinate analysis plots showing samples from Tudhoe Mill (A2) nitrifying activated sludge plant (ASP) treating domestic sewage, Aycliffe (A4) ASP treating heavy industrial and domestic sewage and Browney (A10) ASP treating light industrial and domestic. Figure a shows the microbial community structure identified using amplicons from the V3 region, figure b from the V4 region, figure C from the V5 region and figure d from the V4V5 tandem regions... 67

Figure 11: Taxa summary of amplicon libraries amplified from microbial communities using primers targeting the variable V3, V4 and V5 regions of the 16S rRNA gene of Bacteria and Archaea. Each sample label, A2, A4 and A5 (representing Tudhoe Mill nitrifying plant no industrial influent, Aycliffe nitrifying plant with a large proportion of industrial influent and Browney nitrifying plant with a small proportion industrial influent) has a post-fix label representing the respective variable region. Each colour represents OTUs at 97% similarity cut off grouped according to taxa identity, where colours represent equivalent taxa in each sample. 68

Figure 12: Boxplots a - d show the total abundance of indicated bacterial groups in samples from Aycliffe industrial nitrifying activated sludge plant (ASP), Tudhoe Mill

nitrifying ASP and Browney nitrifying ASP. One sample from each plant was PCR amplified using V3, V4 and V5 primers producing nine amplicon pools. Each pool was sequenced and the data clustered in to operational taxonomic units (OTUs) at the 97% cut off level and assigned to a taxonomy. Those taxa known to be ammonia oxidisers, nitrite oxidisers, involved in foaming or bulking were identified and their cumulative abundance recorded. 71

Figure 13: Number of operational taxonomic units at 97% cut-off identified to species level..... 73

Figure 14: Boxplots e – h show the number of OTUs identified to species level in a sample from Aycliffe industrial nitrifying activated sludge plant, Tudhoe Mill nitrifying activated sludge plant and Browney nitrifying activated sludge plant. One sample from each plant was PCR amplified using V3, V4 and V5 primers and taxonomy assigned to each amplicon sequenced. The number of OTUs to which a species had been assigned was recorded. 75

Figure 15: Taxonomic summary of three activated sludge samples extracted using three DNA extraction methods. Columns labelled Z.X where Z = the sample and X = the extraction method. Sample 1 was Tudhoe Mill nitrifying plant treating domestic sewage and prone to foaming. Sample 2 was Amble non-nitrifying sequencing batch reactor treating domestic sewage. Sample 3 was from Aycliffe nitrifying plant treating domestic and light industrial sewage. Method 1 was the manufacturer's protocol. Method 2 increased bead beating duration to 160 seconds. Method 3 included a final incubation of DNA binding matrix at 55 degrees centigrade. Each colour band represents a taxon and the height of the band represents its relative abundance. 78

Figure 16: DNA extract concentration in nanograms per microliter where one sample from each of 3 ASPs was extracted using three methods; Method 1: standard manufacturer's protocol, Method 2: standard protocol with increased velocity and duration of bead-beating compared to method 1 (6 ms-1 for 160 secs), Method 3: as Method 2 with the addition of a 5 minute incubation of DNA binding matrix at 55 °c. The sampled sites were Tudhoe Mill nitrifying domestic activated sludge plant (ASP), Amble non-nitrifying domestic sequencing batch reactor ASP, and Aycliffe nitrifying ASP treating influent with some domestic and significant industrial effluent. 80

Figure 17: Relative abundance of foaming bacteria in a sample from three activated sludge plants using standard manufacturer's protocol (Method 1), standard protocol with bead beating at 6 ms-1 for 160 seconds (Method 2), and Method 2 with the addition of a 5 minute incubation of DNA binding matrix at 55 °c (Method 3). Sample 1 was from Tudhoe Mill nitrifying domestic activated sludge plant (ASP), sample 2 was from Amble non-nitrifying domestic sequencing batch reactor ASP and sample 3 was from Aycliffe nitrifying ASP treating influent with some domestic and significant industrial effluent. Each sample was extracted using each method producing 3 groups of 3 DNA extracts. The bottom line of each box is the first quartile, the middle line the median, the top line the third quartile and the cross in a circle is the mean. 82

Figure 18: Relative abundance of bulking bacteria in a sample from three activated sludge plants using standard manufacturer's protocol (Method 1), standard protocol with bead beating at 6 ms-1 for 160 seconds (Method 2), and Method 2 with the addition of a 5 minute incubation of DNA binding matrix at 55 °c (Method 3). Sample 1 was from Tudhoe Mill nitrifying domestic activated sludge plant (ASP), sample 2 was from Amble non-nitrifying domestic sequencing batch reactor ASP and sample 3 was from Aycliffe nitrifying ASP treating influent with some domestic and significant industrial effluent. Each sample was extracted using each method producing 3 groups of 3 DNA extracts. The bottom line of each box is the first quartile, the middle line the median, the top line the third quartile and the cross in a circle is the mean. 82

Figure 19: Relative abundance of ammonia oxidising bacteria extracted by the manufacturer's protocol (Method 1), standard protocol with bead beating at 6 ms-1 for 160s (Method 2), and Method 2 plus 55 degree c incubation of DNA binding matrix (Method 3). Sample 1 was from Tudhoe Mill nitrifying domestic activated sludge plant (ASP), sample 2 was from Amble non-nitrifying domestic sequencing batch reactor ASP and sample 3 was from Aycliffe nitrifying ASP treating influent with some domestic and significant industrial effluent. Each sample was extracted using each method producing 3 groups of 3 DNA extracts. The bottom line of each box is the first quartile, the middle line the median, the top line the third quartile and the cross in a circle is the mean. 83

Figure 20: Relative abundance of nitrite oxidising bacteria in a sample from three activated sludge plants using standard manufacturer's protocol (Method 1), standard

protocol with bead beating at 6 ms-1 for 160 seconds (Method 2), and Method 2 with the addition of a 5 minute incubation of DNA binding matrix at 55 °c (Method 3). Sample 1 was from Tudhoe Mill nitrifying domestic activated sludge plant (ASP), sample 2 was from Amble non-nitrifying domestic sequencing batch reactor ASP and sample 3 was from Aycliffe nitrifying ASP treating influent with some domestic and significant industrial effluent. Each sample was extracted using each method producing 3 groups of 3 DNA extracts. The bottom line of each box is the first quartile, the middle line the median, the top line the third quartile and the cross in a circle is the mean. 83

Figure 21: Principal coordinate analysis plot showing weighted UniFrac phylogenetic distance (beta diversity) between samples and DNA extraction methods. Method 1 was the standard manufacturer's protocol, method 2 was as method one with bead beating at 6 ms-1 for 160 seconds, method 3 was as method 2 with the addition of a 5 minute incubation of the DNA binding matrix at 55 °c 85

Figure 22: Principal coordinate analysis plot showing unweighted UniFrac phylogenetic distance (beta diversity) between samples and DNA extraction methods. Method 1 was the standard manufacturer's protocol, method 2 was as method one with bead beating at 6 ms-1 for 160 seconds, method 3 was as method 2 with the addition of a 5 minute incubation of the DNA binding matrix at 55 °c 85

Figure 23: Cluster analysis using weighted UniFrac distance between samples from Tudhoe Mill nitrifying activated sludge plant (ASP) treating domestic sewage (S1), Amble non-nitrifying sequencing batch reactor ASP treating domestic sewage (S2) and Aycliffe ASP treating heavy industrial and domestic sewage (S3). DNA from each sample was extracted according to the standard protocol (method 1, M1), the standard protocol except bead beating at 6 ms-1 for 160 s (method 2, M2) and as method 2 with an additional incubation at 55 °c for 5 minutes (method 3, M3). The length of each branch is proportional to the UniFrac distance between samples. The indicated p values represent the significance of the UniFrac distance determined by 100 Monte Carlo random sequence assignments. Data points S1.M3, S2.M3 and S3.M3 in the right hand box represent samples from each plant where DNA extracted using the same method and were included to demonstrate the distance between samples within extraction method. 86

Figure 24: Regression analysis of abundance of ammonia oxidising bacteria (AOB) measured by the Personal Genome Machine (PGM ALL where the values indicate relative abundance) and fluorescence in situ hybridization (FISH where values indicate the number of cells per millilitre) in 10 lab scale and 12 full scale nitrifying activated sludge reactors. Each data point represents the number of AOB cells per ml as measured by FISH and NGS. The number denoted by the letter S is the regression equation standard error. Regression equation: log10(PGM r.a. (all)) = - 4.241 + 0.2541 log10(FISH)..... 88

Figure 25: Regression analysis of abundance of ammonia oxidising bacteria (AOB) measured by quantitative polymerase chain reaction (qPCR where the values are number of cells / ml) targeting the amoA gene and Personal Genome Machine (PGM where the values are relative abundance of AOB) in 10 lab scale and 12 full scale nitrifying activated sludge reactors. Each data point represents the number of AOB cells per ml as measured by FISH and NGS. The letter S denotes the regression standard error. Regression equation: log10(PGM r.a. (all)) = - 4.189 + 0.2781 log10(amoA qPCR)..... 89

Figure 26: Regression analysis of the abundance of ammonia oxidising bacteria (AOB) measured by quantitative polymerase chain reaction (qPCR) targeting the 16S rRNA gene in 10 lab scale and 12 full scale nitrifying activated sludge reactors. Each data point represents the number of AOB cells per ml as measured by FISH and NGS. Regression equation: log10(PGM r.a. (all)) = - 4.859 + 0.4705 log10(16S qPCR)..... 89

Figure 27: Coefficient of variation of total ammonia oxidising bacteria relative abundance determined by next generation sequencing at multiple read depths where reads were selected by random rarefaction with replacement 10 times at each read depth..... 91

Figure 28: Enterococci and *E. coli* colony forming units (CFU) per 100 ml seawater sample from Saltburn bathing water where analysis was undertaken at the Environment Agency (EA). Samples were taken at the same time and place. n = 23 samples..... 93

Figure 29: Enterococci colony forming units (CFU) per 100 ml seawater sample from Saltburn bathing water where analysis was undertaken at the Environment Agency

(EA) or Northumbrian Water (NWL) labs. Samples (n=5) were taken at the same time and place..... 93

Figure 30: Enterococci colony forming units (CFU) per 100 ml seawater sample from Saltburn bathing water where analysis was undertaken at the Environment Agency (EA) or Newcastle University (NU) laboratories. EA measurements used standard culturing techniques to enumerate Enterococci. NU methods used 16S rRNA gene DNA sequencing by Ion Torrent Personal Genome Machine (PGM) producing relative abundance data. Samples were taken at the same time and place. n = 23 samples... 94

Figure 31: SourceTracker results showing the relative contribution of microbial communities from different sources (identified in the key which reads left to right) to the bathing water (sink) at Marsden Bay. STW = Sewage Treatment Works, Surface = Surface water drains. CSO = Combined Sewer Overflow. 100

Figure 32: Taxa summary of all activated sludge plant mixed liquor samples where each colour represents a summarised OTU grouping. Selected taxa are identified to illustrate the large abundance of bacteria whose function remains unknown e.g. *Hydrogenophaga* and the comparatively low abundance of vital bacteria e.g. Nitrosomonadaceae. All but one plant was sampled once – the time series plant data is indicated..... 102

Figure 33: Cell Specific Ammonia Oxidation Rate (CSAOR) for 25 activated sludge plants. A plant was considered stable if it had not breached its ammonia effluent consent in the previous 12 months. 103

Figure 34: Relative abundance of foaming bacteria in activated sludge plants with samples grouped according to whether they do not foam (NF, n=20), occasionally foam (OF, n=20) or are currently foaming (CF, n=1). Note that only one plant was sampled when currently foaming. 104

Figure 35: Boxplot showing the percent relative abundance of bulking bacteria from 32 activated sludge plant samples. Plants were grouped by operational status based on reports from plant operators. NB = Not bulking, OB = occasional bulking, CB = currently bulking..... 105

Figure 36: Percent relative abundance of bulking bacteria in samples from 32 activated sludge plants versus effluent suspended solids concentration in mg/L. 106

Figure 37: Linear regression analysis between number of Enterococci colony forming units (CFU) measured by the Environment Agency (EA) and total percentage of faecal source in sink determined by the SourceTracker tool. n = 23 samples 109

Figure 38: Linear regression analysis between number of *E. coli* colony forming units (CFU) measured by the Environment Agency (EA) and total percentage of faecal source in sink determined by the SourceTracker tool n = 23 samples 109

Figure 39: Partial least squares analysis between the number of Enterococci colony forming units measured by the Environment Agency (EA) and total percentage of faecal source in sink determined by the SourceTracker tool. n = 23 samples. 110

Figure 40: Partial least squares analysis between the number of E. coli colony forming units measured by the Environment Agency (EA) and total percentage of faecal source in sink determined by the SourceTracker tool. n = 23 samples. 110

Figure 41:SourceTracker results; the relative contribution of microbial communities from different sources (identified in the key which reads left to right) to the bathing water (sink) microbial community at Saltburn: Top, during 10th August 2014 storm before combined sewer overflow (CSO) operation; middle, during storm and CSO operation; Bottom, on 11th August, 6 hours after CSO cessation. STW = Sewage Treatment Works, Surface = Surface water drains 114

Figure 42: Abundance of ammonia oxidising bacteria measured by fluorescence in situ hybridization (FISH) and Personal Genome Machine (PGM) where only those taxa targeted by FISH probes Neu, NSO1225 and 6a192 were included in PGM data. S denotes the regression standard error. Regression equation: log10(PGM r.a. (FISH)) = -6.436 + 0.4098 log10(FISH)..... 164

Figure 43: AOB measured by FISH and PGM where PGM data is relative abundance for all AOB taxa detected multiplied by total cell count..... 165

Figure 44: AOB measured by FISH and PGM where PGM data is relative abundance for only those AOB taxa targeted by FISH probes NSO1225, Neu and 6a192 multiplied by total cell count..... 165

Figure 45: Abundance of ammonia oxidising bacteria (AOB) by quantitative polymerase chain reaction (qPCR) targeting the ammonia mono-oxygenase A gene (amoA) versus

the relative abundance of AOB detected by Personal Genome Machine (PGM) multiplied by total cell counts 166

Figure 46: Abundance of ammonia oxidising bacteria measured by quantitative polymerase chain reaction (qPCR) targeting the 16S gene versus AOB relative abundance measured by PGM multiplied by total cell count 166

Chapter 1. Introduction

The water industry relies on a number of biological treatment processes to provide potable water which is safe to drink and wastewater which is safe to discharge to the environment. The activated sludge wastewater treatment process is thought to be the most important biotechnological process in the world, treating over 330 trillion litres of sewage per year (Horan 2014; Drechsel et al. 2015). However the process is designed and operated empirically, with little understanding of the underlying microbiological processes. ASM1 (Activated Sludge Model) and ASM3 are based on Monod-type kinetics (Henze et al. 2000) and do not take account of the individual substrate affinities and metabolic rates present in each species nor the detrimental effects of over population by foaming and bulking bacteria. This leads to over engineered solutions which occasionally fail unexpectedly with no obvious cause. The process is energy intensive due to aeration of the mixed liquor, with one company spending approximately £5m per annum on aeration alone (Northumbrian Water 2011; Soares 2008). Given the significant expenditure on electricity and associated CO2 emissions due to aerating the mixed liquor and environmental impact of treatment failure, improved process control would be beneficial to the water industry, the environment and the consumer. A similar situation exists with bathing water quality as measured by the abundance of certain faecal indicator bacteria. In the past when wastewater was insufficiently treated it was clear that sewage was the primary source of faecal indicator bacteria in the bathing water. However as investment by the industry has led to widespread collection and secondary treatment of sewage, the source of these bacteria is no longer clear. Demand for cleaner bathing waters is high, leading to attempts to address the problem by hydrological modelling of sewerage catchments. This provides valuable insight in to the operation of a catchment under storm conditions, but does not directly investigate the flow of bacteria. Nevertheless the results of these modelling studies have been used to justify significant investment in sewerage infrastructure. Northumbrian Water built a storm water storage tank to reduce the number of spills from a combined sewer overflow at a cost of £6m with no clear understanding of the impact this action would have on bathing water quality.

These situations have arisen due to the lack of a cost effective, accurate method for the identification and enumeration of bacteria with sufficient throughput to provide

timely and operationally relevant data to the water industry. One method which can identify and enumerate bacteria is DNA Sanger sequencing of the 16S ribosomal RNA (rRNA) gene (Woese et al. 1975), which is considered the best evolutionary marker for distinguishing bacterial taxa. This technology identifies the bacteria present in a sample by comparing the sequence of the 16S rRNA gene against a database of 16S rRNA gene sequences from known bacteria (Woese et al. 1977). However this technique was costly and low throughput, and therefore not suitable for wide scale microbial population surveys. The advent of next-generation DNA sequencing changed this, with costs falling quicker than Moore's law (Glenn 2011). For the first time many tens of thousands of 16S rRNA genes can be sequenced quickly and cheaply, facilitating surveys on a scale that could bring about new insights in to the microbial functions so vital to the water industry

One next generation DNA sequencing platform is the Ion Torrent Personal Genome Machine (PGM, Life Technologies, California USA). The PGM is able to sequence up to 5m reads, comparable to identifying 5m bacteria per run. The process can be made more cost effective by multiplexing many samples per run, with each run taking only 3 hours. Therefore for the first time the challenges detailed above can be addressed in a cost effective and timely manner. However the protocols, data analysis procedures and concordance with existing regulatory techniques has yet to be established.

1.1 Aim and Objectives

Aim: To establish the suitability of DNA sequencing by PGM as a monitoring tool for the UK water industry for ASP monitoring and control and MST for faecal contamination source apportionment.

Objectives:

1. Develop laboratory and data analysis protocols that enable the identification and enumeration of relevant bacterial taxa in activated sludge and microbial source tracking
2. Collect data on the abundance of nitrifying, foaming and bulking bacteria from Northumbrian Water activated sludge plants along with effluent chemical parameters and qualitative assessment of plant failure risk

3. Collect data on the abundance and source apportionment of faecal indicator organisms in the Saltburn catchment.
4. Evaluate DNA sequencing by PGM in comparison with existing techniques as a cost effective tool for process monitoring and microbial source tracking.

Chapter 2. Molecular Techniques Overview

2.1 Isolation and enrichment of bacterial cells

Bacteria were discovered in 1676 by Antoni van Leeuwenhoek, however it was not until the late 19th century that bacteria were isolated and grown in pure culture (Madigan & Brock 2012). Robert Koch was the first to do so, culturing Anthrax bacilli and inoculating mice to demonstrate that the bacterium was the infectious agent (Madigan & Brock 2012). However this approach did not shed light on the metabolic requirements of bacteria. Further work by Martinus Beijerinck and Sergei Winogradsky developed the concept of the enrichment culture (Madigan & Brock 2012), in which an inoculum thought to contain the microbe of interest is enriched in conditions which most closely match its ecological niche. The microbe can then be isolated, usually by subculturing using growth media and conditions which select for the target microbe and against others. Growth conditions manipulated to achieve this aim include pH, organic substrates present and temperature amongst others. These physiological traits can be used in combination with observation of cell morphology to identify microbes.

However, it is thought that greater than 99% of bacterial and archaeal species are "unculturable" (Martins et al. 2004; Amann et al. 1995). This may be due to inability to provide a suitable growth medium or poor understanding of required environmental characteristics such as optimum temperature for growth (Amann et al. 1995). Therefore investigations based solely on culturing techniques will fail to characterise the importance of many taxa. Morphological techniques are unreliable as bacterial cells can be morphologically similar yet be different species; such as *Gordonia amarae* and *Rhodococcus* spp, or be morphologically dissimilar yet of the same species; for example the pleomorphic *Sphaerotilus natans*, whose morphology is known to vary according to environmental conditions (Martins et al. 2004). The morphotype *Nostocoida limicola* has recently been shown to describe a grouping of different taxa from four phyla (Nielsen, Kragelund, et al. 2009), illustrating the difficulty in attempting to identify cells using morphology alone.

Early classification of microbial species was based upon morphology and biochemical testing (Haeckel 1866). This included well known staining techniques such as the Gram stain (Gibbons & Murray 1978) or testing for the ability to metabolise particular

compounds such as acetate or lactate. However our burgeoning understanding of microbial genetics has revealed the limitations of these techniques (Woese 1987). Horizontal gene transfer enables bacteria to acquire genes from other distantly related taxa. These newly acquired genes may encode enzymes enabling the cell to metabolise compounds which previously it could not, altering the outcome of biochemical testing and thereby confounding the use of these techniques for identification.

Attempts to classify bacteria based on their genomic DNA sequence build on the concept that classification is based on shared evolutionary history. Therefore cells from the same species will share highly similar genomic DNA sequences. This can be tested using DNA hybridization studies where greater than 70% hybridization indicates cells are from the same species (Moore et al. 1987). However DNA hybridization tests are slow, low throughput and work only for pure cultures. The discovery of the highly conserved 16S ribosomal RNA gene and its utility in identifying bacteria and archaea has revolutionised microbial classification (Woese et al. 1975; Fox et al. 1977). By sequencing the 16S rRNA gene from a bacterial or archaeal cell and comparing this sequence against that of known taxa, one can deduce it's taxonomy. However 16S rRNA gene sequence similarity between taxa of the same species varies with no fixed thresholds defined (Janda & Abbott 2007). To circumvent this issue, the concept of the operational taxonomic unit (OTU) is used to describe a group of bacteria at a particular taxonomic level. The OTU is created based on 16S rRNA gene sequence similarity, where 97% similarity generally equates to 70% DNA-DNA hybridization, and therefore the sequences within that OTU are generally considered as belonging to a single species (Janda & Abbott 2007). This threshold must be used with caution, since some cells of the same species feature greater 16S rRNA similarity and some cells from different species feature greater than 99% similarity. See sections 2.5.2 and 2.5.3 for discussion on the mitigation of this issue.

2.2 Polymerase chain reaction

The polymerase chain reaction (PCR) was developed in 1983 by Kary Mullis (Bartlett & Stirling 2003). PCR utilises the natural mechanism of DNA replication to selectively amplify DNA. Selectivity is achieved by using oligonucleotides known as primers which are complementary to the DNA sequence at either end of the region of interest. These

primer pairs bind to the template DNA molecule when the double strand denatures and allow the polymerase to attach and begin replication. Therefore by designing the primer pairs to be complementary only to the area of interest the amplified DNA will be the target DNA fragment. A number of parameters can affect primer binding, of which annealing temperature is the most important. At a lower than optimum temperature the primer can anneal to non-target areas whilst at higher temperatures the primer can fail to anneal entirely. Therefore when using a new primer design an annealing temperature gradient is undertaken and the size of amplified fragments (amplicons) observed. The optimum temperature is the highest temperature at which the target amplicon size is produced with no non-target size amplicons observed.

The selection of primer pair dictates what can be concluded from the presence of amplicons at the end of the reaction. For example, where the primer pair is specific to a single gene the presence of amplicons indicates that bacteria containing the gene were in the sample and thus in the template DNA. This concept can be utilised to investigate the presence of bacteria with a particular metabolic capability, such as ammonia oxidation, where a single protein (and thus gene) is indicative of the metabolic pathway. For example, amplicons generated from a primer specific to the amoA gene would indicate the presence of ammonia oxidising bacteria in the sample (Rotthauwe et al. 1997). This can be used to establish the presence of a group of bacteria in a sample, but gives no indication of quantity.

2.2.1 Quantitative PCR

Quantitative PCR (qPCR) harnesses the exponential nature of PCR DNA amplification along with an optical reporter which provides a quantitative indication of DNA presence (Higuchi et al. 1993). These optical reporters can either be generic fluorescent DNA stains or labelled oligonucleotides which specifically anneal to the amplified target gene (Livak et al. 1995). In either case the intensity of fluorescence is proportional to the concentration of DNA present. The increase in fluorescence intensity is compared against the increase in intensity produced by a range of concentration standards. These standards contain a known starting concentration of DNA template such as 10^7, 10^8 and 10^9 copies. Therefore by comparing the point at which the sample fluorescence intensity enters the exponential phase and the point at which the standards enter the exponential phase the concentration of the sample DNA

can be determined. qPCR uses primers for the same purposes as PCR i.e. to selectively target genes of interest. Therefore qPCR products indicate both presence and abundance of the target gene, and as such can be used to establish the number of cells per ml of sample.

2.3 Fluorescence In Situ Hybridization

Fluorescence In Situ Hybridization (FISH) is currently considered the best tool for activated sludge microbial ecology studies (Nielsen, Daims, et al. 2009; Bellucci & Curtis 2011). This technique enables identification of taxonomic groups at all levels from kingdom to species. The 16S subunit of the bacterial ribosome is targeted using a probe with a fluorescent moiety attached. A number of moieties are available such that up to 4 probes can be used simultaneously to distinguish several OTUs. The probe is a small piece of DNA which binds specifically to the 16S rRNA sequence under investigation. This probe can be designed to target specific taxa, which can be achieved due to the varying evolutionary conservation of 16S rRNA sequence. Due to the vital nature of ribosomal function, the overall sequence of the 16S rRNA is highly conserved. However there are areas not involved in 16S rRNA folding (and thus function) which are significantly less conserved, and thus more variable than others (Bellucci & Curtis 2011). Consequently, certain sections of the sequence are identical amongst all taxa within a group whilst other sections of the sequence are specific to subgroups such as species. It is this knowledge that enables the design of probes specific to the taxonomic level desired, for example facilitating the identification of all AOB or just *Nitrosomonas eutropha*.

Once the probes have been applied, the sample is then visualised using a fluorescence microscope. Abundances can be estimated by viewing a number of focal planes and counting cells. Cell clustering and flocs can be accounted for using equations specific to cluster type e.g. the “footballs” formed by AOB (Coskuner et al. 2005).

2.4 Sequencing

2.4.1 The 16S ribosomal RNA gene

The 16S rRNA gene is a highly conserved gene encoding the 16S ribosomal subunit (Woese et al. 1975). The approximately 1500 base pair gene contains regions which are hypervariable and thus feature greater evolutionary divergence than the other 16S rRNA regions (Woese et al. 1975) as described in section 2.1. This phenomenon can be

utilised for the identification of bacterial species since the evolutionary distance and thus sequence divergence between two cells of the same species is less than the sequence divergence between two cells of different species (Fox et al. 1977). Therefore one can identify taxa by comparing its 16S rRNA gene sequence against 16S rRNA sequences in a database of known taxa. For further discussion of 16S rRNA primer sequence selection see section 5.2.

2.4.2 Sequencing platforms

Culturing and morphological observation are laborious, low throughput techniques which may give false negative results due to the fastidious nature of taxa present (Mancini et al. 2010). Culture techniques may also produce incorrect or no identification of taxa with atypical or uncharacterised phenotypes (Drancourt et al. 2000). However culturing remains a key tool in understanding the morphology and physiology of bacteria which is important in identifying their ecological role and industrial relevance. For microbial community surveys one might wonder why 16S rRNA gene sequencing has not swiftly replaced these traditional approaches. In part this is down to the capital cost of DNA sequencers and the reagent cost per sample. Until recently the dominant platform was the 454 platform, which cost $530,000 to purchase excluding ancillary lab equipment and $6200 per run (Glenn 2011, Table 1). In combination these issues meant DNA sequencing was not feasible for routine operation. However recent advances in technology have significantly reduced the capital and reagent costs, meaning investigations can now begin in to how DNA sequencing can inform the operation of wastewater assets.

The Ion Torrent Personal Genome Machine (PGM) is a unique new DNA sequencing platform, costing only £50,000 and with reagent costs as low as £2 per sample (Glenn 2011, Table 1)). The cost reduction is primarily achieved by avoiding the use of complex optical components in the detection of nucleotide incorporation. Instead the PGM detects the release of a hydrogen ion from the 3’ hydroxyl group as a pH change when a nucleotide hybridizes to the growing polymer. The reactions take place within microwells on a complementary metal–oxide–semiconductor chip which is electrically connected to the PGM (Rothberg et al. 2011). In the main, it is the number of wells that can be fabricated on one chip which determines the throughput of the machine.

Therefore as semiconductor technology improves, it is expected that throughput will increase and costs consequently decrease.

Table 1: Comparison of next generation sequencing platforms. Mbp = mega base pairs.

Platform	Capital cost	Cost / run	Cost / Mbp	Reads/ run (m)	Run time	Error rate
PGM (318)	$68k	$925	~$0.93	8	2h	~1%
454 FLX Titanium	$530k	$6200	$12.4	1	10h	~1%
MiSeq	$125k	$750	$0.74	3.4	26h	~0.1%

2.5 Bioinformatics

The DNA sequencing platform produces "raw" data in the form of signal amplitudes representing pH change. This signal data is then converted to DNA sequence data known as "reads" via the proprietary software package "Torrent Suite" (Life Technologies 2014). There are no other options available for this step. DNA sequence data must be processed in order to produce bacterial identification and abundance data. A number of separate tools have been developed to address different steps in this process, such as BLAST for taxonomy assignment (Altschul et al. 1990) or RaXML for building phylogenetic trees (Stamatakis 2006). In addition there are multiple tools for achieving the same outcome, each with their own claimed benefits. For example, sequence alignment can be achieved using PyNAST (J. Gregory Caporaso et al. 2010), Infernal (Nawrocki et al. 2009), ClustalW (Larkin et al. 2007), MUSCLE (Edgar 2004) and MAFFT (Katoh & Standley 2013) to name but a few (Figure 1).

2.5.1 Pipeline choice

The large number of software packages required to achieve common steps in 16S rRNA data processing combined with their differing parameter formats has led to the development of data processing pipelines. These pipelines combine a wide range of software packages in to a small number of commands. However just as there are many bioinformatics tools available for the same task, there are multiple 16S rRNA sequence data analysis pipelines. The choice between these pipelines is largely based on personal preference as there is often overlap in capabilities. The Quantitative

Insights Into Microbial Ecology package (QIIME) (J Gregory Caporaso et al. 2010) is widely used, regularly updated and offers a number of software packages for each analysis step. There is also a streamlined installation process to enable easy local installation, as well as virtual machine images and Amazon Web Services images. mothur (Schloss et al. 2009) is another widely used software package which is packaged as a single binary. The mothur author claims that mothur is superior to QIIME in speed and uses only open source software. At the time of writing, QIIME was cited 2685 times whilst mothur was cited 3676 times giving an indication of popularity. A number of other less popular pipelines are available, including LotuS (Hildebrand et al. 2014) and Clovr (Angiuoli et al. 2011).

2.5.2 OTU picking

The 16S rRNA gene (hereby referred to simple as 16S) sequence varies between and within species and therefore the reads in a DNA sequencing run must be clustered such that reads from the same species are grouped together (Bond et al. 1995). The 16S sequence variance within a single species can be from 90% to 100% similarity (Janda & Abbott 2007). In order to cluster 16S reads from the same species, a percentage similarity must be selected which represents the best balance between including 16S sequences from the same species and excluding 16S sequences from different species. Generally 16S reads are grouped (picked) at 97% similarity (Janda & Abbott 2007) in to operational taxonomic units (OTUs). We refer to OTUs as opposed to species since a group of sequences greater than 97% similar may or may not represent a single species due to the aforementioned variance in multiple 16S genes in a single species. One must remember OTU picking may cause over inflation of the observed number of OTUs, and failure to resolve OTU taxonomy to species level. For example, two different species may have 99% similar 16S genes (Fox et al. 1992) and conversely some species may feature multiple 16S sequences which are only 95% similar (Eren et al. 2013). It is possible to pick OTUs at 99% similarity; however this may result in over inflation of the number of OTUs present in the sample since a greater proportion of the aforementioned highly divergent taxa will form artificially distinct OTUs. Furthermore in a 450bp read only 5 bases need differ to create 1% difference which may lead to false OTU identification due to sequencing or PCR errors. Therefore the choice of OTU similarity cut off must be selected with reference to the taxa under investigation.

2.5.3 Taxonomy assignment

Once the OTUs have been picked, a representative set must be created since in any one OTU cluster the sequences within could be up to 3% dissimilar. A 16S sequence which best represents the original 16S reads clustered in to the OTU is selected for each OTU. Common approaches for selection of this "best representative" include the most abundant read in the cluster, the longest read, a random read or the first read (J Gregory Caporaso et al. 2010). The first read is used to avoid further computational complexity if the reads were sorted by abundance before clustering, since the first read will therefore be the most abundant. Once chosen, each read in the representative set is then assigned to the highest taxonomic group possible with reference to a taxonomy database (see section 2.5.4). For example where the sequence represents a well-studied taxon it is likely that the taxon and thereby its 16S sequence has been characterised to species level and named accordingly. Thus the query sequence will be assigned to species level. In other cases the query sequence may represent a novel taxon which has yet to be fully described. In this case the read will be assigned to species level but the name of the species will not appear in the data set. Lastly, where the sequence cannot be reliably classified to species level it is classified to genus level and the label "other" added for species. This process applies such that the highest taxonomic classification possible is achieved, so that where the data only supports classification to Phylum level the taxon would be classified as for example *Chloroflexi* : Other.

There are many methods for taxonomy assignment, including BLAST (Altschul et al. 1990) and RTAX (Soergel et al. 2012). The default taxonomy assignment tool in QIIME 1.7 is RDP (Wang et al. 2007). This tool uses naive Bayesian classification to assign taxonomy to a read and repeats this classification 100 times for each read. The user can then set the minimum confidence level (number of times the same taxon is selected) with a default setting of 0.8. This would mean that 80 times out of 100 the same taxon would be selected. However Claesson et al (2009) found that a bootstrap cutoff of 0.5 produced correct classifications at genus level for 95% of reads.

2.5.4 Databases

The taxonomic database used when assigning taxonomy is vital, since if the taxa of interest do not appear in the database it is impossible for them to appear in the

results. Greengenes is a highly curated database containing mostly full length 16S rRNA gene sequences (DeSantis et al. 2006), is checked for chimeras (see 2.5.5) and is relatively small and thus can be loaded in to 4GB RAM. However Greengenes is not regularly updated; the most recent release was May 2013. The number of OTUs in Greengenes is also relatively limited at 99322 (97% cutoff). SILVA is another taxonomic database in widespread use which is highly curated (Pruesse et al. 2007). SILVA includes rRNA genes predicted from genomic sequence data as well as those annotated as rRNA. All sequences are checked for acceptable alignment scores, thereby removing incorrectly annotated or predicted genes. The SILVA taxonomy is fully manually curated, and therefore can contain newly discovered, functionally important taxa before other databases. An example is *Nitrotoga*, recently confirmed as an important nitrite oxidising genus (Lücker et al. 2014) which appears in the SILVA taxonomy and is absent in Greengenes. This reaffirms the importance of selecting the best database for each study.

2.5.5 Chimeras

Chimeras occur when a partially extended PCR fragment binds to another fragment with similar sequence but which originates from a different taxon (Kopczynski et al. 1994). The second fragment acts as a primer, leading to the creation of a single fragment which contains 16S rRNA gene sequences from two organisms (Smyth et al. 2010). It is important to remove these sequences as they can over inflate diversity by appearing to be novel taxa (Wang & Wang 1997). There are many bioinformatic approaches to achieve this either by reference to a “gold” standard database or by de novo detection. One such method is ChimeraSlayer (Haas et al. 2011) which effectively splits the query reads in half and aligns each half to the “gold standard” i.e. the best available database. If the two halves align to the same gold standard database sequence, the query sequence is deemed non chimeric. If the two halves of the query sequence align to two different gold standard sequences, the query sequence is possibly chimeric.

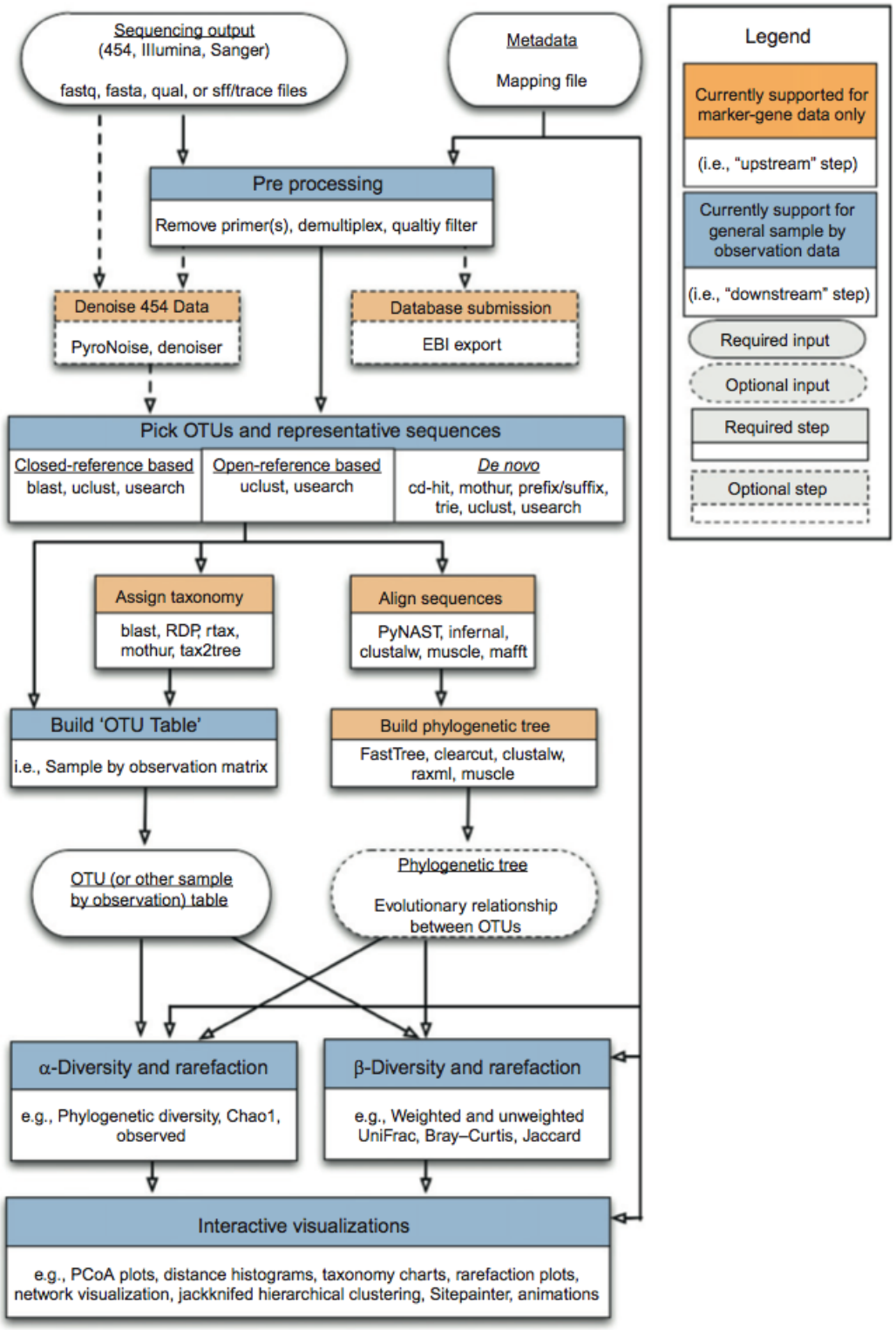


Figure 1: Flow diagram showing bioinformatics steps when processing 16S rRNA data. From https://sites.google.com/site/knightslabwiki/qiime-workflow

Chapter 3. Activated Sludge Process Monitoring Literature Review

3.1 Introduction

Activated sludge was first conceived by Ardern and Lockett in the early 20th century (Ardern & Lockett 1914). It is a sewage treatment process which is designed to retain settled biomass whilst discharging the supernatant as treated sewage (see Figure 2). In this way the retention time of the bacterial community can be greater than the hydraulic retention time of the reactor. Therefore slow growing bacteria can be retained and utilised in the oxidation of pollutants such as complex organic matter and ammonia. Activated sludge is now the most important biotechnological process in the world, treating over 330 trillion litres of sewage per year (Horan 2014; Drechsel et al. 2015), yet despite our reliance on this process the underlying biology remains poorly understood. Models such as ASM3 (International Water Association 2000) are able to inform plant design but are based on simple Monod kinetics (Monod 1949) and therefore cannot fully encapsulate the complex activated sludge plant (ASP) community dynamics. This leads to ASP designs featuring large margins of error and inefficient operation. Issues such as foaming and bulking as well as over aeration and unpredictable nitrification failure continue to impact on effluent quality and operational costs (Rossetti et al. 2005; Davenport et al. 2008). It is expected that effluent quality requirements will become more stringent in the near future both in terms of existing regulated parameters and new regulatory limits on micropollutants. At the same time wastewater companies are seeking to reduce their electricity consumption to minimise costs and carbon footprints. In order to meet these challenges using existing infrastructure we must gain a greater understanding of microbial community dynamics in ASPs and how they relate to effluent quality.

The UK water industry currently accounts for 5 million tonnes of CO_2 emissions per annum, of which 1.7 million tonnes can be attributed to sewage treatment (Environment Agency 2009). A significant proportion is due to electricity consumption by pumps, blowers and other plant machinery. Northumbrian water treated 184,690 megalitres of sewage in the regulatory period 2010-11 with energy costs of £7.9 million (Northumbrian Water 2011). Approximately 55% of energy use is due to aeration (Soares 2008) equating to £4.35m per annum. Carbon dioxide emissions due to aeration per megalitre treated are estimated at 88kg (Georges et al. 2009), thus

Northumbrian Water (NWL) activated sludge aeration resulted in the emission of 16,253 tonnes of CO2 in the same period. It is postulated that there are significant efficiency improvements to be made in ASP process control. This could be achieved by adjusting aeration such that dissolved oxygen (DO) levels are maintained at the minimum required for sufficient COD reduction or ammonia oxidation (Coskuner et al. 2005). Recent work by Severn Trent (2008) to adjust aeration levels according to ammonia concentration in the influent reduced aeration energy use by 10% in the winter and 20% in the summer. If similar savings were achieved using NGS as a monitoring tool, NWL could save £0.65 million and reduce CO2 emissions by 2,438 tonnes per annum along with the associated PR benefits. Given that 10 billion litres of sewage are produced per day in England and Wales (Georges et al. 2009) the potential for reduction in energy usage across the UK is substantial. In addition, a key driver in the water and wastewater industry is public perception (Waites 2012). Thus the PR benefits of reducing energy consumption and preventing pollution events are primary motivators for industry sponsorship of this research.

Many studies have attempted to understand the ecology of activated sludge over the last seventy years (Allen 1944; Dias & Bhat 1964; Balakrishnan & Eckenfelder 1969; Prakasam & Dondero 1970). Until recently studies of ASP ecology relied upon culturing techniques and morphological approaches to identify bacteria (see Chapter 2), limiting the ability of such studies to investigate the bacteria responsible for important process functions such as nitrification. However the recent and significant reduction in cost of DNA sequencing along with increase in throughput (see 2.4) has enabled the use of next generation sequencing (NGS) in understanding the microbial community dynamics of activated sludge plants.

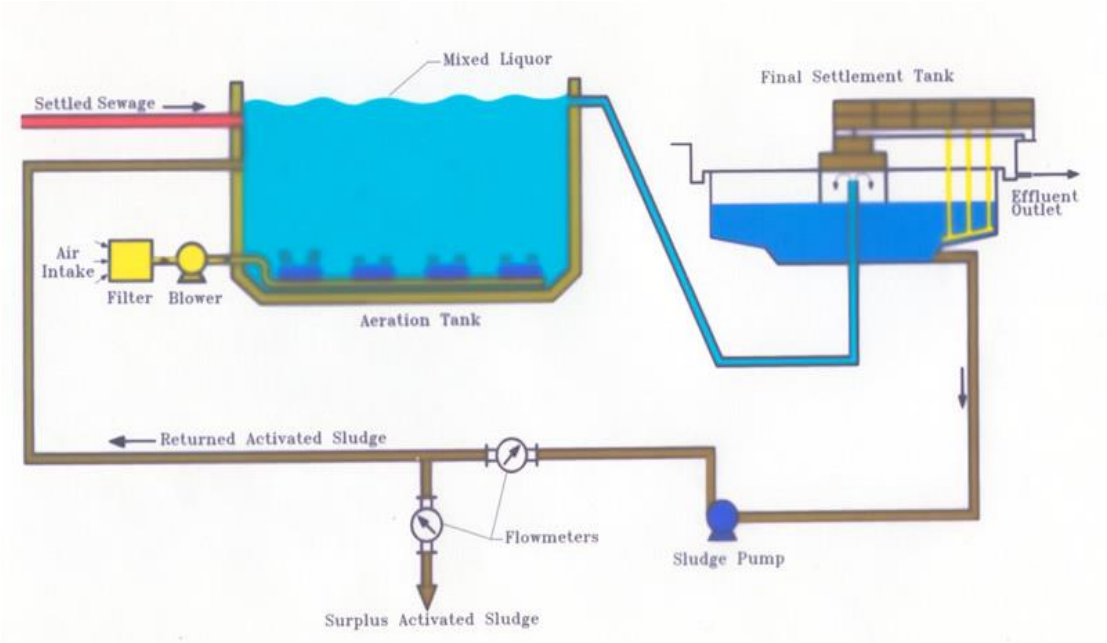


Figure 2: Schematic of the activated sludge secondary sewage treatment process. Raw sewage is first settled in a primary clarifier to remove solids before being pumped in to the aeration tank. The activated sludge microbial community then oxidises complex organic matter and in some cases ammonia. The suspended sludge and treated sewage is then pumped to the final settlement tank where the activated sludge settles out and the supernatant is discharged as final effluent. The majority of the settled activated sludge is pumped back to the aeration tank, with the remainder "wasted" out of the system. The wastage rate influences the mean sludge age and is a key control parameter. Image from http://www.staffs.ac.uk/schools/sciences/consultancy/dladmin/zCIWEMWWT/Activity5/Images/bigasp.jpg

3.2 Carbonaceous activated sludge plants

Sewage contains a large concentration of readily biodegradable organic matter such as carbohydrates, lipids and proteins. Aerobic chemoorganoheterotrophs are the dominant consumers of organic matter in oxic environments such as rivers, consuming oxygen as they metabolise organic matter. This gives rise to biological oxygen demand (BOD), which correlates with the amount of organic matter available for consumption. If sewage with a high BOD is discharged in to a river the resultant increase in oxygen consumption by the aerobic chemoorganoheterotrophs may totally deplete dissolved oxygen in the water, leading to asphyxiation of fish and invertebrates living in the receiving water. Therefore it is important that the BOD of the final effluent is sufficiently low to prevent deleterious effects on the ecological status of the receiving water body. A carbonaceous activated sludge plant (ASP) is designed and operated to

achieve this aim, generally with a sludge age of around 3 days (Metcalf & Eddy. et al. 2003). The microbes which oxidise complex organic matter are quick growing and due to the wide phylogenetic diversity of the taxa involved the process is generally very stable requiring little monitoring and intervention.

3.3 Nitrifying activated sludge plants

Nitrogen can be inert or a significant pollutant depending on the form in which it is present. Around 60% of nitrogen in sewage is in the form of ammonia with the remaining 40% as organic nitrogen (Rittmann et al. 1999). Heterotrophic bacteria quickly convert organic nitrogen to ammonia, and it is ammonia which is the most environmentally damaging form of nitrogen as it is toxic to aquatic life at low concentrations. Therefore the EU Urban Wastewater Treatment Directive (EEC Council 1991) was enacted to protect the environment from excessive nitrogenous inputs. The directive stipulates a total nitrogen limit of 15 mg/L for sewage from 10,000 to 100,000 population equivalent and 10 mg/L for greater than 100,000. These limits are further reduced where the effluent is discharged in to "sensitive areas" on a case by case basis. Specific ammonia limits set by the Environment Agency can be as low as 2 mg/L and given that typical domestic sewage contains 25 – 45 mg/L of ammonia (Metcalf & Eddy. et al. 2003), sewage must be treated to achieve the required effluent quality. In these cases an ASP can be designed and operated in a way that encourages the growth of nitrifying bacteria. This requires a sludge age of greater than 5 days (Metcalf & Eddy. et al. 2003) as the nitrifying bacteria grow more slowly than the heterotrophic bacteria. The dominant nitrifying bacteria in ASPs can be split in to two groups – those which carry out the first stage of nitrification known as ammonia oxidising bacteria (Equation 1), and those which carry out the second stage of nitrification known as nitrite oxidising bacteria (Equation 2). Whilst there are other microbes such as ammonia oxidising archaea and anaerobic ammonium oxidation (annamox) bacteria, these have not been identified in significant numbers in ASPs (Limpiyakorn et al. 2011; Wells et al. 2009).

$$Ammonia\ (NH_3) \xrightarrow{Ammonia\ oxidising\ bacteria} Nitrite\ (NO_2^{-})$$

Equation 1: First step in nitrification

$$Nitrite\ (NO_2^-) \xrightarrow{Nitrite\ oxidising\ bacteria} Nitrate\ (NO_3^-)$$

Equation 2: Second step in nitrification

Until recently it was thought that a strict phylogenetic delineation existed between those aerobic bacteria which could oxidise ammonia and those which could oxidise nitrite. However recent studies have found a number of taxa which can oxidise both ammonia and nitrite (Pinto et al. 2016; van Kessel et al. 2015). Furthermore it has recently been established that bacteria outside of the traditional AOB clade *Nitrosomonadales* (phylum Proteobacteria) can oxidise ammonia, with some taxa being significantly phylogenetically distant (*Rhodococcus* from phylum Actinobacteria) (LaPara et al. 2015; Fitzgerald et al. 2015). These developments bring in to question the idea that AOB or NOB are phylogenetically distinct groups. However as the focus of this thesis is on AOB and NOB in the context of wastewater treatment it is reasonable to maintain this delineation for the purposes of discussion given that the primary concern is the abundance and function of bacteria found in ASPs which are phylogenetically distinct.

3.3.1 Ammonia oxidising bacteria

Ammonia oxidising bacteria (AOB) are chemolithoautotrophic oxidisers of ammonia to nitrite, the abundance of which is directly related to the nitrifying capacity of the plant (Coskuner et al. 2005). Ammonia oxidation is a two step process (see Equation 3 and Equation 4). In combination these equations dictate the maximum energy release per ammonium molecule oxidised and thereby the maximum mass of ammonia which can be treated by a given abundance of AOB. Ammonia oxidation is thought to be the limiting step in nitrification (Kowalchuk & Stephen 2001), though Graham et al (2007) propose that the abundance and stability of the NOB community directly impacts the abundance and stability of the AOB community. In any case, understanding the dynamics of this group is an important component in monitoring plant performance since without AOB nitrification cannot occur.

$$NH_3 + O_2 + 2H^+ + 2e^- \rightarrow NH_2OH + H_2O$$

Equation 3: First step in ammonia oxidation. Ammonia is oxidised to hydroxylamine

$$NH_2OH + H_20 \rightarrow NO_2{}^- + 5H^+ + 4e^-$$

Equation 4: Second step in ammonia oxidation. Hydroxylamine is oxidised to nitrite

The AOB found in activated sludge consist of two groups (Woese et al. 1984). The Nitrosomonads group are Betaproteobacteria and contain the genera *Nitrosomonas*, *Nitrosospira*, *Nitrosolobus* and *Nitrosovibrio*. The second group contains only the genus *Nitrosococcus*. The phylogenetic coherence of the AOB in ASPs has made the group relatively easy to study; for example probes for Fluorescence In-Situ Hybridization (FISH) studies can be designed to capture all ASP AOB (Coskuner et al. 2005). As a result, the operational factors which most strongly affect the AOB have been well studied (Coskuner & Jassim 2008; Amand & Carlsson 2012; Kowalchuk & Stephen 2001; Juretschko et al. 1998; Harms et al. 2003). Quantitative polymerase chain reaction (qPCR) probes have been developed to target and enumerate the AOB (Baptista et al. 2014; Bellucci & Curtis 2011). However until these techniques can be automated or require hands on time and skill equivalent to current ammonia testing, they remain too costly and slow to inform the operational control of plants.

The AOB found in ASPs are limited in diversity with only two genera thought to be responsible for the majority of ammonia oxidation (Seviour et al. 2010). Of these, *Nitrosomonas* was thought to be dominant (Wagner & Loy 2002; Prosser & Prosser I. 1989). However dominant taxa would appear to depend on the nature of the influent, with Juretschko et al (1998) noting dominance of *Nitrosococcus mobilis* in an industrial ASP treating influent with a high salt content. A number of studies have noted dominance of *Nitrosospira* (Siripong & Rittmann 2007; Hiorns et al. 1995; Yu & Zhang 2012; Coskuner & Curtis 2002) despite assertions elsewhere that *Nitrosospira* is only rarely observed in ASPs (Seviour et al. 2010; Koops et al. 2006; Wagner & Loy 2002). It seems logical given the diverse nature of influent composition and abiotic factors that the nitrifying community structure will vary globally.

AOB cells tend to form “football” shaped clusters that are quite distinctive. The number of cells within the cluster can be estimated using the methods of Coskuner et al (2005) as follows. A calibration curve correlating micro-colony volume with cell count is determined empirically. The mean volume and number of micro-colonies is

then observed enabling calculation of AOB cells per ml of mixed liquor. AOB abundance can then be used to calculate the cell specific ammonia oxidation rate (see Equation 5). The cell specific ammonia oxidation rate has been shown to correlate with ASP nitrification stability (Pickering 2008), defined as no nitrification failures in the 12 months prior to sampling (see Figure 3). AOB abundance monitoring could therefore be utilised in improving AS plant control if a monitoring technology can be developed which is cost effective and rapid. This could enable proactive management of insufficient AOB abundance by adjustment of plant operation control such as dissolved oxygen and sludge age, or potentially bioaugmentation using sludge from stable plants. The technology could also be used to provide early warning of nitrification failure. Plant operators could then initiate alternative emergency ammonia removal methods such as struvite precipitation whilst the AOB recover.

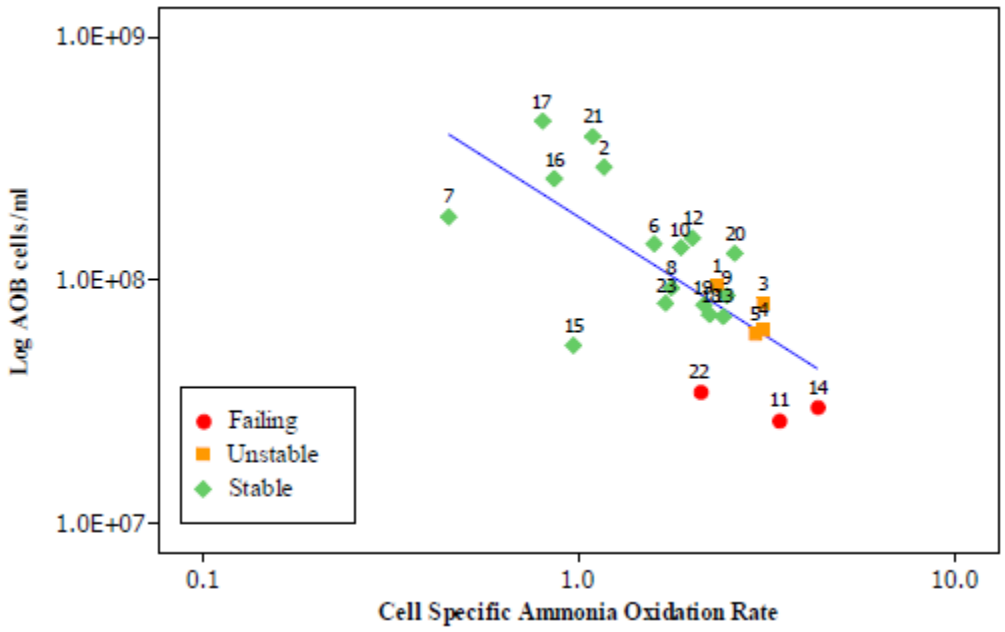


Figure 3: Samples from several nitrifying activated sludge plants where ammonia oxidising bacteria have been quantified using fluorescence in-situ hybridization. Plant operators classified the plants as stable (no failures in 12 months), unstable (at least one failure in 12 months) or failing (consistently failing) where failure is final effluent ammonia concentration exceeding consent. From Pickering et al (2008).

3.3.2 Nitrite oxidising bacteria

Nitrite oxidising bacteria undertake the second step in nitrification – oxidising nitrite to nitrate (see Equation 2). This group is generally not considered to be rate limiting, though as previously discussed NOB abundance may impact AOB abundance (Graham

et al. 2007). In addition high concentrations of nitrite are inhibitory to AOB growth. Thus whilst important, NOB abundance is of less significance to AS plant control than AOB abundance.

NOB are a more phylogenetically diverse group than the AOB, however the number of species found in AS plants remains small and thus have been well studied (Lücker et al. 2014; Juretschko et al. 1998; Siripong & Rittmann 2007; Harms et al. 2003; Coskuner & Curtis 2002; Saunders et al. 2015; Eyice et al. 2007; Mobarry et al. 1996; Daims, Nielsen, et al. 2001). Early culture based studies identified *Nitrobacter* as the primary genus in AS NOB. However molecular techniques did not corroborate this observation and it is now understood that *Nitrobacter* spp. are not the dominant NOB (Wagner et al. 1996). In fact *Nitrospira spp* are dominant though resistant to culturing, with only one species – *N. defluvii* – being cultured so far (Seviour et al. 2010; Juretschko et al. 1998). Kinetics studies have demonstrated that *Nitrobacter* is able to grow substantially quicker than *Nitrospira* in excess nitrite, however when nitrite concentration is low *Nitrospira* is able to outcompete *Nitrobacter* (Boon & Laudelout 1962; Blackburne et al. 2007), explaining *Nitrospira* dominance in AS plants due to the rate limiting AOB. It has been noted that several *Nitrospira* species can occupy different positions within the floc and thus different ecological niches, since the concentration of nitrite varies according to proximity to the AOB (Maixner et al. 2006; Gieseke et al. 2003). It can be postulated that this ability to support NOB diversity enables more stable plant operation, particularly in those which receive occasional ammonia spikes in the influent.

3.4 Foaming

Foaming refers to the presence of persistent biological foam in AS plants (Soddell & Seviour 1990). This may limit oxygen transfer by reducing effective tank surface area and increases suspended solids in the effluent and hence the receiving water course (Soddell & Seviour 1990). Operating costs may be affected due to greater aeration required (Seviour et al. 2010), whilst pollution incidents due to solids consent breaches may result in fines (Parliament 2010).

In order for foaming to occur, the following must be present: gas bubbles, hydrophobic particles, surfactants and solids (Petrovski et al. 2011). The gas bubbles are assumed to be from aeration, whilst the solids are assumed to be those present in the primary

settled influent. However the source of the surfactants and hydrophobic particles is uncertain (Petrovski et al. 2011; Blackall et al. 1991). For example, mycolic acids are hydrophobic compounds found in the cell walls and extracellular polymeric substances created by the Mycolata, a group universally present in AS plants yet only sporadically causing foaming (de los Reyes 2010). Similarly, *Actinomycetes* produce surfactants but it is not known whether this is a significant source within activated sludge (Soddell & Seviour 1990; Blackall et al. 1991). Critical micellar concentration (CMC) is the minimum concentration of surfactant required to form micelles which aggregate to form foams. Mycolic acid CMC can be as low as 7.4×10^{-5} mol/L (Lee et al. 2005) compared with a CMC of 1.3×10^{-3} for the commonly used detergent linear alkylbenzene sulfonate (Chauhan & Sharma 2014). It has been suggested that total surfactant concentration in the influent may be sufficient for the formation of foam (Narayanan et al. 2010). However McAvoy et al (1998) observed surfactant concentrations of up to 5.6×10^{-6}, 7.9×10^{-6} and 1.7×10^{-5} for alkyl ethoxylate sulfonates, alcohol ethoxylates and linear alkylbenzene sulphonates respectively in wastewater influent, lower than their CMC. Therefore the primary source of surfactants with respect to foam causation remains unclear.

Davenport et al (2008) identified an abundance threshold for mycolata cells below which foaming does not occur. It is hypothesised that the hydrophobic mycolata cell walls contribute to stabilisation of foams. When investigating this concept it is important to recognise that viability of cells appears to be irrelevant (Petrovski et al. 2011). Therefore the cells must be physically removed from the system as opposed to simply killing them. However as filamentous bacteria are involved in the degradation in a number of micropollutants complete removal of foaming bacteria may not be desirable (see section 3.6). The dominant species in foaming plants vary according to geography (Naidoo et al. 2011), potentially due to variance in average temperature (Guo & Zhang 2012). There remains some doubt regarding the full range of species implicated in foaming as it is difficult to establish whether species enriched within foams are causative or coincidental (de los Reyes 2010). Investigating only the enriched species discounts the impact of bacteria producing surfactants in the mixed liquor as discussed above. Additionally, in light of the doubt cast upon morphological identification systems, some of the assertions relating to species physiology such as

preferred electron acceptors cannot be relied upon (Martins et al. 2004). Therefore until this knowledge is verified using molecular methods, it must be used with caution.

Microthrix parvicella is regularly encountered in foaming events throughout the world (Soddell & Seviour 1990; Mielczarek et al. 2012; Nielsen, Kragelund, et al. 2009; Guo & Zhang 2012; Erhart et al. 1997) and along with the Mycolata are the most significant group in terms of foaming (Nielsen, Kragelund, et al. 2009). In particular *Gordonia spp.* and *Skermania piniformis* (previously *Nocardia pinensis*) (Seviour et al. 2008; Kragelund et al. 2007; de los Reyes et al. 1997) are regularly associated with severe foaming, with *Gordonia spp*. occurrence potentially being underestimated due to incorrect morphological identification using light microscopy (Nielsen, Kragelund, et al. 2009; Marrengane et al. 2011). It is recommended that molecular techniques are used as filament morphology is an unreliable tool for species identification (Soddell & Seviour 1990; Mielczarek et al. 2012; Nielsen, Kragelund, et al. 2009). Foaming bacteria ecology appears to be very complex, with conflicting evidence from culture and in situ studies. For example *Microthrix parvicella* is thought to exclusively feed on long chain fatty acids with surface lipase activity being detected in-situ. However this is not always detected in pure culture (Rossetti et al. 2005). Similarly discrepancies between in situ and pure culture observations have been described for oleic acid uptake in *Gordonia amarae* (Kragelund et al. 2007). It appears that *G. amarae* and *S. piniformis* can oxidise a wide range of substrates such as palmitic acid, ethanol and acetate (Seviour et al. 2008) under both aerobic and anoxic conditions (Kragelund et al. 2007). This metabolic flexibility may explain their global prevalence and highlights their utility in degrading complex organic compounds.

Eikelboom Type 0803 has recently been assigned to the *Chloroflexi* and is known to be physiologically similar to *M. parvicella* (Kragelund et al. 2011). It has been implicated in severe foaming events yet akin to other foamers is important in floc formation and degrades complex compounds (Kragelund et al. 2011). *Nocardia farcinica* has recently been isolated in South African AS plants and identified as a member of the *Nocardiaceae* which is dominant in foaming incidents in the region (Naidoo et al. 2011). As with other foaming filaments, it is able to metabolise complex and poorly biodegradable compounds, further confirming the ecological niche which filaments occupy (Naidoo et al. 2011). These new discoveries of dominant foaming bacteria

illustrate our emerging understanding of the phenomenon and highlight the difficulty of defining the foaming bacteria group.

3.5 Bulking

Bulking refers to the proliferation of filamentous bacteria (filaments) within AS flocs causing sludge to settle too slowly for efficient plant operation (Sezgin et al. 1978). This is caused by growth of filaments such that they extend outside of the floc, leading to significant reduction in gravitational settling rate (Guo & Zhang 2012). Bulking is problematic as suspended solids in the effluent may reach unacceptable levels and control of sludge age is impeded. However it is important to note that filaments are vital to stable floc formation, providing the backbone around which extracellular polymers attach (Guo & Zhang 2012).

Some studies have shown that substrate uptake rate for filaments studied is up to 80 times lower than floc formers, potentially explaining low filament abundance during normal plant operation (Martins et al. 2004). It would appear then that only during some perturbation of the plant are the filaments able to proliferate and become problematic. Several attempts have been made to quantify the parameters and threshold values which lead to such events, with mixed and often contradictory outcomes (Soddell & Seviour 1990; Petrovski et al. 2011; de los Reyes 2010; Davenport et al. 2000). More detailed investigation of the bulking filament phylogeny, in particular correctly identifying the dominant species would bring about greater understanding of the causative agents in bulking. However efforts to do so have been hampered by filament phylogenetic diversity and subsequent difficulty in establishing suitable molecular probes (Zhang et al. 2012), as well as indistinct morphology leading to misidentification of species (Tolvanen & Karp 2011). Isolating species has been challenging due to poor understanding of minimal growth media constituents, though micromanipulation has circumvented this (Lemos et al. 2008; Nittami et al. 2009; Soddell et al. 2006). However difficulties in culturing the isolates remain (Tolvanen & Karp 2011; Nielsen, Kragelund, et al. 2009). This has led to confusion and contradictory statements regarding the identity and ecophysiology of filaments (Naidoo et al. 2011; Seviour et al. 2008), such as whether or not species can grow anoxically (Nielsen, Kragelund, et al. 2009; Soddell & Seviour 1990). The use of next generation sequencing (NGS) circumvents all of the above as culturing is not required

and 16S rRNA analysis is not subjective unlike light microscopy based morphometric identification. NGS is a high throughput method which can enable large scale surveys of bulking plants and subsequent correlation of change in taxa abundance with bulking thereby identifying taxa responsible. In addition metagenomic or single cell whole genome sequencing can identify genes present in species without culturing, and by comparison with homologs give an indication of gene function and thereby species ecophysiology. However new sources of error are introduced (see sections 2.4 and 2.5) and therefore whilst NGS can provide additional insight it is not a panacea. The filamentous taxa that comprise the bulking group are polyphyletic and therefore have a diverse ecophysiology. Some of the taxa involved are described below to illustrate the diverse nature of the group.

Filaments thought to be important in bulking include *Microthrix parvicella , Spaherotilus natans, Thiothrix spp.* and types 021N and 1851. *M. parvicella,* a species in phylum Actinobacteria, is regularly observed throughout the globe and is associated with both bulking and foaming events (Nielsen, Kragelund, et al. 2009). Indeed it has often been found to be dominant in both (Martins et al. 2004; Nielsen, Kragelund, et al. 2009), and thus is a primary target for controlling such outbreaks. Whilst able to reduce nitrate, it is uncertain whether *M. parvicella* can sustain growth (Nielsen, Kragelund, et al. 2009) or not (Martins et al. 2004) under anoxic conditions. Specific physiology varies according to strain (and potentially species) under investigation. Some are able to take up only complex molecules such as lipids and long-chain fatty acids (Nielsen, Kragelund, et al. 2009), whilst others take up glucose, acetate and glycine amongst others (Rossetti et al. 2005). This is clear evidence for the need for clarification of *M. parvicella* physiology, especially given its apparent importance in both bulking and foaming.

S. natans (class *Betaproteobacteria*) is an obligate aerobe able to grow well in all ranges of dissolved oxygen (DO) and substrate concentration, taking up a number of readily biodegradable substrates (RBCOD) (Martins et al. 2004; Nielsen, Kragelund, et al. 2009). It's importance in bulking episodes remains unclear as it is frequently misidentified in morphological studies due to pleomorphism (Nielsen, Kragelund, et al. 2009). Molecular evidence suggests it is not usually present in significant numbers, however further surveys are required to confirm this (Nielsen, Kragelund, et al. 2009).

Type 021N is a morphotype including *Alpha- and Gammaproteobacteria* along with the newly identified *Thiothrix eikelboomii*, *T. disciformis* and *T. flexilis*. These taxa along with the other members of the genus are both mixo- and autotrophic, able to oxidise RBCOD and sulphide at high rates in conditions of low substrate concentration (Martins et al. 2004). Both aerobic and anoxic (nitrate reducing) respiration has been shown to occur in situ (Nielsen, Kragelund, et al. 2009). They are often the primary cause of bulking in plants treating industrial effluents, in which they are most commonly observed (Nielsen, Kragelund, et al. 2009). Some of the *Alphaproteobacteria* of Type 021N have been identified to species level, such as *Sphaeronema italicum*, however they have yet to be cultured and so their exact physiology remains unclear. It seems they oxidise only soluble substrates varying from mainly fatty acids to sugars and amino acids (Nielsen, Kragelund, et al. 2009). However in all cases they are able to take up substrate in anoxic conditions, resuming growth aerobically when conditions permit (Nielsen, Kragelund, et al. 2009).

The morphotype *Nostocoida limicola* appears to delineate a grouping of physiologically similar yet phylogenetically distinct OTUs, able to grow aerobically and fermentatively. *N. limicola* describes a number of unidentified gram-positive bacteria, as well as species from the Planctomycetes and Alphaproteobacteria, illustrating the diversity of this morphotype (Nielsen, Kragelund, et al. 2009; Martins et al. 2004). Other metabolically similar OTUs include Type 0961, 1863 and 1851 (*Chloroflexi*). The groups' flexibility may suggest a competitive advantage under some conditions, yet there is little evidence that abundance of the group is significantly correlated with bulking events (Martins et al. 2004).

There are many other filamentous OTUs known to be present in activated sludge, however these are omitted for purposes of brevity due to low abundance in bulking episodes (Nielsen, Kragelund, et al. 2009). Many of the above observations are based on FISH studies using probes known to hybridize with non-target organisms. Therefore misidentification may still occur. Morphological identification techniques are subject to a greater proportion of misidentification due to similar morphology in distinct species. There remains much confusion regarding fundamental issues such as whether the morphotype under investigation is in fact a single species and difficulty establishing in situ metabolism using axenic cultures. Therefore the need for studies which

investigate the full breadth of species present will greatly aid the understanding of bulking events.

3.6 Control of foaming and bulking

In both foaming and bulking, there are many possible approaches to prevention, with some more successful than others. Current control measures focus on generic approaches such as the application of chlorine, however it is now recognised that due to the diverse range of substrates which filaments can metabolise (Kragelund et al. 2007; Zang et al. 2008; Seviour et al. 2008; Yoon et al. 2010) it may not be desirable to eliminate them from activated sludge. This is of particular importance when considering the removal of endocrine disruptors and other novel pollutants (Soddell & Seviour 1990). Other chemicals used for control include polyaluminium chloride which flocculates dispersed filaments thus reducing foaming potential (Narayanan et al. 2010), and compounds such as 3-hydroxyhexanoic acid which inhibit the production of mycolic acid (Davenport et al. 2005) all with mixed success.

As little as 1% of sludge biomass as filaments can lead to bulking (Martins et al. 2004), however the causative environmental conditions are poorly understood. It is thought that increasing the food to microorganism ratio (F/M) can help control bulking and foaming as the causative species have a low substrate affinity for readily biodegradable compounds (Ehlers et al. 2012). Selectors are based on this theory; a plug flow system is created such that the influent enters a region of high RBCOD, ensuring floc formers outcompete filaments (Chudoba et al. 1973). Mixed results have been observed in applying this method, with little data available illustrating potential causes for success or failure (Martins et al. 2004; Nielsen, Kragelund, et al. 2009). Other non-specific control methods include increasing DO to favour floc formers (Rossetti et al. 2005) or adjusting sludge age to flush out slow growing species (Martins et al. 2004).

It has been suggested that if we can understand the biochemistry of each species, we could model the system and thus create an environment unfavourable to foaming and bulking. The difficulty with this approach is that this would require investigation of each species' substrate affinity for all substrates under varying environmental conditions such as temperature, pH, presence of protagonists or antagonists and concentration of innumerable chemicals in the influent amongst other factors. The

author does not believe this approach is viable due the complexity and variability of influent composition and the large number of species interactions involved. Instead I suggest we identify the species causing bulking or foaming and develop specific control measures such as phage application (Withey et al. 2005; Lukacik et al. 2012; Petrovski et al. 2012).

3.7 Removal of micropollutants

The removal of endocrine disrupting compounds (EDCs) from influent is a novel problem brought about by the large number of chemical compounds being created each year; the current documented total exceeds 200,000 (European Commission 2008). A significant number of these compounds are known to disrupt the endocrine system in many animals (Vos et al. 2000). Of particular concern are compounds such as oestrogens and oestrogen mimics that lead to the feminization of males (Sonnenschein & Soto 1998; Hayes et al. 2002). Therefore there is a clear need to investigate the ability of microbes present in sewage treatment plants (STPs) to remove EDCs. It has been shown that STPs with a longer sludge age are often able to remove a larger proportion of EDCs (Clara et al. 2005; Falas et al. 2012), potentially due to the greater microbial diversity observed in such plants (Ziembinska et al. 2007).

As alluded to above, the significant diversity of both EDCs and the bacteria which oxidise them result in difficulties in studying and thus describing this group. Indeed when investigating only one natural oestrogen (E1), Zang et al (2008) found that species from the Alpha-, Beta- and Gammaproteobacteria were capable of assimilating the compound. Some broad conclusions can be made, such as EDC removal is strongly linked to heterotroph abundance and diversity (Falas et al. 2012). It would also seem that for many EDCs there is a species capable of its assimilation or mineralization (Navacharoen & Vangnai 2011; Isabelle et al. 2011; Sang et al. 2011; Wei et al. 2011; Liao et al. 2010). For example *Pseudomonas* strain MI14.1 is able to degrade oestrone (Isabelle et al. 2011), and *Deinococcus radiodurans* is able to degrade di-n-butyl phthalate (Liao et al. 2010). However encouraging these populations within an AS plant may prove challenging, as demonstrated by several failed attempts at bioaugmentation (Boon et al. 2000; Bouchez et al. 2000; Tchelet et al. 1999).

3.8 Operational utility of molecular methods in activated sludge process control

The diversity of filamentous bacteria involved in foaming and bulking creates difficulty in designing FISH probes which accurately target taxa of interest. The number of probes required to target phylogenetically divergent taxa of importance in foaming such as *Microthrix* (Phylum Actinobacteria) and Type 0803 (Phylum Chloroflexi) is infeasibly large. A further difficulty is the low throughput and highly specialised nature of FISH analysis. These two factors combine to prohibit its use at individual treatment plants, and central sample analysis would require a large team producing results several days after the event occurs. qPCR whilst more rapid would require a large number of probes and concordant standard dilutions of target taxa. For routine use, analyses must be as cheap and simple as standard chemical tests such as ammonia test kits. Therefore NGS also fails to fit these criteria. However, NGS is more cost effective than qPCR when considering the wide range of taxa that can be enumerated per run and when sufficient samples are multiplexed can achieve higher throughput. Therefore NGS can provide useful insights when troubleshooting long term operational issues such as repeated foaming. NGS is particularly suitable for research projects which require community surveys, such as establishing the metabolic potential of the microbial community across several ASPs. In this way NGS can contribute to the efficient operation and process improvement of ASPs.

Chapter 4. Microbial Source Tracking Literature Review

4.1 Concept

In 1854 John Snow established that the source of a Cholera outbreak was a public water pump in Broad Street, Soho London. By doing so he invented the field of epidemiology and established the link between faecal contamination of drinking water and the spread of gastrointestinal disease. This new understanding highlighted the importance of minimising public exposure to faecal contamination. As public health in the developed world continued to improve, the proportion of disease burden arising from contact with water borne pathogens began to shift from potable water to a broad range of environments. In the 1970s the European Economic Community began to focus on the quality of the water in which we swim, after it was established that gastrointestinal disease was common after bathing (Brown 1964). In 1976 the Bathing Water Directive was introduced in order to further protect public health. The directive introduced limits on the abundance of faecal indicator organisms in bathing waters, including *E. coli* and Enterococci (Table 2).

Bathing water quality is significantly improved today (European Environment Agency 2015). However the impact of faecal contamination of bathing waters on public health remains significant and is estimated to result in global annual economic losses of $12bn and three million disability-adjusted life years per annum (Shuval 2003). In order to further reduce faecal contamination it is necessary to establish the relative contributions from point and diffuse sources. When the 1976 Bathing Water Directive came in to force, the primary source of faecal contamination in bathing waters was the discharge of insufficiently treated sewage (Jones & Obiri-Danso 1998). By 1998, of the 9.5 million UK residents whose sewage discharged to sea 11 percent was untreated, 68 percent primary or secondary treated and 21 percent tertiary treated (Water Services Association 1998). It is now becoming clear that agricultural runoff and non-human faeces are significant sources of faecal indicator bacteria in bathing waters (Converse et al. 2012; Dickenson & Sansalone 2012; Edwards et al. 2012). Microbial source tracking attempts to elucidate these sources and their relative contribution to faecal bacteria in the bathing water.

Since privatisation, the UK water industry has invested over £2.5bn in bathing water quality improvement schemes (Department for Environment Food and Agriculture

2015). These have largely involved minimising the frequency and volume of combined sewer overflow spills, as well as installing secondary and in some cases tertiary sewage treatment at coastal works. It is important to understand what proportion of faecal indicator bacteria currently found in bathing waters arises from sewage, and the relative contributions from each point source in a catchment, in order to direct future asset investment strategies. In industry this is primarily achieved by building models which reflect the hydraulics of wastewater networks and simulating the response of the network under varying rainfall conditions (Waites 2012). However this approach relies on estimates of faecal influx from agricultural runoff and the volume of surface water entering the system during real world storm events. The construction of a small storm water retention tank costs approximately £6m (Waites 2012) and may not significantly reduce FIB in the bathing water. An improved approach is needed to increase asset investment efficiency, and achieve maximum reduction in public health risk per unit cost spent on the catchment. MST can assist in catchment management by identifying the proportion of faecal contamination from diffuse and point sources. Clarification of the proportion of contamination from human sources would inform cost benefit analyses when planning interventions such as storm tank installation. Conversely with clear data identifying the impact of diffuse pollution on bathing water quality land owners may be more motivated to take action. This would provide wider benefits since run off contaminated with faecal bacteria from farms may contain other substances such as phosphorus which are relevant to the overall ecological status of the catchment.

4.2 Legislation

The original Bathing Water Directive (1976) required EEC member states to identify bathing waters, instigate a monitoring programme and ensure bathing waters met minimum criteria by 1986. These criteria were specified at two levels – guideline and mandatory (Table 2). Samples were required every fortnight where 95% of samples must have met the mandatory standards.

Table 2: Microbiological parameters for bathing water quality in EEC Bathing Water Directive (1976)

Parameter	**Guideline**	**Mandatory**
Total coliforms / 100ml	500	10000
Faecal coliforms / 100ml	100	2000
Faecal streptococci / 100ml	100	-
Salmonella / 1L *	-	0
Enteroviruses PFU/ 10 L *	-	0

PFU = Plaque forming units. Those parameters marked * are checked only where there is suspicion that an event has occurred which would cause deviation from the required standard.

In 2006 the revised bathing water directive came in to force, and featured new bacteriological water quality parameters (Table 3) designed to simplify legislation and further reduce the risk to public health from bathing (European Commission 2016). This change was in part politically motivated by public demand for greater environmental information access and protection. The revised bathing water directive proposal document highlighted the fact that Enterococci and *E. coli* abundance thresholds chosen were a political rather than scientific decision (European Parliament and Council 2012). Cost benefit analysis broadly supported the introduction of more stringent standards though this was not unequivocal (Georgiou & Bateman 2005). This uncertainty may reflect the different risks posed by viruses and bacteria from faecal sources. Wyer et al (2012) showed that FIB appear to be inactivated in seawater whilst human adenovirus is not, suggesting FIB may not be suitable indicators of coastal bathing water health risk. This is highlighted by surveillance studies showing that enterovirus can be detected even when FIB standards are met (Hughes et al. 1992; Muscillo et al. 1997). However the aetiological agent in bathing water acquired gastroenteritis is not clear (Fleisher et al. 1996), and a meta-analysis by Wade et al (2003) found that Enterococci abundance correlate with gastroenteritis relative risk. The public health risk associated with bathing in faecally contaminated waters should also take account of diseases such as hepatitis (Gammie 1997). The revised directive proposal document highlighted the need for further research in to the risk from viruses and their correlation with FIB (European Parliament and Council 2012)

The new directive featured five parameters (Enterococci CFUs / 100ml, *E. coli* CFUs / 100ml, pH between 6 and 9, and visual inspection for oils and debris) as opposed to 19 parameters in the 1976 directive. Public health protection was further improved by reducing maximum acceptable *E. coli* abundance from 2000 to 500 per 100 ml and introducing *Enterococci* abundance as an additional indicator of faecal contamination. However as samples are not replicated, results may not reflect true contamination levels. Variance in FIB abundance in ten minutes can be greater than the difference between bathing water classification levels (Boehm 2007).

The EU recognised that the changes required would take time to implement and thus the 2006 directive would not fully replace the 1976 directive until 2016 (European Parliament 2006). The revised directive establishes four classification levels (Table 3) and requires public signage indicating the most recent results and whether bathing is recommended. Where a bathing water fails to meet the "Sufficient" standard, it will be classified as "Poor" and bathing will not be advised. If a bathing water is classified "Poor" for five consecutive years it will be permanently closed to bathers. Should an EU member state fail to adhere to the requirements of the directive, the member may be fined. However this has occurred only once in 2003 (Court of Justice of the European Union 2003) when Spain was ordered to pay €624,150 per year per 1% of bathing waters not conforming to the mandatory parameters (Table 2).

Table 3: Required parameters for bathing water quality in EU Revised Bathing Water Directive (2006)

Parameter	**Excellent**	**Good**	**Sufficient**
Intestinal enterococci (CFU/100ml)	100	200	185
E. coli (CFU/100ml)	250	500	500

Excellent and Good status is based on 95% of samples meeting the stated parameter standards, whilst Sufficient status is based on 90% of samples meeting the stated parameter standards.

4.3 Techniques

4.3.1 Phenotype analysis

The general principles and methods for culturing bacteria are discussed in chapter 2.1. *E. coli* and intestinal *Enterococci* are faecal indicator bacteria (FIB) that are cultivated and enumerated to quantify the risk to public health from faecal contamination of

bathing waters (see section 4.2). This data provides little information on the potential source of contamination. Early attempts to track the source of these FIB included antibiotic resistance assays (ARA) and carbon source utilisation (CSU). Wiggins et al (2003) used the ARA approach to classify the source of FIB isolates with 57% average rate of correct classification and a 25% minimum detectable percentage of contamination. Correct classification rates of up to 95% have been claimed using the ARA method (Hagedorn et al. 1999), however the accuracy of this data has been questioned (Stoeckel & Harwood 2007). The CSU approach has been rarely used and found to have similar correct identification challenges as ARA. (Wallis & Taylor 2003; Donald M. Stoeckel et al. 2004) Due to the large number of isolates required to build libraries and the associated labour and cost, these methods are no longer pursued.

4.3.2 Sterols

Faeces contains many sterol compounds which can be quantified using methods such as gas chromatography. The cumulative ratios of some sterols are thought to be specific to the host animal, and therefore can be used to identify the source of faecal contamination (Furtula et al. 2012; Gilpin et al. 2002; Gilpin et al. 2003). However these ratios work only in simple watersheds, such as those impacted only by treated human sewage. Where there are multiple faecal inputs, the method fails due to accumulation of sterols from different sources confounding the end ratio (Shah et al. 2007). Recent developments have used principal component analysis to group sterol profiles between known and unknown samples, thereby identifying source (Solecki et al. 2011). Sterol profiles may be altered by sewage treatment (Furtula et al. 2012), sterol decay (Solecki et al. 2011) and diet (Derrien et al. 2011). Therefore sterol profiles cannot be used quantitatively. A comparison of multiple MST approaches found that sterols do not improve upon microbial methods (Blanch et al. 2006).

4.3.3 Mitochondria

During defecation epithelial cells from the lining of the intestine are passed out with the faeces (Iyengar et al. 1991). These host cells contain mitochondria which harbour genetic content unique to the host species. Therefore mitochondrial genome sequencing can be used to identify the source of faecal contamination. Kapoor et al (2013) sequenced the human mitochondrial hypervariable region 2 to track faecal contamination in an urban creek system. It was found that human mitochondrial

abundance correlated with *Bacteriodales* 16S rRNA sequence abundance, suggesting that mitochondria are a useful tool in tracking faecal source. However the catchment investigated was relatively simple, with no agriculture or recreational use and therefore it is unclear whether mitochondrial data correlates with FIB abundance in complex catchments.

It has been suggested that this approach is of limited utility because there are many factors that can confound the presence and abundance of mitochondrial DNA . These include:

- Mitochondria are present in skin cells sloughed off during recreation in water bodies, and therefore the presence of mitochondria from a species does not necessarily indicate faecal contamination.
- There are between one and five thousand mitochondrial genome copies per epithelial cell (D'Erchia et al. 2015) and few studies investigating the number of epithelial cells shed per host per day other than in humans (Kamra et al. 2005). Thus the abundance of mitochondrial DNA may not be directly related to abundance of faecal bacteria.
- A mitochondrial signal from consumed meat can be present in the consumers' faeces, for example when a human eats pork. Martellini et al (2005) observed this phenomenon in untreated sewage from a catchment with no animal husbandry present.

Therefore given that the presence of mitochondria from a host does not necessarily indicate faecal contamination from that host, and that the abundance of mitochondrial gene copies does not necessarily correlate with the extent of faecal contamination, this approach was not pursued.

4.3.4 Host associated bacteria

It is thought that some bacterial species exist only in the guts of certain groups of animals, such as *Catellicoccus marimammalium* in gulls (Roslev & Bukh 2011; Seyrig et al. 2011). Therefore the detection of these host-associated bacteria can indicate the presence of faecal contamination by specific host groups.

By designing primers which target the 16S sequence unique to host-associated taxa, the polymerase chain reaction (PCR) can be used to identify the presence or absence

of faecal contamination from a host. Quantitative PCR (qPCR) can also quantify host-associated bacteria abundance, however the abundance of host marker bacteria in the host gut may not be stable across or within hosts (Roslev & Bukh 2011). In particular this is problematic in that bathing water quality is regulated on the basis of faecal indicator bacteria (FIB) abundance (see section 4.2), and therefore qPCR data cannot be used to apportion FIB to hosts. The quantitative polymerase chain reaction (qPCR) is described in section 2.2.1. The use of the 16S rRNA gene for the identification of taxa is discussed in section 2.4.1.

The marker bacteria approach has been used with some success across a number of catchments, with the relationship between host and bacteria occurring globally (Reischer et al. 2006; Flynn et al. 2016; Diston et al. 2015; Oyafuso et al. 2015; Ridley et al. 2014). However qPCR is thought to be sensitive to chemical contamination which inhibits PCR amplification and therefore biases results. Ebentier et al (2013) found that coefficient of variation was low (< 7%) when protocols and reagents were standardised, yet Oliver et al (2014) found that results from qPCR were variable and sometimes contradictory. A large scale study reported that qPCR data interpretation was a significant source of variance in assay performance (Boehm et al. 2013).

Next generation sequencing (NGS) uses end point PCR which is less sensitive than qPCR to chemical contamination and therefore inhibition (Schrader et al. 2012). However NGS introduces new uncertainty due to sequencing error or misidentification due to bioinformatics issues. See section 2.4 and 2.5 for further discussion.

Digital PCR is end point PCR multiplexed across a large number of microwells whose volume is such that a single DNA molecule is present in each well. The number of microwells in which DNA replication takes place is approximately equivalent to the number of target DNA molecules in the sample. Therefore digital PCR can achieve qPCR-like quantification without introducing error from partial inhibition and comparison against standard curves (Hindson et al. 2011). Loop mediated isothermal amplification (LAMP) is a DNA amplification technique which does not require expensive and delicate thermal amplification devices as in PCR, qPCR and dPCR (Notomi et al. 2000). Furthermore LAMP amplification can be identified by eye or simple optics, reducing the cost of field suitable apparatus (Mori et al. 2001). However

LAMP requires multiple primers per reaction and multiplexing can be challenging (Tanner et al. 2012). Further work is required to validate the utility of LAMP for field MST.

4.3.5 Whole community

Whole community microbial source tracking attempts to circumvent the difficulties of proving unique host-species taxa association by identifying community wide features specific to each source. One approach uses a marker gene sequence (such as the 16S rRNA gene) along with the relative abundance of each sequence to produce a pattern unique to each source. This method can identify faecal contamination source but is not quantitative as the test is based on the presence or absence of a pattern. By comparison of the patterns, it is possible to match faecally contaminated samples (the "sink") to faecal source. Cao et al (2013) employed terminal restriction fragment polymorphism and phylogenetic microarrays to generate such patterns and thereby attempt to match challenge samples to a source. They also used next generation sequencing (NGS) to identify 16S rRNA sequences unique to faecal sources, as well as using OTU (see chapter 2.5.2) taxonomy and relative abundance to cluster faecal source and challenge samples by principal component analysis. Challenge samples were composed of single or dual source composite samples in comparison to 12 reference samples from single host species. All three methods were found to be of use in differentiating sources, each correctly categorising 95% of samples. The statistical approach for NGS used Bray-Curtis distance to establish similarity between samples and therefore did not allow for proportional contribution of different sources to the sink.

Knight et al (2011) further developed the OTU approach by modelling the sink as a mixture of sources, where each observation of an OTU is assigned to a source. The probability of each OTU observation in one source being observed in the sink, and the probability of each OTU observation in the sink coming from that source are both Dirichlet distributions. Bayesian Gibbs sampling of the joint distribution enables the estimation of the contribution to the total sink community from each source. Uniquely this tool allows for a proportion of OTUs in the sink to be assigned to an unknown source, thereby enabling the analysis of environmental sinks where all sources cannot easily be sampled. Newton et al (2013) employed this tool in a wastewater context by

using human faecal samples as sources and both sewage influent and harbour water samples as sinks. They created a human faecal "signature" by identifying those OTUs in the sinks which were assigned to human faecal origin with greater than 10% probability. This low threshold was used due to the high variability observed in faecal samples from individual humans. This limited the power of the study, as did the lack of any samples from non-human faecal material. However it was possible to subtract the human faecal OTUs from non-human, giving an indication of the proportion of total faecal contamination coming from humans. Using samples from sewage outfalls as opposed to individual humans may provide better results, given that the large number of individuals per catchment and the relatively stable sewage treatment community may lead to a more stable final effluent community. In addition, it is the contribution to faecal contamination from sewage outfalls as opposed to raw human faeces which is relevant to the water industry.

Henry et al (2016) analysed the impact of SourceTracker parameters on prediction accuracy, precision and repeatability. They also used SourceTracker to estimate the source proportions in an artificially created sink sample from raw sewage and duck faeces, and then applied their findings to two catchment in Australia. Henry et al established that default parameters produce the best results (average sensitivity >97%, avg specificity 80%, avg precision 79%, avg accuracy 87%) unless approximate sink proportions are known in advance. They noted that SourceTracker occasionally detected non-existent sources at very low proportions (<1%) with high variability. Therefore they concluded that by running the model five times and calculating relative standard deviations (RSD), it was possible to exclude these false positives by excluding sources whose RSD was greater than 100%. The analysis of two bathing water catchments largely mirrored expectations in terms of source proportions detected, with human sources dominating but at low levels. An interesting finding was the persistent treated effluent signal in sand samples, highlighting the potential for delayed impacts from sewage due to intense wave action.

Further work is required in order to establish whether whole community analysis can differentiate between similar sources such as different sewage outfalls in the same catchment. Whole community analysis can be of use where a unique host associated bacteria can not be identified, however they are most powerful when used in

combination. A significant limitation of whole community analysis is the additional sampling and sequencing effort and therefore cost of characterising each potential source in a catchment.

4.3.6 Whole genome

Several studies have investigated whether genomic divergence within an OTU may result in detectable differences between e.g. *E. coli* from cows and sheep (Gomi et al. 2014; Fu et al. 2011). This could enable direct MST of FIB and thereby avoid the marker to FIB ratio problem, enabling absolute quantification of FIB per source (see 4.3.4). Gomi et al (2014) sequenced 22 *E. coli* whole genomes from humans, cows, pigs and chickens to identify regions of sufficient sequence divergence to enable source differentiation. They subsequently developed PCR primer pairs targeting these areas and attempted to classify 549 environmental isolates using this method. Only 48% of isolates could be classified, of which 4.4% had markers from multiple hosts. Gomi et al suggested this may have been due to insufficient sampling of sources or contribution from sources not considered in the study. However the group limited their search to significant variance in individual genes rather than cumulative divergence across the entire genome, which may also have limited their ability to distinguish sources. Sequencing the entire genome of an unknown isolate may find more subtle divergence such as single nucleotide polymorphisms which cumulatively would enable differentiation of a faecal host. It may even be possible to distinguish between host animal groups e.g. herds of cows, given that animals living in close contact harbour more similar gut genomes than highly dispersed animals (Song et al. 2013; Shanks et al. 2011; Thompson et al. 2008). Fu et al (2011) used (GTG)5 PCR to generate 1363 unique fingerprints from 1104 samples across 9 different host animals. This form of PCR uses primers for the bacterial GTG repetitive sequence to generate amplicons the size distribution of which forms a fingerprint unique to the taxa at the strain level (Versalovic et al. 1994). Fu et al (2011) attempted to classify isolates from 1600 water samples by hierarchical cluster analysis of the isolate fingerprints and their constructed host isolate fingerprint database. 3% of environmental isolates could not be classified, indicating the importance of analysing the whole genome. The rate of false negatives was not reported. Sampling effort when building their database was significantly higher than Gomi et al (1104 samples versus 33), which may explain the difference in classification rate.

4.3.7 Host associated viruses

Some viruses are specific to a host e.g. porcine adenoviruses (Bofill-Mas et al. 2011). Therefore detection of these viruses indicates contamination from a specific host. However quantifying this contamination in terms of proportional contribution per host and correlation with FIB can be challenging (Bofill-Mas et al. 2010) which may explain the paucity of studies using viruses for MST. For example human adenoviruses are present year round in human faeces, whereas rotaviruses are found more sporadically (Pina et al. 1998). Sample preparation of viruses is also more challenging than bacteria, as the virus particles must be concentrated from several litres of water using complex protocols. Intra-laboratory viral DNA extraction efficiency varies by 2 orders of magnitude (Girones et al. 2010) (Figure 4) and therefore this method cannot be used quantitatively. Further research is required in order to understand the utility and cost effectiveness of viral MST in comparison to other methods.

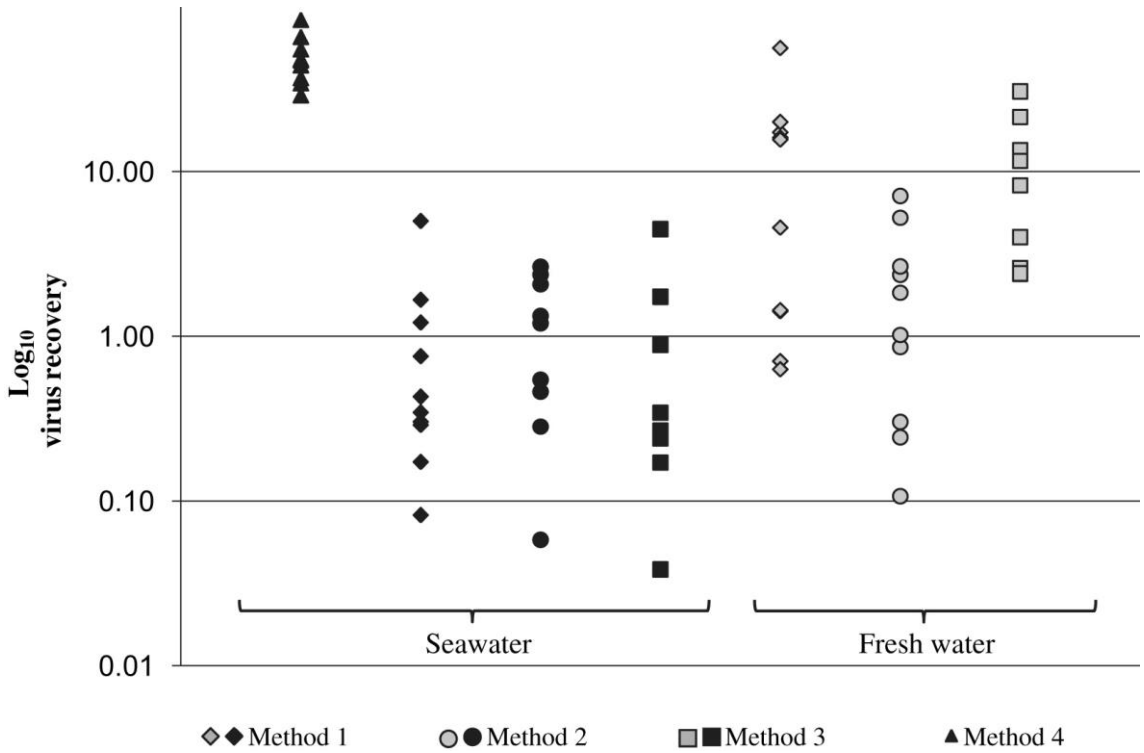


Figure 4: Intra-laboratory variability of viral concentration methods in artificial seawater and freshwater. Virus recovery values obtained after spiking sets of ten 10-l samples with HAdV 2, concentrating by: Method 1: electronegative filters of nitrocellulose and glycine 0.05 M pH 9.5 – skimmed milk buffer,2: electronegative filters of nitrocellulose and glycine 0.25 M pH 9.5 – beef extract buffer, 3: Glass wool column and glycine 0.25 M pH 9.5-beef extract buffer, 4: Direct organic flocculation with skimmed milk. Taken from Girones et al (2010).

4.3.8 Phage plaques

A bacteriophage is a virus which infects bacteria. It is thought that some phages are specific to certain hosts, such as group 2 and 3 F-RNA phages largely occurring in humans. However phage host specificity is unclear, and the utility of phages in differentiating between animal hosts other than humans is unclear, with a number of contradictory studies (Harwood et al. 2013; Diston et al. 2015; Diston & Wicki 2015; Purnell et al. 2011). In addition there seems to be differential survival between group 1 and 4 (animal specific) and group 2 and 3 (human specific) F-RNA phages thereby leading to false negatives for animal faecal contamination (Byappanahalli et al. 2012). Some studies have found that 16S rRNA gene data shows higher sensitivity to faecal contamination than phages (Diston et al. 2015), possibly due to UV phage inactivation (Diston et al. 2014). Given these complications and their low environmental abundance as for viruses, there appears to be few advantages for phage plaques over 16S rRNA gene sequencing.

In summary, the only MST method to date able to quantify the contribution to faecal contamination from individual point sources is whole community 16S rRNA sequencing. As this one method can also identify host marker bacteria using the same data, I suggest that 16S rRNA gene sequencing is a suitable tool for MST and will enable wastewater companies to quantify the impact of their assets on bathing water quality for the first time. However there are numerous research gaps including the temporal stability of whole community fingerprints from single point sources, the capability of whole community analysis to accurately apportion the contribution to faecal bacteria in a sink from point sources and a cost-benefit analysis of this approach.

Chapter 5. Methodological development

5.1 Sampling

30 ml grab samples of activated sludge were aseptically collected from the aeration basins (Eaton & Franson 2005) when the blowers were operating to ensure that the liquor was fully mixed (Table 4 and Figure 5). This is considered to be sufficient for collecting a representative sample (Bellucci & Curtis 2011). The samples were fixed on site with an equal volume of 0.2 micron filtered molecular grade absolute ethanol (Guo & Zhang 2013) and transported at 4 °C (International Standards Organisation 2003; Guo & Zhang 2012; Zhang et al. 2012). The samples were then stored at -20 °C until analysis (Guo & Zhang 2012; Juretschko et al. 2002). The sampling programme was designed to coincide with the Urban Wastewater Treatment Directive (UWWTD) sampling undertaken by NWL. UWWTD data includes ammonia concentration after the primary and secondary treatment stages, which along with flow and basin volume data enables the calculation of the mass of ammonia removed in the 24 hours prior to sampling. In addition to UWWTD data, plant operators were asked to classify the plant on the basis of nitrification stability and foaming propensity, where stable plants have not failed their ammonia consent in the last 12 months and unstable plants have suffered one or more failures.

Of the sampled ASPs (Table 4), those designed and operated to nitrify show ✔ in the nitrifying field. Plants with a centrifuge may show an increase in ammonia, solids and BOD added after inlet. There were 9 samples of foam and 11 samples of mixed liquor during a foaming event from 9 different plants. There were two mixed liquor samples during bulking events from two plants.

Figure 5: Map showing location of activated sludge plants sampled

Table 4: Activated sludge samples sites

Name	Site code	Owner	P.e.	Nitrifying	Centrifuge	Problems	Notes
Washington	A1	NWL	74916			Prone to foaming and bulking (3 samples)	
Tudhoe Mill	A2	NWL	22493	✔	✔	Foaming sample	Sample 1 in DNA extraction comparison
Amble	A3	NWL	16607				SBR. Sample 2 in DNA extraction comparison
Aycliffe	A4	NWL	61106	✔			Tertiary. Significant industrial influent. Sample 3 in DNA extraction comparison.
Berwick	A5	NWL	15537				
Billingham	A6	NWL	35293			Foaming sample	
Blyth	A7	NWL	37859	✔			SBR
Bran Sands	A8-A	NWL	391142	✔	✔		Industrial (coke & acid)
Bran Sands	A8-B	NWL	391142	✔	✔	Foaming sample	Municipal
Bran Sands	A8-C	NWL	391142		✔	Foaming sample	Municipal and industrial
Broomhaugh	A9	NWL	7095				
Browney	A10	NWL	21586	✔		Foaming	
Cambois	A11	NWL	28655				
Cramlington	A12	NWL	45309				
Haggerston Castle	A13	NWL	2040	✔			
Hendon	A14	NWL	229108		✔		
Hexham	A15	NWL	29714				60/40 flow to ASP/ Trickling Filter
Horden	A16	NWL	100299				
Howdon	A17	NWL	947811		✔		
Marske	A19	NWL	93556				SBR
Newbiggin	A20	NWL	38487			Foaming	

Seaham	A21	NWL	23595				
Seahouses	A22	NWL	11213				SBR
Seaton Carew	A23	NWL	120222				
Sedgeletch	A24	NWL	51152	✔			Tertiary
Branxton	A25	NWL				Foaming	Descriptive
Bowsden	A26	NWL					Descriptive
Worsley	A27	UU	18133	✔		Failed NH3 by 0.3	
Oldham	A28	UU	154000	✔		Foaming sample	
Kidsgrove	A29	UU	26000	✔		Foaming sample	
West Houghton	A30	UU	33000	✔		Bulking sample	
Skelmersdale	A31	UU	51000	✔			
Horwich	A32	UU		✔		Failed NH3 by 4.5 mg/L	
A wastewater treatment plant must have a consent to discharge to a water course. Where the population equivalent (the number of people whose waste the plant treats or the equivalent where there is an industrial input) is less than 250 the consent is "descriptive" i.e. a piece of text stating that the discharge must have no visual impact on the water course. Where the population equivalent is over 250, the consent will be "numerical" and state the maximum permissible values for a number of determinants, such as biological oxygen demand and suspended solids. Where a numerical consent includes a maximum ammonia concentration the plant may be required to nitrify (see section 3.3)							

The microbial source tracking work focussed on the Saltburn catchment in North Yorkshire. There are three principal water courses which discharge in to the bathing water: Skelton Beck, Saltburn Gill and Pit Hills Stell. Skelton Beck and Saltburn Gill form a confluence approximately 100m upstream of the bathing water (Figure 6, Table 5). The sewer network drains a total of 385 hectares extending 8km inland with a population of 22,418. In dry weather conditions the majority of sewage is pumped to Marske-by-the-Sea for treatment and UV disinfection. The small rural villages of Dunsdale and Upleatham have separate sewage networks and treatment plants that discharge in to Skelton Beck. There is also a private treatment works treating sewage from the Tockett's Mill holiday park which discharges in to Skelton Beck. There are 24 combined sewer overflows (CSOs) which discharge indirectly in to the bathing water via the river network. Figure 6 is a schematic diagram showing the river network and NWL wastewater assets in the catchment.

The catchment was selected as historic bathing water quality data indicated that the beach was at risk of being rated "Poor" under the revised bathing water directive. The bathing water quality was rated "Sufficient" in 2011, "Poor" in 2012 and "Sufficient" in 2013 at which point the catchment was selected. Subsequently the bathing water was rated "Good" both in 2014 and 2015 after planned asset interventions which included the creation of additional storm storage at Guisborough CSO and the diversion of Dunsdale wastewater treatment plant effluent to Marske for UV disinfection and subsequent discharge through a long sea outfall.

Table 5: Saltburn bathing water subcatchment animal density per hectare (ha). STW = Sewage Treatment Works, CSO = Combined Sewer Overflow.

	No. of STWs	No. of CSOs	Catchment area (ha)	Cattle	Sheep	Pigs	Total
Skelton Beck	3	13	5503	0.36	0.37	0	0.73
Saltburn Gill	0	8	1891	1.53	0.35	4.23	6.11
Pit Hills Stell	0	3	275	0.45	1.94	0.72	3.11

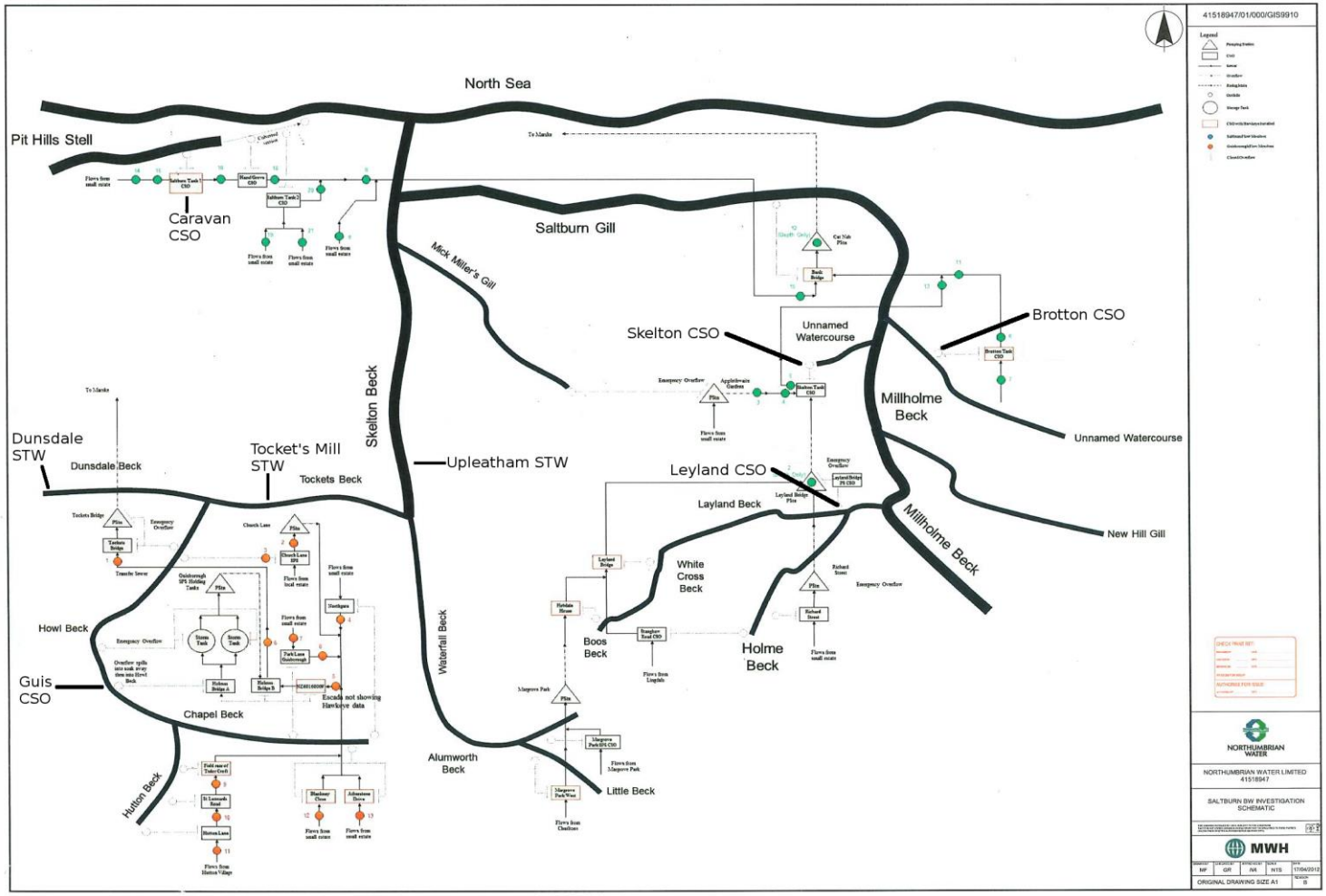


Figure 6: Schematic of the Saltburn catchment and NWL assets. Triangle = pumping station. Rectangle = combined sewer overflow (CSO). Large circle = storage tank. Small circle = outfall. Solid line = sewer. Dashed line = rising main. Adapted from MWH Saltburn bathing water report 2012.

It was decided that the regular sampling regime should match that of the Environment Agency (EA) to enable comparison of the results with the EA regulatory results. Therefore each planned sampling effort began with a sample taken from Saltburn bathing water at the same time and place as the EA. On the same day a further nine samples were taken from the catchment at strategic locations within the catchment to establish the contribution to faecal contamination in the bathing water from known faecal sources, as well as to partition the contributions from the Skelton Beck, Saltburn Gill and Pit Hills Stell sub catchments (Table 6 and Figure 7). On five occasions, simultaneous additional samples were taken (see column NWL labs in Table 6) and sent to Northumbrian Water Limited laboratories for the selective culturing and quantification of *E. coli* and Enterococci using the same method as the EA (Environment Agency 2012). This was undertaken to provide abundance data for faecal indicators in comparison with Personal Genome Machine (PGM) data and to keep isolates for potential whole genome analysis in the future. A total of ten samples per sampling event was chosen as this was the maximum number of samples the Northumbrian Water laboratory could process. However more than ten strategic sampling points were identified, and therefore there was a "core" set of sampling points and an "additional" set which were sampled at varied times indicated in the results section. Planned sampling events were considered "non-storm" samples if they were taken more than 48 hours after CSO flows. This time was chosen based on empirical observation of EA sample results in combination with CSO flow notifications. In addition the impact of any potential residence greater than 48 hours would be limited by microbial decay dependant on solar radiation levels (Whitman et al. 2004). Due to lack of data for river flow and morphology no attempt to calculate hydraulic residence times was made.

In addition to planned sampling events, samples were collected during storm events from combined sewer overflows (CSOs). This enabled the quantification of the impact of CSOs on bathing water quality before, during and after a storm. During storm events sampling was attempted at five CSOs identified by previous hydraulic studies of the catchment as primary contributors to faecal pollution in the bathing water (see Table 6). In addition, samples were taken from the river network and the bathing

water itself on the same day. Further details regarding these samples can be found in the results section.

5.1.1 Samples taken from the bathing water

Samples were collected as per EA procedures (Adam & Walmsley 2001) from Saltburn coastal waters at latitude 54.59, longitude -0.9698, where possible at the same time as those taken by EA sampling staff. The EA sample bathing waters once per week during the 20 week bathing period that runs from May to September. A single sample is collected aseptically according to the Annex V of the EU Bathing Water Directive (European Parliament 2006). Approximately 100 ml of seawater was collected in to a 250 ml sterile pot to which an equal volume of 0.2 µm filtered absolute ethanol was added. Ethanol is a widely used sample preservation reagent which minimises changes in microbial community structure during sample transportation and storage (Hale et al. 2015; Murphy et al. 2002). PGM cost limitations dictated a target of 10^5 reads per sample, which is considered sufficient to be representative of the microbial community structure in the sample (Caporaso et al. 2011; Lemos et al. 2011). Samples of 100 ml were deemed sufficient given that there would be approximately 10^8 bacteria (10^6 bacteria per ml in sea water) in the sample. However the PGM limit of detection, calculated as estimated number of bacteria in the sample divided by the target number of reads would be 10^3 bacteria per sample as opposed to the EA limit of 10 *E. coli* per sample. As the NGS analysis was not intended as a replacement for culture based FIB enumeration, the discrepancy in lower limits of detection was not relevant. The samples were transported at 4°C and stored at -22°C until filtration.

5.1.2 Other aqueous samples

For rivers, CSOs and aqueous sources other than the bathing water (Table 6), approximately 100 ml of liquid was collected aseptically in to 250 ml containers to which an equal volume of 0.2 µm filtered molecular grade absolute ethanol was added. The samples were transported at 4°C and stored at -22°C until filtration.

5.1.3 Samples from solid matrices

For faecal matter, sand from the bathing water site, and agricultural samples from the catchment (Table 6), a 60 ml sterile pot was half filled aseptically by hand and topped up with 0.2 µm filtered molecular grade absolute ethanol, thoroughly mixed, and

transported at 4 °C. The sample was subsequently stored at -22 °C until DNA extraction.

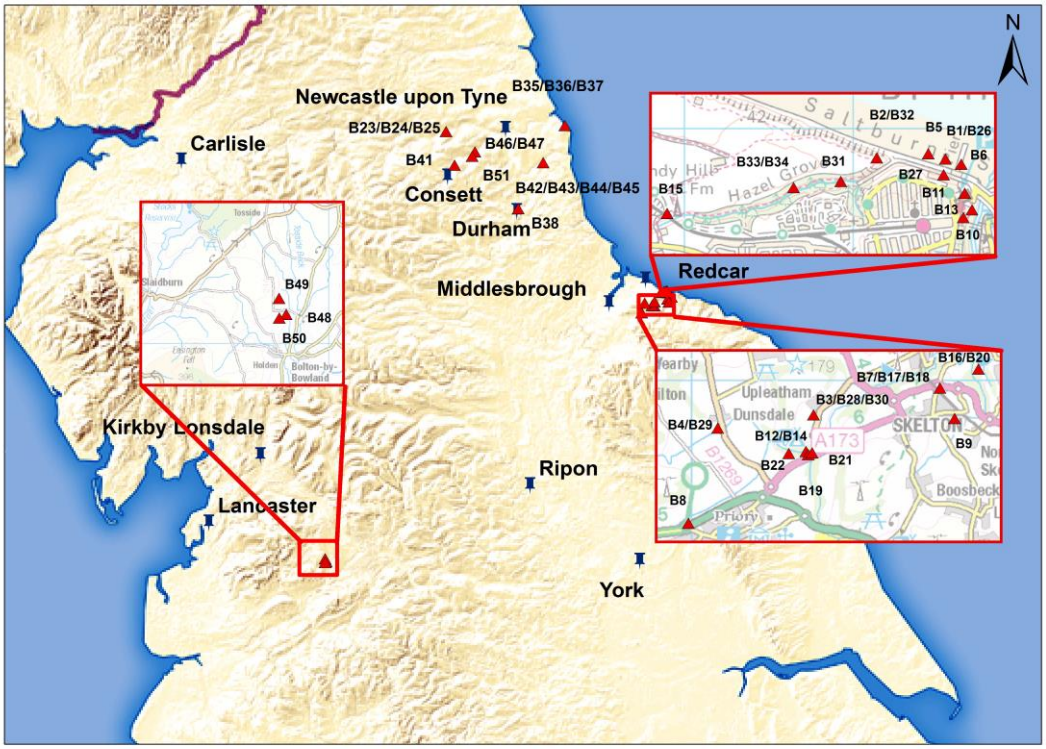


Figure 7: Map showing the location of sampling points in the Pit Hills Stell, Skelton Beck and Saltburn Gill catchments, as well as faecal samples from across the North of England. Pins show a selection of conurbations. Red triangles show the sampling points with their respective identification code (see Table 6).

Table 6: Sampling points, locations and types of samples taken for microbial source tracking. CSO = combined sewer overflow. NWL = Northumbrian Water Limited, where sample points marked "yes" were sampled at least once for microbial culturing at NWL laboratories.

Sample code	Sample point name	Sample source	Sample type	Latitude	Longitude	NWL labs
B1	Saltburn Bathing Water	Bathing Water	Aqueous	54°35'13.45"N	0°58'12.46"W	Yes – Core
B2	Pit Hills Stell outfall	River	Aqueous	54°35'13.85"N	0°58'35.14"W	Yes – Core
B3	Upleatham	Sewage Treatment Works	Aqueous	54°33'46.27"N	1° 1'18.27"W	Yes – Core
B4	Dunsdale	Sewage Treatment Works	Aqueous	54°33'36.02"N	1° 3'40.74"W	Yes – Core
B5	West of Saltburn bathing water	Seawater	Aqueous	54°35'14.50"N	0°58'18.09"W	
B6	East of Saltburn bathing water	Seawater	Aqueous	54°35'12.36"N	0°58'7.24"W	
B7	Skelton tank	CSO	Aqueous	54°34'7.74"N	0°58'10.06"W	Yes
B8	Guisborough	CSO	Aqueous	54°32'14.07"N	1° 4'25.87"W	Yes
B9	Leyland Bridge	CSO	Aqueous	54°33'41.31"N	0°57'48.90"W	Yes
B10	Skelton Beck	River	Aqueous	54°35'2.14"N	0°58'6.86"W	Yes – Core
B11	Skelton Beck / Saltburn Gill confluence	River	Aqueous	54°35'6.81"N	0°58'6.30"W	Yes – Core
B12	Tockett's Mill surface water	Run off	Aqueous	54°33'14.80"N	1°1'30.40"W	Yes
B13	Saltburn Gill	River	Aqueous	54°35'3.63"N	0°58'3.95"W	Yes – Core
B14	Tockett's Mill	Sewage Treatment Works	Aqueous	54°33'14.80"N	1°1'30.40"W	Yes
B15	Pit Hill Stell upstream of allotments	River	Aqueous	54°35'3.75"N	0°59'44.45"W	Yes
B16	Brotton	CSO	Aqueous	54°34'23.68"N	0°57'12.67"W	Yes
B17	Skelton land drain	Run off	Aqueous	54°34'7.74"N	0°58'10.06"W	
B18	Skelton road drain	Run off	Aqueous	54°34'7.74"N	0°58'10.06"W	
B19	Waterfall Beck	River	Aqueous	54°33'12.01"N	1° 1'26.83"W	Yes
B20	Brotton surface water	Run off	Aqueous	54°34'23.68"N	0°57'12.67"W	

B21	Skelton Ellers Beck	River	Aqueous	54°33'13.52"N	1° 1'21.02"W	Yes
B22	Tockett's Beck	River	Aqueous	54°33'13.19"N	1° 1'55.66"W	Yes
B26	Sand under pier (wet)	Sand	Solids	54°35'13.45"N	0°58'12.46"W	
B27	Sand adjacent to path (dry)	Sand	Solids	54°35'10.43"N	0°58'13.10"W	
B28	Cow pat from Upleatham	Faecal	Solids	54°33'46.27"N	1° 1'18.27"W	
B29	Dunsdale inlet	Raw sewage	Aqueous	54°33'36.02"N	1° 3'40.74"W	
B30	Upleatham inlet	Raw sewage	Aqueous	54°33'46.27"N	1° 1'18.27"W	
B31	Pit Hills Stell top of culvert	River	Aqueous	54°35'9.40"N	0°58'47.09"W	Yes
B32	Pit Hills Stell outfall (foam)	Foam	Solids	54°35'13.85"N	0°58'35.14"W	
B33	Caravan park	CSO	Aqueous	54°35'8.38"N	0°59'2.67"W	Yes
B34	Pit Hills Stell upstream of B33	River	Aqueous	54°35'8.38"N	0°59'2.67"W	Yes
B35	Marsden	Bathing water	Aqueous	54°58'38.09"N	1°22'30.15"W	
B36	Seagull	Faecal	Solids	54°58'38.09"N	1°22'30.15"W	
B37	Seagull + sand	Faecal	Solids	54°58'38.09"N	1°22'30.15"W	
B39	Chicken faeces	Faecal	Solids	Unknown UK	Unknown UK	
B40	Chicken litter	Faecal	Solids	Unknown UK	Unknown UK	
B41	Dog	Faecal	Solids	54°53'10.35"N	1°49'28.64"W	
B42	Horse 1	Faecal	Solids	54°53'26.98"N	1°27'54.23"W	
B43	Horse 2	Faecal	Solids	54°53'26.98"N	1°27'54.23"W	
B44	Horse 3	Faecal	Solids	54°53'26.98"N	1°27'54.23"W	
B45	Horse 4	Faecal	Solids	54°53'26.98"N	1°27'54.23"W	
B46	Bridge run off (liquids)	Run off	Aqueous	54°55'1.50"N	1°44'28.30"W	
B47	Bridge run off (solids)	Run off	Solids	54°55'1.50"N	1°44'28.30"W	
B48	Spring fed pool	Run off	Aqueous	53°57'29.14"N	2°20'10.17"W	
B49	Field drain	Run off	Aqueous	53°57'54.25"N	2°20'28.90"W	
B50	Sheep pen wash down	Run off	Solids	53°57'23.47"N	2°20'27.80"W	
B51	Poultry & horse manure	Run off	Solids	54°54'24.96"N	1°45'9.01"W	

5.1.4 Membrane filtration and sample processing

Samples were processed according to the United States Environmental Protection Agency standard for quantitative PCR enumeration of faecal indicators (Oshiro 2010). To summarise, samples were filtered through sterile 0.2 µm polyethersulfone membrane filters (Supor filters, Pall Corporation, Port Washington, USA) using a 250 ml glass membrane filter holder (P/N 16307, Sartorius, Gottingen, Germany) with glass frit membrane support. Before each use the filter holder was autoclaved and dried. After passing the sample through the filter, the holder was thoroughly washed down with 18 mΩ water three times to ensure no particles were retained on the walls of the filter holder. The filter was then folded according to the United States Environment Protection Agency protocol (United States Environment Protection agency 2012) and inserted in to a 2 ml Lysing Matrix E tube (MPBio, Santa Ana, USA) and stored at -22 °C until DNA extraction.

In the case of solids, 0.5 grams were added to a Lysing Matrix E tube (MPBio, Santa Ana, USA). The DNA from both aqueous (membrane filters) and solids were extracted using the "Karst et al + 55" protocol (section 5.3, Table 8). Primers targeting the V3-V4 hypervariable region of the 16S rRNA gene (section 5.2, Table 7) were used in PCR as described in section and the subsequent amplicons sequenced as described in section 5.4.

5.2 Primer selection

The 16S rRNA gene conserved and hypervariable regions differ in number of nucleotides and sequence variance (Klindworth et al. 2013; Guo et al. 2013; Mizrahi-Man et al. 2013). A single primer targeting a hypervariable region cannot encompass all bacterial clades (coverage) (Hong et al. 2009). This is as one unique sequence in a region may exist in only one species, whilst another unique sequence may exist in many species. The number of nucleotides dictates the number of unique DNA sequences the region can contain. Therefore the longer the region the greater the number of unique sequences and thus the more likely it is that a unique sequence is shared by fewer taxa. Therefore the taxonomic resolution which can be achieved by sequencing a region is generally greater for longer regions. The taxonomic coverage achievable by a primer pair cannot necessarily be predicted in silico (Morales & Holben

2009; Claesson et al. 2010) and therefore it is vital to empirically test potential primer pairs.

It is also important to ensure the amplicon produced by a primer pair will contain sufficient sequence diversity to differentiate between target and non-target species, and is of a suitable length for the chosen sequencing platform. Three primers, suitable for the Ion Torrent Personal Genome Machine (PGM) 200bp chemistry (Life Technologies, Carlsbad USA), were compared (Table 7). In order to multiplex samples in one sequencing run, 50 twelve base pair Golay error-correcting barcodes (Hamady & Knight 2009) were incorporated in to fusion primers along with the Ion Torrent sequencing adaptors (Life Technologies 2012). The delta G of these 50 unique oligonucleotide sequences were checked using OligoAnalyzer 3.1 (Integrated DNA Technologies 2015) to ensure that their values were greater than – 10 kcal/mol to prevent hairpin and primer dimer formation (William & Ronald 1982). The criterion for selecting a primer pair was the ability to generate the greatest number of OTUs identified to species level for selected bacterial groups important in the activated sludge process (Table 9).

Table 7: Primers selected for the amplification of 16S rRNA gene fragments, which were compared for their ability to resolve taxa among the ammonia oxidising bacteria, nitrite oxidising bacteria and mycolata. All sequences are presented in 5' to 3' orientation

16S hypervariable region	**Amplicon length in base pairs**	**Sequence (forward)**	**Reference (forward)**	**Sequence (reverse)**	**Reference (reverse)**
V3	192	CCTAYGGGRBGCASCAG	341f (Behrendt et al. 2012)	TTACCGCGGCKGCTGNCAC	515r is the reverse complement of 515f
V4	291	GTGNCAGCMGCCGCGGTAA	515f (Quince et al. 2011)	GGACTACNNGGGTATCTAAT	806r (Behrendt et al. 2012)
V5	139	ATTAGATACCCNGGTAG	787f (Roesch et al. 2007)	CCGYCAATTYMTTTRAGTTT	926r (Quince et al. 2011)

5.3 DNA extraction comparison

DNA extraction should lyse all cells with equal efficacy such that no bias is introduced. The FastDNA SPIN kit for soil (MPBiomedicals, Santa Ana USA) has been shown to produce the purest and least biased DNA extraction from activated sludge (Guo & Zhang 2013) and is widely used (Guo & Zhang 2012; Ye et al. 2012), although key steps of the protocol can be performed differently. These include the vigour and duration of the bead-beating step (Karst et al 2013), the size of the tube when mixing the lysate with DNA binding matrix, and the timing and temperature of the DNA elution step (see section 13.1.2). The sample was inverted several times to mix and 250µl added to the lysing matrix tube. The standard protocol for this kit within the department was based on internal optimisation of the manufacturers' protocol and specified bead beating at 6.5 ms-1 for 30 seconds. Karst et al (2013) extracted DNA using multiple bead beating settings and compared the subsequent 16S rRNA sequencing data against gold standard fluorescent in situ hybridisation (FISH) for the same sample set. The best correlation between FISH and DNA sequencing was observed with bead beating at 6 ms-1 for 160 seconds. The manufacturer's protocol suggests that incubating the DNA binding matrix for five minutes at 55 degrees will increase DNA yield, yet Karst *et al* did not test this approach. I therefore included the incubation step in this study.

DNA was extracted and sequenced from three samples (see notes column in Table 4) using the three changes to the standard protocol (Table 8). Subsequent DNA extracts were sequenced using barcoded V4 primers (Table 7) on the PGM using 200bp chemistry and OneTouch v1 DL 200bp. The sequence data was processed as per section 5.5 and the taxonomy summary data manually investigated for species level identification of relevant functional groups (see Table 9).

Table 8: DNA extraction using the MPBio FASTSpin kit for Soil. Major variation from the manufacturer's protocol shown.

Protocol	**Bead beating intensity and duration**	**Volume of DNA binding tube**	**Final incubation**
Departmental	6.5 ms^{-1} / 30 s	2 ml	None
Karst et al	6 ms^{-1} / 160 s	15 ml	None
Karst et al + 55	6 ms^{-1} / 160 s	15 ml	5 minutes, 55 °C

5.4 PCR, clean up and sequencing protocol

16S rRNA gene V4-V5 regions were PCR amplified from DNA extracts using Fast Start High Fidelity polymerase (Roche, Basel Switzerland) and buffer containing 18 $MgCl_2$ with Roche PCR Nucleotide Mix. The following PCR protocol was used: 95 °C for 2 minutes followed by 30 cycles of 95 °C for 30 s, 56 °C for 30 s and 72 °C for 45 s and a final elongation of 7 minutes at 72°C. The PCR products where visualised using a 1.5% agarose gel with Nancy-520 DNA stain (SigmaAldrich, Bumbleby, Ukraine) and included a 50 – 2000 bp PCR marker (SigmaAldrich). The PCR products were cleaned with Ampure XP (Beckman Coulter, Brea USA) according to the manufacturers' protocol, except at a ratio of 1.1. Cleaned PCR products were then quantified with the Qubit dsDNA High Sensitivity dye-based DNA quantification kit (Life Technologies, Carlsbad USA) on a Qubit 2.0 Fluorimeter (Life Technologies, Carlsbad USA). The minimum PCR product concentration required for the subsequent step is 100 picomolar. No PCR products were of insufficient concentration. The PCR products from different samples were pooled in equimolar amounts according to their individual DNA concentrations and then re-quantified.

On the day of sequencing, the amplicon pool was diluted to 100 pM and added to the OneTouch v2 400 bp amplification library as per the manufacturer's protocol (Life Technologies 2013b). The library was then enriched with the OneTouch ES according to the manufacturers' protocol (Life Technologies 2013b). The non-enriched library was checked using the Ion Torrent Quality Control Kit according to the manufacturers' protocol (Life Technologies 2013b). The enriched library was sequenced on the Ion Torrent Personal Genome Machine according to the manufacturers' protocol (Life

Technologies 2013a).

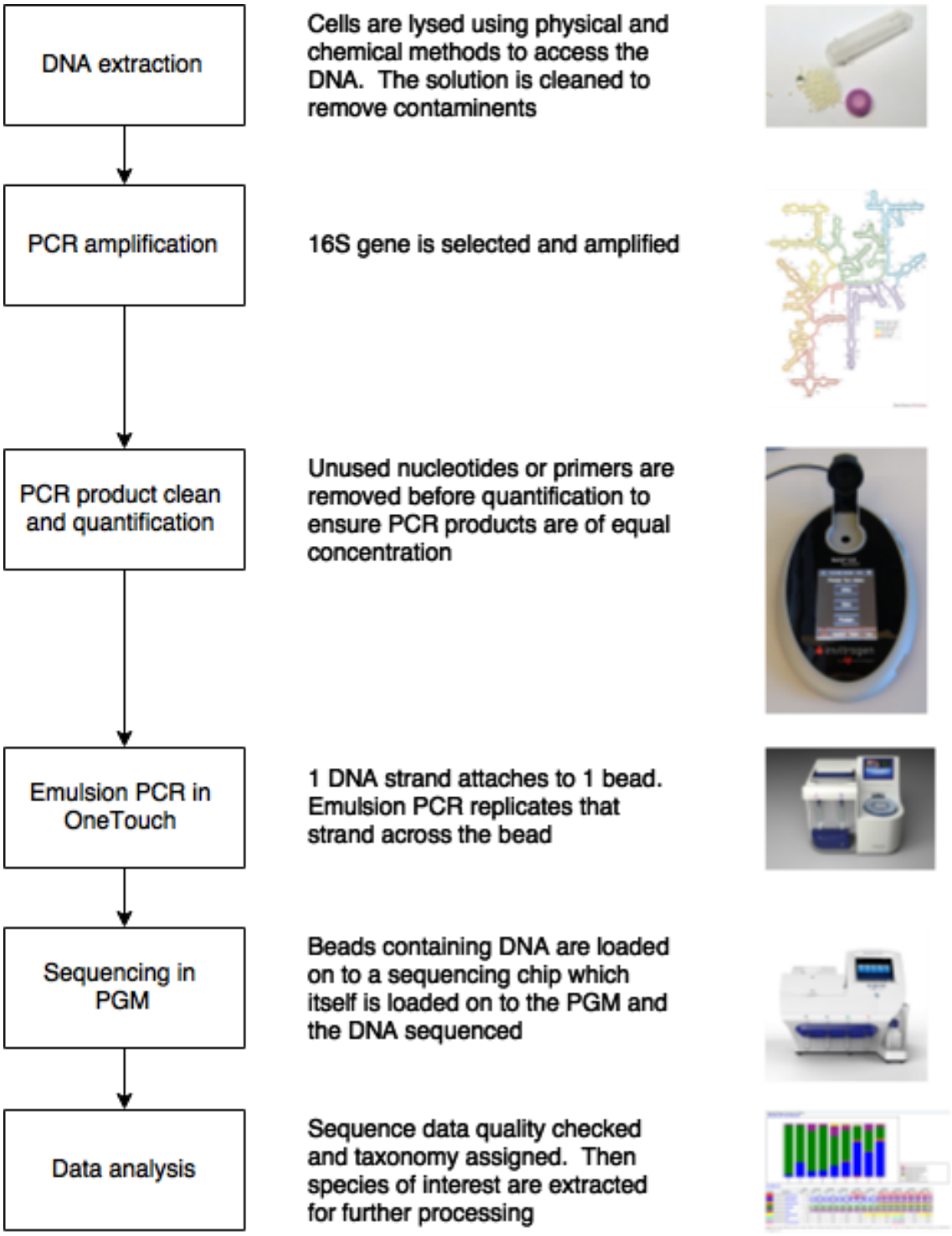


Figure 8: Flow diagram showing sample processing sequence

See 5.1 for an overview of sample processing.

5.5 Bioinformatics and selection of a taxonomic database (DB)

Raw data was processed by Torrent Suite software v4.4.2. This software converts the PGM signal data to DNA sequence data and filters the sequence data by trimming from the 3' end of the amplicons using a moving window in which the minimum average

base quality score is Q15 (Life Technologies 2015 page 22). The sequence data is exported as a FASTQ file containing the sequence data and corresponding base quality scores.

The Quantitative Insights Into Microbial Ecology pipeline v1.7 (J Gregory Caporaso et al. 2010) was used to analyse the data. The data was further filtered by removing sequences less than 100 bp or with a mean quality score less than 20, then demultiplexed once barcodes had been checked for errors. OTUs were picked using the open reference method in which OTUs are clustered at 97% similarity by the UCLUST package (Edgar 2010) against the Greengenes 13_08 database (DeSantis et al. 2006; McDonald et al. 2012; Werner et al. 2012) and SILVA 119 database (Pruesse et al. 2007). Any reads that did not match the database were clustered de novo. Representative sequences for OTUs were picked by selecting the first read in the OTU cluster, as this is the OTU seed used by UCLUST. Taxonomy is assigned using the RDP classifier (Wang et al. 2007) using both Greengenes 13_8 (DeSantis et al. 2006; McDonald et al. 2012; Werner et al. 2012) and Silva 119 (Pruesse et al. 2007) in order to gain the best possible identification of OTUs present. OTU tables were built in the biom format where one table features the Greengenes taxonomy and the other the SILVA taxonomy. The representative sequences were aligned using PyNAST (J. Gregory Caporaso et al. 2010) and the alignment filtered for gaps present in all sequences and for non-conserved sequence data according to the Lane mask (Lane 1991). Sequences that failed to align were removed from the OTU table. PCR chimeras were identified with the ChimeraSlayer package (Haas et al. 2011) and removed from the OTU table and the alignment. A phylogenetic tree was built from the chimera checked OTU table with FastTree (Price et al. 2010). The taxa in the OTU table were summarised by taxonomic level and the abundance of taxa matching those in Table 9 extracted by loading the OTU table in to a spreadsheet and manually copying and pasting relative abundances of taxa of interest. The bacteria of interest with respect to the MST case study are listed in Table 10. The taxa to be used as faecal host markers were based upon a literature review at the time of designing the study. Whilst a number of papers described PCR primers purportedly specific to hosts, very few identified the taxonomy of the target OTUs. Therefore the taxa included in this study were those whose

taxonomy as well as host specificity had been identified. See Table 10 for references for each marker

Table 9: Bacterial functional groups of interest in the activated sludge process and their constituent species **(Seviour et al. 2010; Lücker et al. 2014; Martins et al. 2004)**.

Functional group	**Genera in group**
Ammonia oxidising bacteria / archaea	*Nitrosomonas*
	Nitrosospira
	Nitrosovibrio
	Nitrosococcus
	Nitrosolobus
	Nitrosopumilus
	Nitrosoarchaeum
	Cenarchaeum
	Nitrosotenuis
	Nitrososphaera
Nitrite oxidising bacteria	*Nitrospira*
	Nitrotoga
	Nitrobacter
	Nitrospina
	Nitrococcus
Foaming bacteria	*Caldilinea*
	Microthrix
	Dietzia
	Gordonia
	Skermania
	Corynebacterium
	Mycobacterium
	Nocardia
	Rhodococcus
	Tsukamurella
	Millisia
	Williamsia
	Segniliparus
Bulking bacteria	*Microthrix*
	Tetrasphaera
	Haliscomenobacter
	Trichococcus
	Nostocoida
	Meganema
	Defluviicoccus
	Sphaerotilus
	Thiothrix

.

Table 10: Host specific and faecal indicator marker organisms used in the microbial source tracking case study. References: 1: (Gómez-Doñate et al. 2012), 2: (Wéry et al. 2010), 3: (Lee & Lee 2010), 4: (Johnston et al. 2013), 5: (Weidhaas et al. 2010), 6: (Marti et al. 2011), 7: (Reischer et al. 2006)

Marker	Kingdom	Phylum	Class	Order	Family	Genus	Species	Reference
Human 2	*Bacteria*	*Actinobacteria*	*Actinobacteria*	*Bifidobacteriales*	*Bifidobacteriaceae*	*Bifidobacterium*	*adolescentis*	1
Human 3	*Bacteria*	*Actinobacteria*	*Actinobacteria*	*Bifidobacteriales*	*Bifidobacteriaceae*	*Bifidobacterium*	*dentium*	1
Human 4	*Bacteria*	*Bacteroidetes*	*Bacteroidia*	*Bacteroidales*	*Bacteroidaceae*	*Bacteroides*	*caccae*	2
Human 5	*Bacteria*	*Bacteroidetes*	*Bacteroidia*	*Bacteroidales*	*Bacteroidaceae*	*Bacteroides*	*fragilis*	3
Human 1	*Archaea*	*Euryarchaeota*	*Methanobacteria*	*Methanobacteriales*	*Methanobacteriaceae*	*Methanobrevibacter*	*smithii*	4
Chicken 2	*Bacteria*	*Actinobacteria*	*Actinomycetales*	*Micrococcineae*	*Brevibacteriaceae*	*Brevibacterium*	*avium*	5
Chicken 1	*Bacteria*	*Actinobacteria*	*Actinobacteria*	*Bifidobacteriales*	*Bifidobacteriaceae*	*Bifidobacterium*	*saeculare*	1
Cow 1	*Bacteria*	*Actinobacteria*	*Actinobacteria*	*Bifidobacteriales*	*Bifidobacteriaceae*	*Bifidobacterium*	*pseudolongum*	1
Pig 1	*Bacteria*	*Actinobacteria*	*Actinobacteria*	*Bifidobacteriales*	*Bifidobacteriaceae*	*Bifidobacterium*	*PG10*	1
Pig 2	*Bacteria*	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Lactobacillaceae*	*Lactobacillus*	*amylovorus*	6
Gull 1	*Bacteria*	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Enterococcaceae*	*Catellicoccus*	*marimammalium*	3
Ruminant 1	*Bacteria*	*Bacteroidetes*	*Bacteroidia*	*Bacteroidales*	*Bacteroidaceae*	*5-7N15*		7
E. coli	*Bacteria*	*Proteobacteria*	*Gamma-proteobacteria*	*Enterobacteriales*	*Enterobacteriaceae*	*Escherichia*	*coli*	
Enterococci	*Bacteria*	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Enterococcaceae*	*Enterococcus*		

5.6 Statistical analyses

5.6.1 UniFrac phylogenetic distance

The unweighted UniFrac metric gives an indication of the dissimilarity between microbial communities, taking account of both the phylogenetic distance between communities and the presence or absence of taxa (Lozupone & Knight 2005). In calculating the UniFrac distance, a phylogenetic tree is built from the 16S rRNA sequence data corresponding to taxa found in one or both samples. Branches are categorised as shared where they lead to taxa observed in both samples, and unshared where they do not. The length of each branch is proportional to the evolutionary distance between sequences and thereby taxa, since the tree is built using the maximum likelihood method (Price et al. 2009). The distance between samples is calculated as the sum of unshared branch length divided by the sum of all branch lengths. The weighted UniFrac metric also takes account of the relative abundance of taxa present. Since bacteria with similar optimum growth conditions are likely to be closely related, the UniFrac metric can indicate the difference in environmental conditions experienced by two communities in addition the difference in taxa present (Hamady & Knight 2009).

5.6.2 Beta significance tests

The beta significance test investigates the statistical significance of pairwise UniFrac distances by applying Monte Carlo randomisation in which sequences are randomly assigned to a sample 100 times. The probability that the distance between samples is greater than it would be between the random samples is calculated and gives an indication of whether there is an underlying phenomenon driving community divergence.

5.6.3 Accurate estimation of AOB abundance

In order to establish how many reads were required to accurately estimate AOB abundance, a sample from the nitrifying plant Tudhoe Mill (site code A2) was sequenced without sample multiplexing in order to acquire the maximum number of reads possible from the sequencing platform. The data was processed as described in section 5.5. The subsequent OTU table was randomly rarefied with replacement over 10 iterations for read depths ranging from 10,000 to 1.1m reads in steps of 10,000. The abundance of AOB in each iteration at each read depth was extracted by manual comparison of rarefied OTU table taxonomy against the taxa specified in Table 9.

5.6.4 Comparison of AOB abundance as measured by qPCR, FISH and PGM

Baptista et al (2014) compared the abundance of AOB detected by a qPCR assay targeting the amoA gene and AOB detected using FISH probes. The samples used in their study were sequenced on the PGM and the abundance of AOB extracted as described earlier. As the PGM data results in percent relative abundance for each bacteria detected, this was converted to an absolute abundance by undertaking total cell counts (see section 13.1.4) for each sample and multiplying the percent relative abundance by the total cell count.

5.6.5 Calculation of cell specific ammonia oxidation rate (CSAOR)

In an attempt to quantify ASP plant nitrification stability several authors have calculated the CSAOR using quantitative molecular techniques such as FISH. However NGS data are relative abundances as opposed to absolute cell counts. Therefore in order to use NGS data without additional time consuming total cell count analysis, the following equation was used in this study adapted from Daims et al (2001):

$$CSAOR = \frac{(A \times M \times 10^6) \times r}{AOB \times MLSS \times V}$$

Equation 5: Cell specific ammonia oxidation rate

Where A = grams of ammonia, M = the number of moles of ammonia in a gram, r = correction factor of 0.9 due to some ammonia removal by adsorption and assimilation, AOB = the total relative abundance of ammonia oxidising bacteria, MLSS = mixed liquor suspended solids in mg/L and V = the volume of the aeration basin in litres. The section in parentheses includes a multiplier of 10^6 such that the units become micromoles. The ratio of micromoles of ammonia to milligrams of AOB biomass is broadly equivalent to the ratio of femtomoles of ammonia to single AOB cells in Daims et al (2001), since a single bacterial cell is thought to weigh approximately one picogram. It is recognised that this interpretation is tenuous, however keeping the ratios the same may enable comparison of results between methods.

5.6.6 Correlation statistics

Regression analysis of Enterococci and *E. coli* abundance measured by PGM and culture data (from the EA and Northumbrian Water Ltd) was used to identify any correlation between FIB enumeration using the two methods. The data was checked for normality using Minitab 17.

As faecal host marker abundance does not correlate with faecal indicator organism abundance (Stoeckel & Harwood 2007; Lu et al. 2008; Lee et al. 2013), the co-incidence in the detection of faecal indicators with faecal host markers were analysed by correlation analysis. A binary matrix was constructed showing detection of faecal contamination by the EA (*E. coli* or Enterococci abundance >10 CFUs / 100 ml) or PGM (presence of one or more faecal markers in Table 10). Binary logistic regression was used to identify any correlation between occurrence of faecal host markers and faecal indicator organisms.

To investigate the sensitivity and specificity of the faecal markers in Table 10, the abundance of each marker was noted in each collected faecal sample (Table 6). Host marker sensitivity was calculated by dividing the number of samples from the host in which the marker was detected by the total number of samples from that host. Host marker specificity was calculated by dividing the number of samples in which the host marker was not detected by the number of non-host samples tested.

5.6.7 Source Tracker

The SourceTracker tool (Knights et al. 2011) was used to identify the proportion of the bacteria in bathing water samples arising from each point source as well as the agricultural inputs (Newton et al. 2013). Briefly, SourceTracker considers the "sink" community (the bathing water in this case) as consisting of bacteria from a number of "source" samples which may be known (the CSO, sewage treatment works and faecal samples in this case study) or unknown (Table 6). The tool then performs Bayesian iterative Gibbs sampling of the joint Dirichlet distributions between sink and sources in order to estimate the probability of each sequence in the sink coming from each or none of the sources. The results produce an estimated proportion of bacteria in the sink from each source as well as from an unknown source, with concordant probability estimates.

Partial least squares analysis collapses variables in to independent dimensions which are weighted according to their influence on the Y variable (Geladi & Kowalski 1986) in a manner similar to principal component analysis. Therefore the partial least squares approach can take account of the fact that whilst e.g. 20% of all bacteria arise from source A and 2% from source B, source B features greater faecal contamination and thereby contributes a disproportionate amount of faecal contamination to the sink.

This method was used to identify correlation between the relative contribution from faecal sources in the bathing water and the abundance of FIB observed by the EA.

5.7 Primer selection

5.7.1 How does primer pair selection affect observed microbial community structure?

Three PCR primer pairs (Table 7) were used to selectively amplify 16S ribosomal RNA gene fragments from the *Bacteria* and *Archaea* in three ASP samples (Table 4) producing three sets of three amplicon pools. Sequencing and taxonomic classification of each amplicon pool produced three summarised taxa abundance tables from which the abundance of functionally relevant taxa (Table 9) was extracted. The total microbial communities observed in each amplicon pool are visualised in Figure 11 where each colour represents an equivalent taxon in each sample. The figure provides a qualitative illustration of the taxonomic bias introduced by primer pair selection

Each pair of amplicon pools was compared using weighted and unweighted UniFrac (see 5.6.1) and the results stored in two distance matrices for each primer set. Principal coordinate analysis (PCoA) of each UniFrac matrix enabled the representation of the UniFrac dissimilarity of each sample within individual primer sets in 2 dimensional space. The PCoA matrices from all primer sets were further transformed using Procrustes analysis (Gower 1975) relative to a reference principal component vector set (unweighted V3) such that all principal component analysis output could be visualised on one graph. Figure 9 shows this analysis in the form of a principal component plot where each amplicon pool is plotted using both weighted and unweighted UniFrac. The red lines connect transformed data to the “parent” unweighted V3 data. The figure demonstrates that whilst primer selection and the UniFrac metric used does impact the observed beta diversity, the observed communities cluster according to sample. This suggests that the diversity and relative abundance of taxa in a sample has a greater impact on beta diversity than primer selection or choice of metric. The data also suggests that each hypervariable region investigated (V3, V4, V4 and V4V5) contains sufficient sequence variance to distinguish ASP samples at the 97% similarity OTU cut off.

Figure 10 shows the results of the weighted UniFrac beta diversity metric for each primer pair individually, showing that the relative distance between each sample varies

dependant on primer pair. Therefore whilst the difference in community composition between samples has the greatest impact on beta diversity, the primer selected also has an impact. The varying relative distance of samples may reflect the difference in sequence divergence of the 16S rRNA hypervariable regions in closely related taxa. Therefore it is important to establish whether taxa of interest can be distinguished by each hypervariable region and therefore which primer should be used in this case study.

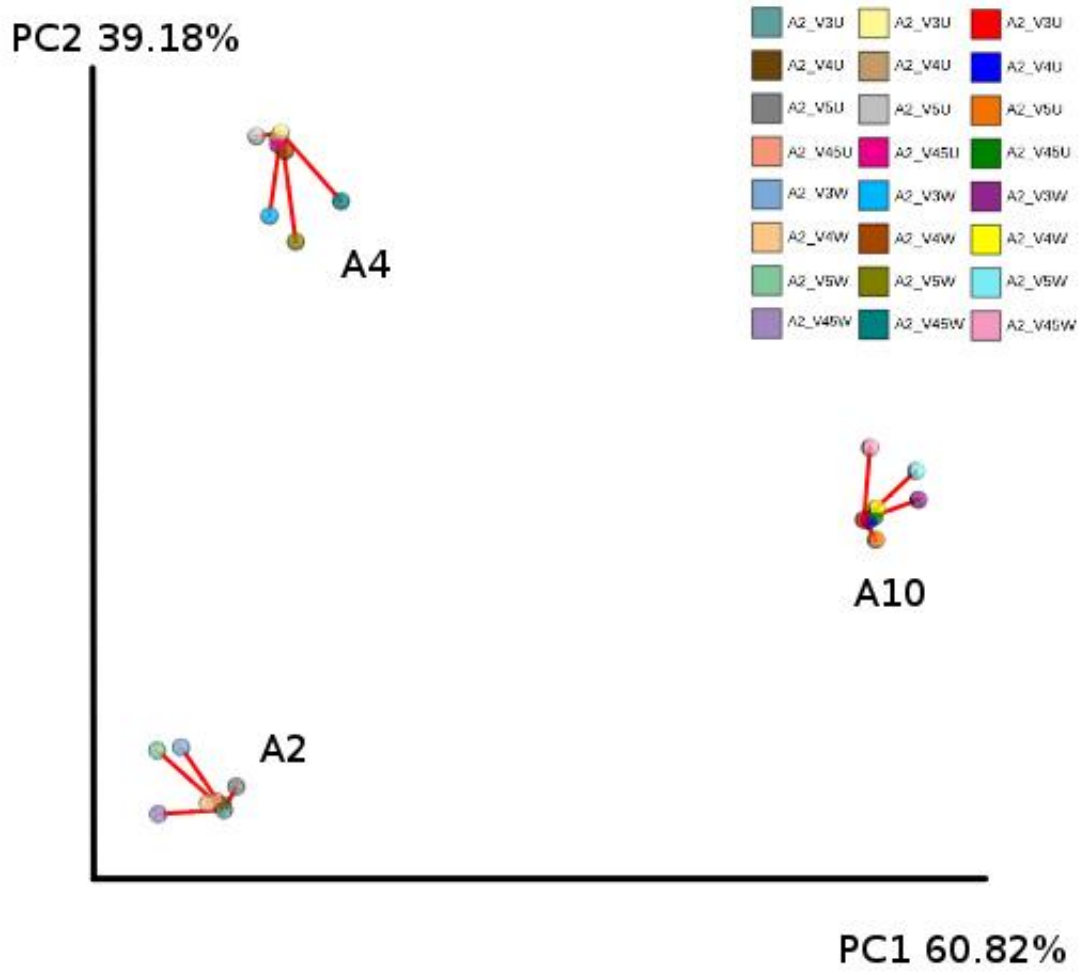


Figure 9: Two dimensional Procrustes analyses plot showing the weighted and unweighted UniFrac beta diversity distance between three activated sludge samples amplified by polymerase chain reaction. Sample A2 is from Tudhoe Mill activated sludge plant (ASP) which treats domestic sewage. Sample A4 is from Aycliffe ASP which treats heavy industrial and domestic sewage. Sample A10 is from Browney ASP which treats light industrial and domestic sewage. Each sample was PCR amplified using primers targeting the V3, V4, V5 and V4V5 hypervariable regions of the 16S rRNA gene.

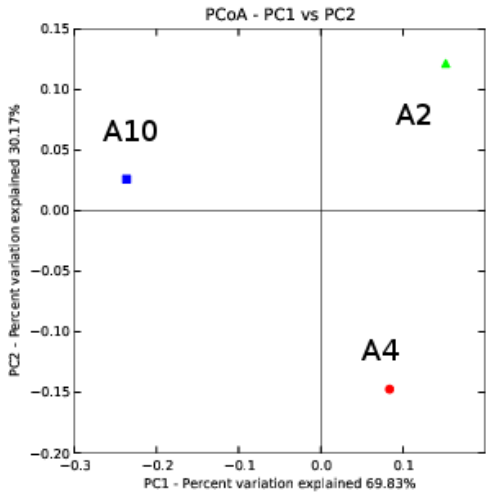


Figure a

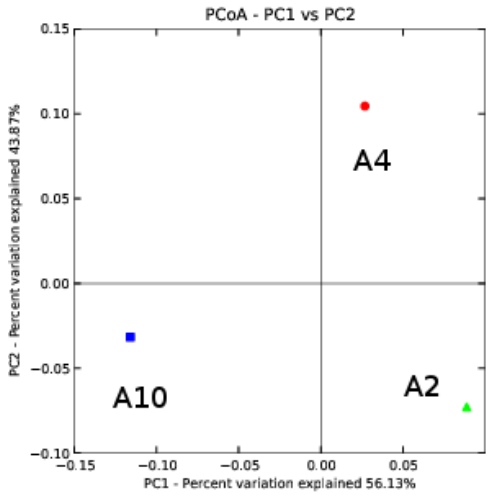


Figure b

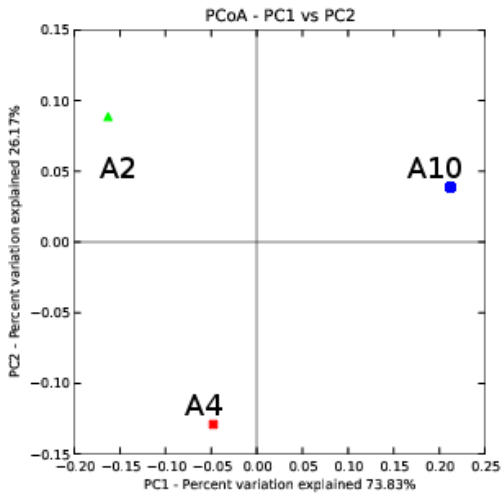


Figure c

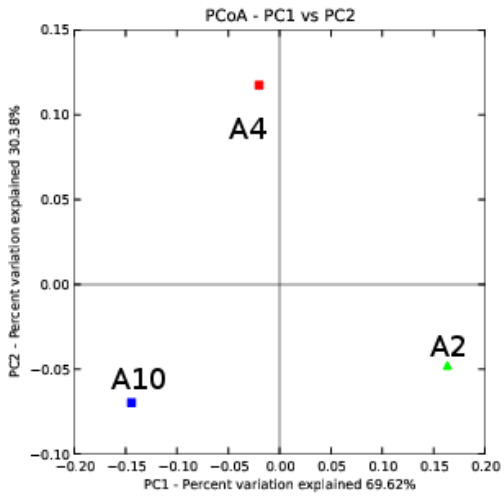


Figure d

Figure 10: Weighted UniFrac beta diversity principal coordinate analysis plots showing samples from Tudhoe Mill (A2) nitrifying activated sludge plant (ASP) treating domestic sewage, Aycliffe (A4) ASP treating heavy industrial and domestic sewage and Browney (A10) ASP treating light industrial and domestic. Figure a shows the microbial community structure identified using amplicons from the V3 region, figure b from the V4 region, figure C from the V5 region and figure d from the V4V5 tandem regions.

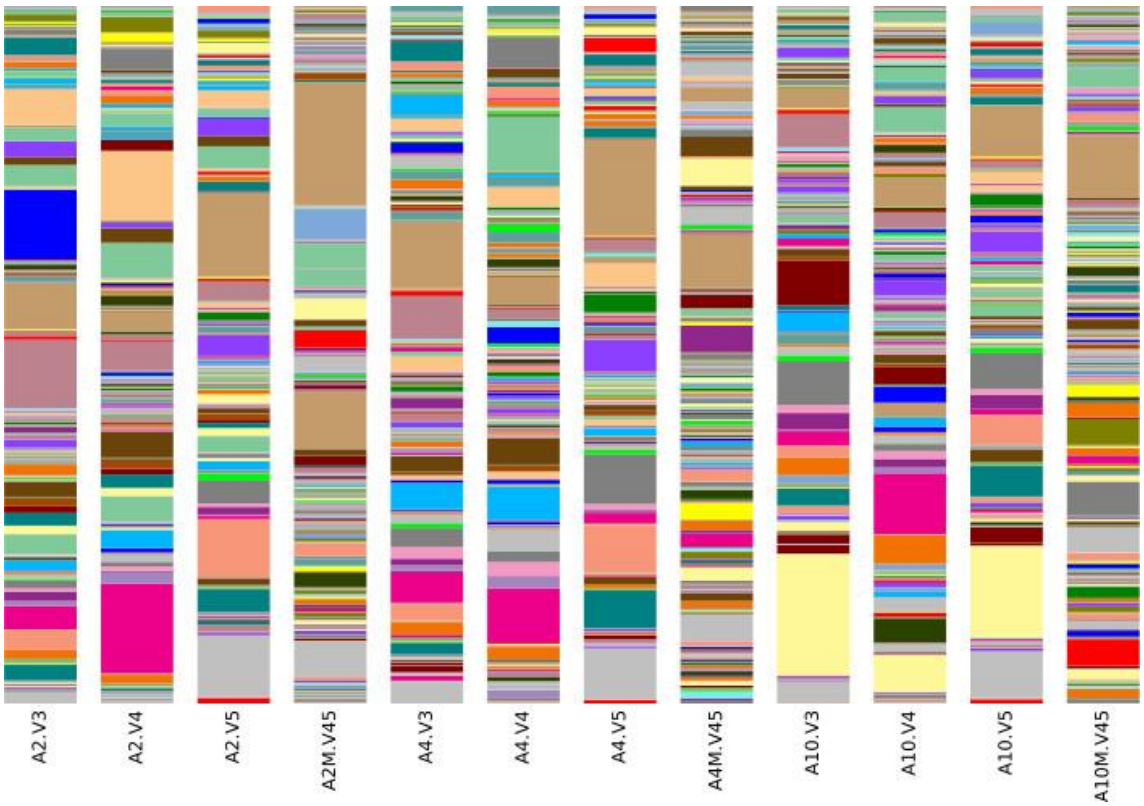


Figure 11: Taxa summary of amplicon libraries amplified from microbial communities using primers targeting the variable V3, V4 and V5 regions of the 16S rRNA gene of Bacteria and Archaea. Each sample label, A2, A4 and A5 (representing Tudhoe Mill nitrifying plant no industrial influent, Aycliffe nitrifying plant with a large proportion of industrial influent and Browney nitrifying plant with a small proportion industrial influent) has a post-fix label representing the respective variable region. Each colour represents OTUs at 97% similarity cut off grouped according to taxa identity, where colours represent equivalent taxa in each sample.

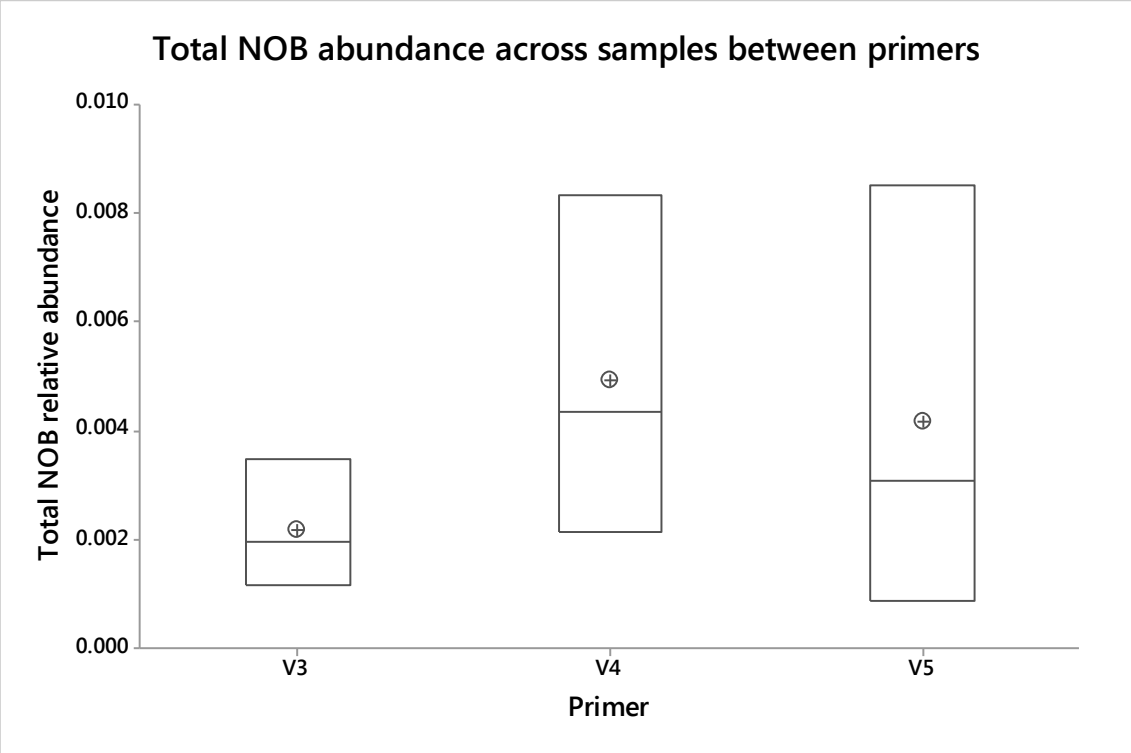
5.7.2 Comparison of the abundance of specific taxa-groups detected by each primer

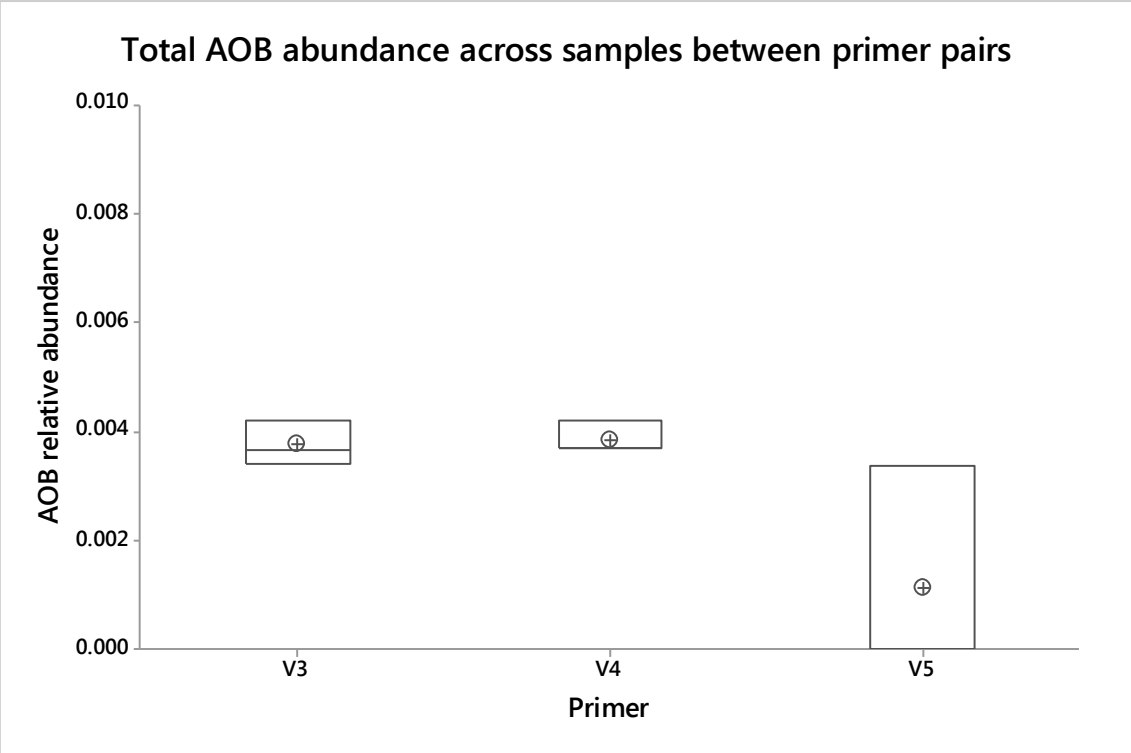
The stability of important activated sludge processes is directly linked to the abundance of the taxa that perform these processes (Table 9). Therefore in order to monitor process stability one must be able to distinguish these taxa within the microbial community. The ability to differentiate between taxa relies on the extent of sequence divergence in the 16S rRNA hypervariable regions between related taxa. Since the sequence read length for NGS is currently limited to about 400 base-pairs, i.e. less than the entire 16S rRNA gene length, we must target a suitable hypervariable region. Each hypervariable region differs in number of base pairs and the extent of sequence divergence within it. Therefore it must be established whether sequence divergence within each hypervariable region is sufficient to distinguish taxa of interest.

As described in the previous section, PCR amplification of three samples using three primers produced nine amplicon pools which were each sequenced. Clustering of the sequence data at the 97% cut off level created nine OTU tables. Taxonomic assignment of the OTUs in each table enabled the identification and extraction of functionally relevant activated sludge OTU abundances.

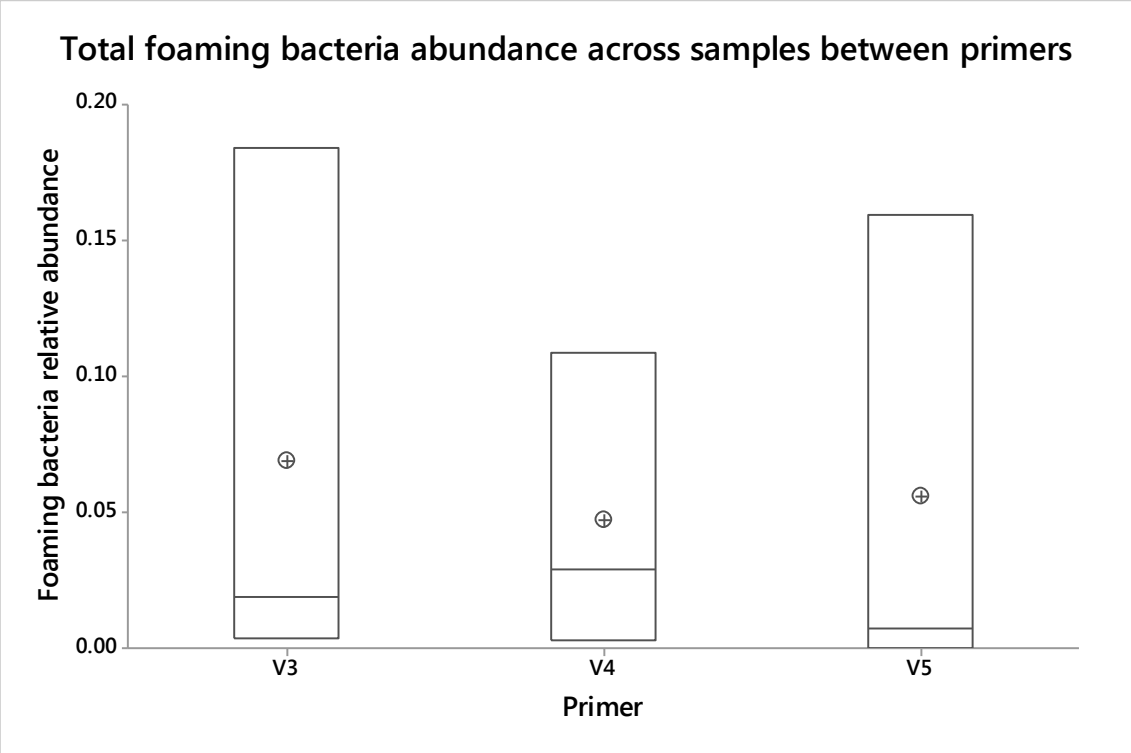
There was no significant difference in the abundance of NOB, foamers or bulkers between primers (ANOVA p=0.545, p=0.95, p=0.832 respectively), however the coefficient of variation (CV) for foamers and bulkers was lowest using V4 primers, suggesting V4 may produce more precise taxa abundance measurements overall (NOB V3 = 54.36%, V4 = 63.74%, V5=94.8%; foamers V3 = 145.25%, V4 = 116.76%, V5 = 161.75%; bulkers V3 = 131.49%, V4 = 94.64%, V5 = 141.97%). These data are represented in Figure 12. Comparison of AOB abundance at genus level was significantly different across samples between primer pairs (ANOVA, p=0.045) however Tukey's pairwise comparison showed no significant difference between primer pairs (V3-V4 p=0.994, V3-V5 p=0.07, V4-V5 p=0.061). This may be due to V5 primers failing to detect any AOB in sample A2 and A10. The coefficient of variation was lowest using V4 primers (V3 = 10.85%, V4 = 7.47%, V5 = 173.21%). Therefore the V4 region appears to be most suitable for measuring AOB abundance. The abundance of AOB, NOB, foamers and bulkers is plotted in Figure 12.

In general, genus level identification is sufficient to establish whether a taxon is relevant to activated sludge processes. For example, all *Nitrosomonas sp.* are ammonia oxidisers. As such the identification of taxa to species level is of less relevance than genus level. However optimum environmental conditions for growth and metabolism vary between species, and therefore this information may be of use in understanding the operational potential of an ASP. For example *Nitrosomonas europaea* is known to degrade halogenated compounds (Vannelli et al. 1990) which may be of relevance when treating industrial effluent.





See next page for figure legend



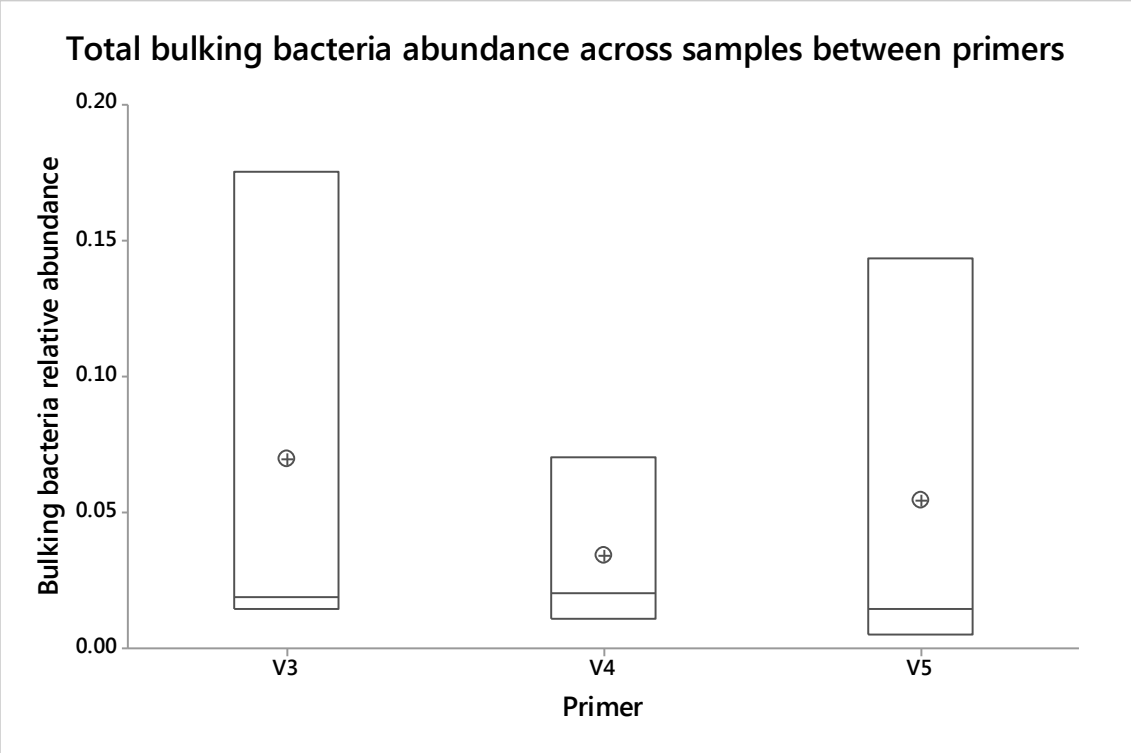


Figure 12: Each boxplot shows the total abundance of indicated bacterial groups in samples from Aycliffe industrial nitrifying activated sludge plant (ASP), Tudhoe Mill nitrifying ASP and Browney nitrifying ASP. One sample from each plant was PCR amplified using V3, V4 and V5 primers producing nine amplicon pools. Each pool was sequenced and the data clustered in to operational taxonomic units (OTUs) at the 97% cut off level and assigned to a taxonomy. Those taxa known to be ammonia oxidisers, nitrite oxidisers, involved in foaming or bulking were identified and their cumulative abundance recorded.

5.7.3 Comparison of the taxonomic resolution observed using each primer

The mean number of AOB OTUs identified to species level was significantly higher using V3 or V4 primers than V5, but not between V3 and V4 (ANOVA $p < 0.00$, Tukey's test V3-V4 $p=0.679$, V3-V5 $P < 0.000$, V4-V5 $p=0.001$, see Figure 13, Figure 14). The coefficient of variation was lowest using V4 (V3 = 14.3%, V4 = 9.1%, V5 = 173.2%).

The mean number of NOB OTUs identified to species level was higher using V3 than V4 or V5 primers, however the difference was only significant between V3 and V5 (ANOVA $p=0.045$, Tukey's test V3-V4 $p=0.098$, V3-V5 $p=0.049$, V4-V5 $p=0.844$, see Figure 13, Figure 14). The coefficient of variation was lowest using V4 (V3 = 28.87%, V4 = 33.33%, V5 = 65.47%).

The difference in mean number of foaming OTUs identified to species level was approaching significance between V3 and V5, significant between V4 and V5 and not significant between V3 and V4 (ANOVA $p=0.047$, Tukey's test V3-V4 $p=0.859$, V3-V5 $p=0.051$, V4-V5 $p=0.1$, see Figure 13, Figure 14). The coefficient of variation was lowest for V4 suggesting this region is best for differentiating foamers (V3 = 40.63%, V4 = 32.54%, V5 = 76.97%).

The mean number of bulking OTUs identified to species level was higher in V4 than V3 or V5 but this difference was not significant (ANOVA $p=0.095$, Tukey's test V3-V4 $p=0.692$, V3-V5 $p=0.254$, V4-V5 $p=0.087$, see Figure 13, Figure 14). The coefficient of variation was highest using V4 suggesting the observed difference in mean abundance may be due to increased variation (V3 = 14.24%, V4 = 32.94%, V5 = 18.37%). Therefore each primer seems equally good at differentiating bulking OTUs.

Overall, the V4 primer appears to give the most precise measurement of abundance and greatest taxonomic resolution, and therefore this primer was selected at the beginning of the case study when 200 bp sequencing chemistry was being used. Morales et al (2009) found taxonomic assignment of V4V5 sequence data produced the same assignment as full length 16S rRNA sequences in 98.5% of cases at phylum level, while Claesson et al (Claesson et al. 2010) found that the V4V5 region produced the most accurate taxa classification at genus level of all tandem variable regions. In 2013 the sequencing platform was upgraded to sequence 400bp amplicons. As I had invested in 50 barcoded V4 primers at a cost of £3500, I investigated whether utilising

the V5 reverse primer would enable greater taxonomic resolution. Therefore the PCR protocol was amended to amplify the V4V5 tandem region.

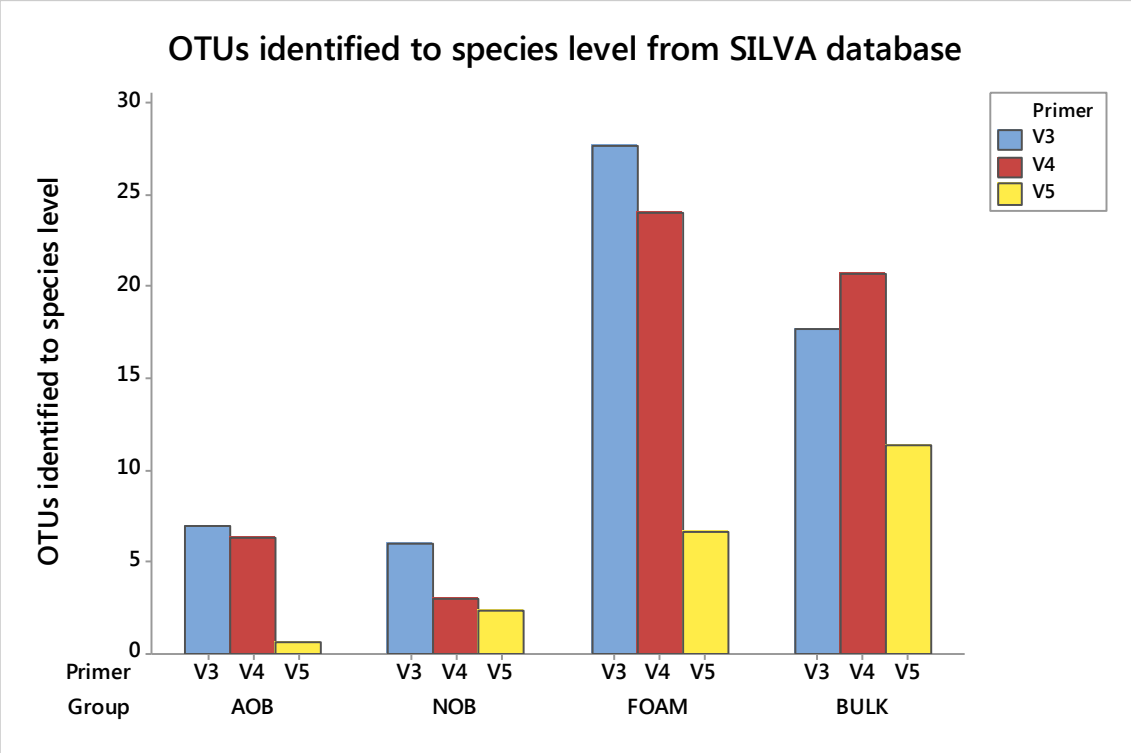
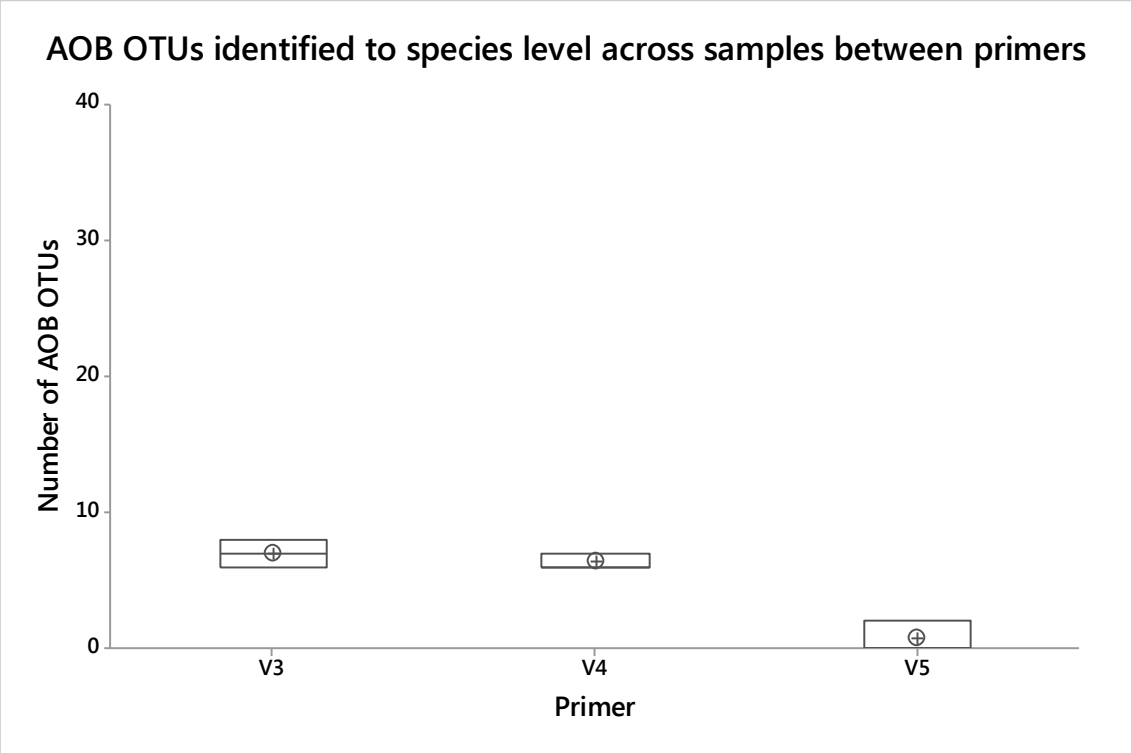
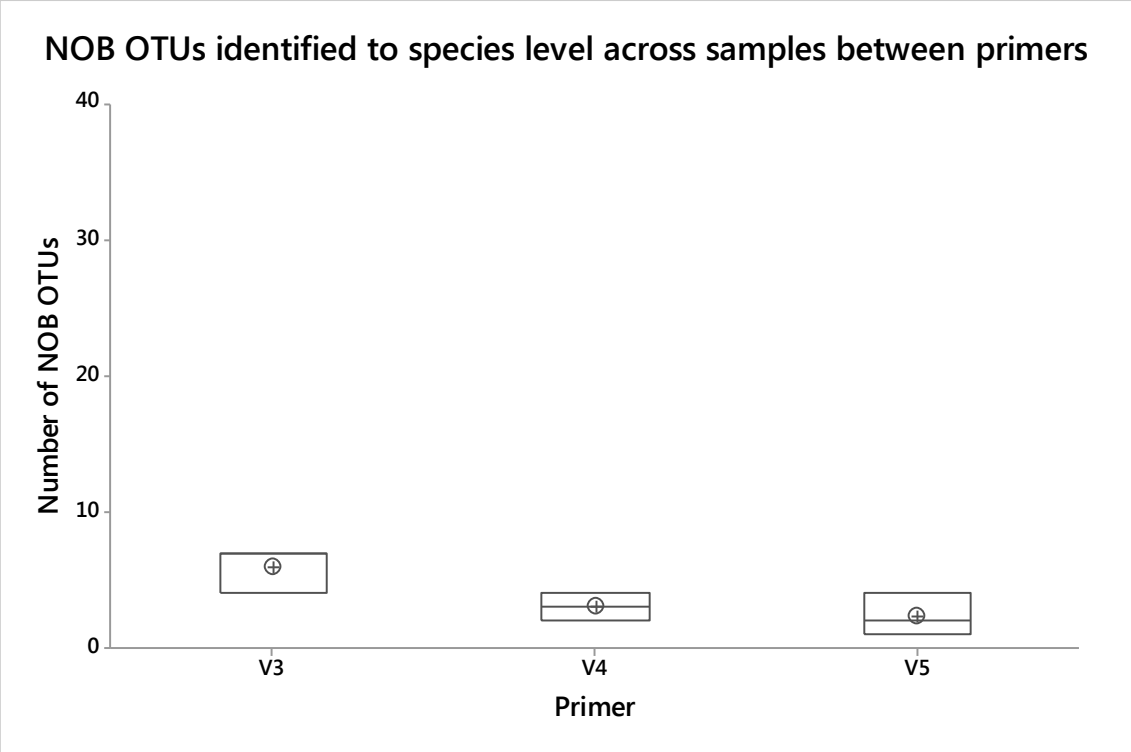
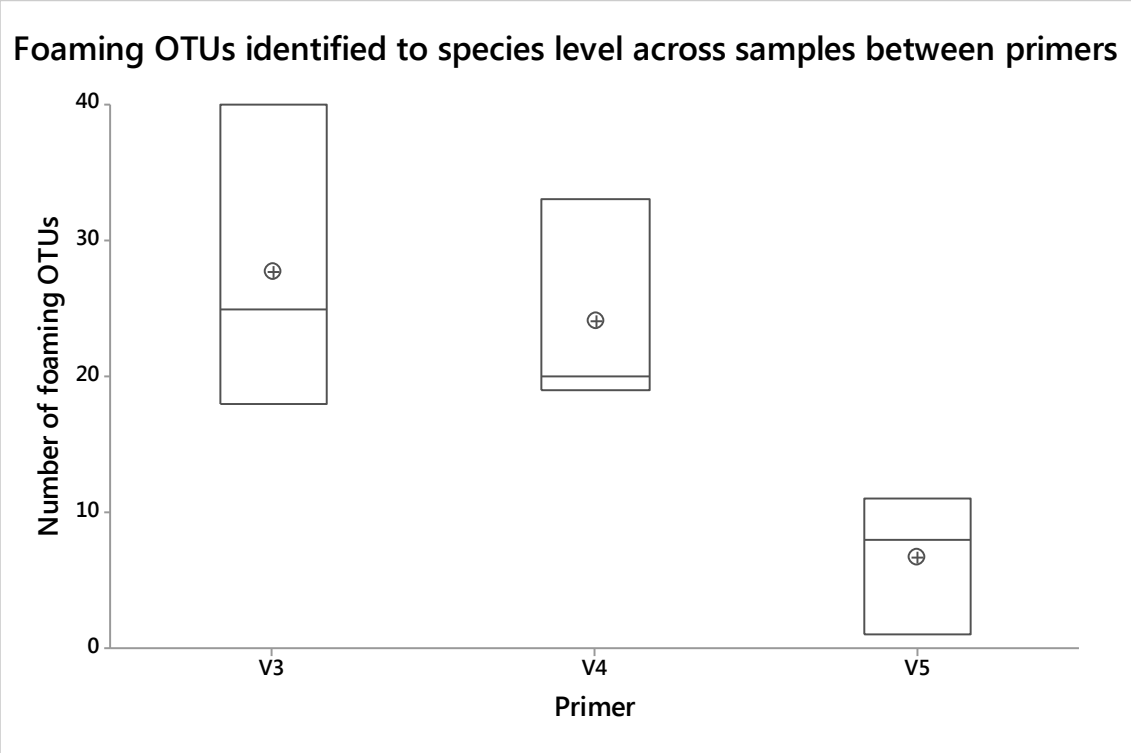


Figure 13: Number of operational taxonomic units at 97% cut-off identified to species level





See next page for figure legend



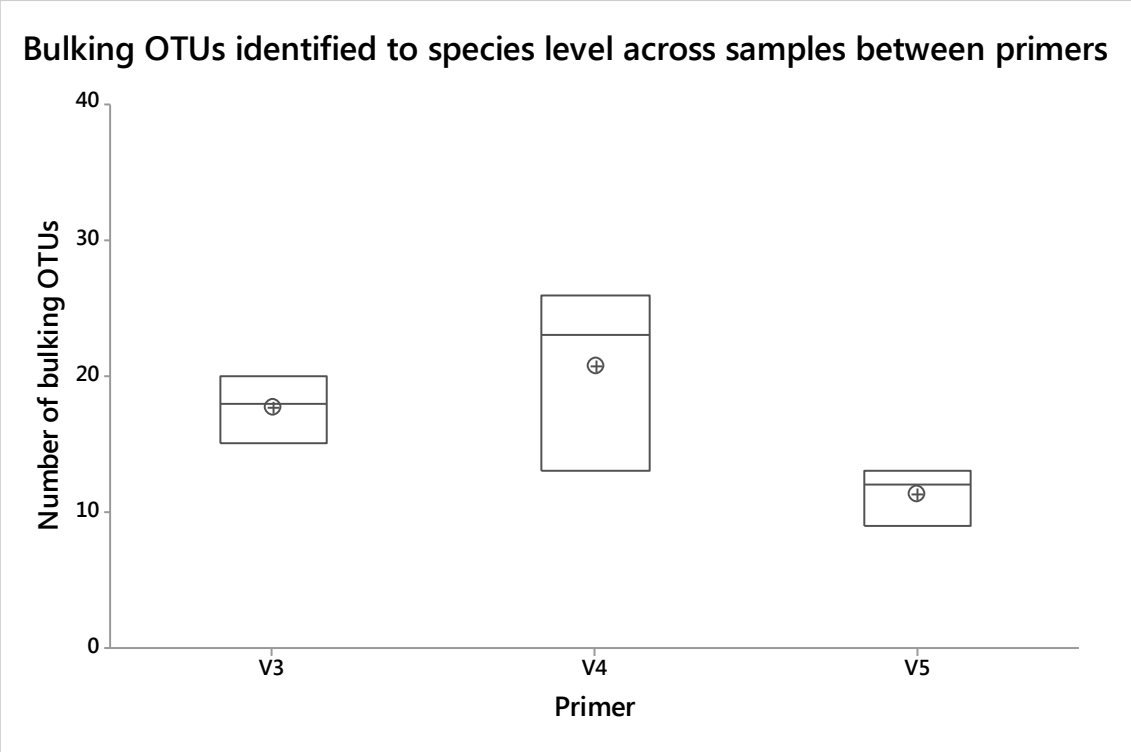


Figure 14: Boxplots e – h show the number of OTUs identified to species level in a sample from Aycliffe industrial nitrifying activated sludge plant, Tudhoe Mill nitrifying activated sludge plant and Browney nitrifying activated sludge plant. One sample from each plant was PCR amplified using V3, V4 and V5 primers and taxonomy assigned to each amplicon sequenced. The number of OTUs to which a species had been assigned was recorded.

5.7.4 Primer selection: Discussion

Given that we cannot sequence the entire 16S rRNA gene due to the limitations of NGS, we must select one or more hypervariable regions to target with specific primer pairs. In general the criterion for making this selection is how well the observed microbial community structure matches the "real" community structure. This implies prior knowledge of the "real" community structure yet when working with environmental samples we cannot know the community structure a priori (Armougom & Raoult 2009). We must therefore base our decision on which primer produces the optimum results for the taxa of interest. The community level results in section 5.7.1 showed that the true difference in community structure between samples has a greater impact on observed community structure than primer choice, which illustrates that the impact of primer choice on observed results is limited and is in line with previous studies (Huse et al. 2008). However in order to use NGS to monitor ASP operation, we must investigate which primer gives the best taxonomic resolution for the functionally relevant taxa listed in Table 9. Therefore in sections 5.7.2 and 5.7.3 the abundance and number of functionally relevant taxa identified to species were extracted from the results for each primer pair. There was no significant difference in the abundance of each functional group across primers, though the V4 data had the lowest coefficient of variation. There have been very few studies which have investigated this phenomenon in activated sludge, however Albertsen et al (2015) found that primer choice had little impact on UniFrac diversity which is in agreement with my findings. Significantly fewer OTUs in V5 data were identified to species level than V3 or V4 data across all functional groups. There was no significant difference between V3 and V4 OTU identification, though the V4 data showed lowest coefficient of variation. There have been no other studies investigating OTU resolution across primers. I suggest that the best primer for differentiating ASP functional groups was V4. However as noted in section 5.7.3 an improvement in NGS read length meant I investigated the possibility of using tandem regions and began using V4V5 which provided even greater taxonomic resolution. These results highlighted the importance of primer selection and demonstrated that our primers were able to differentiate all the functional groups under investigation. This is a vital first step in ensuring NGS is able to monitor ASP function as hypothesised.

5.8 DNA extraction and recovery method selection

Environmental samples contain microbial cells with different cell wall structures, leading to differential cell lysis during DNA extraction such that relative abundances are artificially skewed compared to those in situ. Whilst it is important that target cells are lysed, accurate measurement of relative abundance requires efficient lysis of all cells in a sample. Therefore DNA extraction must be optimised such that target and non-target cells are lysed efficiently. In addition, the fraction of total DNA recovered from the binding matrix is hypothesised to increase when subjected to 55 degree incubation. I therefore investigated the optimum combination of lysing vigour and duration as well as incubation temperature.

Three DNA extraction methods (Table 8) were used to extract DNA from three ASP samples (Table 4) producing three sets of three amplicon pools. Sequencing and taxonomic classification of each amplicon pool set produced three summarised taxa abundance tables from which the abundance of functionally relevant taxa (Table 9) was extracted. The total microbial communities observed in each amplicon pool are visualised in Figure 15 where each colour represents an equivalent taxon in each sample. The figure provides a qualitative illustration of the taxonomic bias introduced by DNA extraction method.

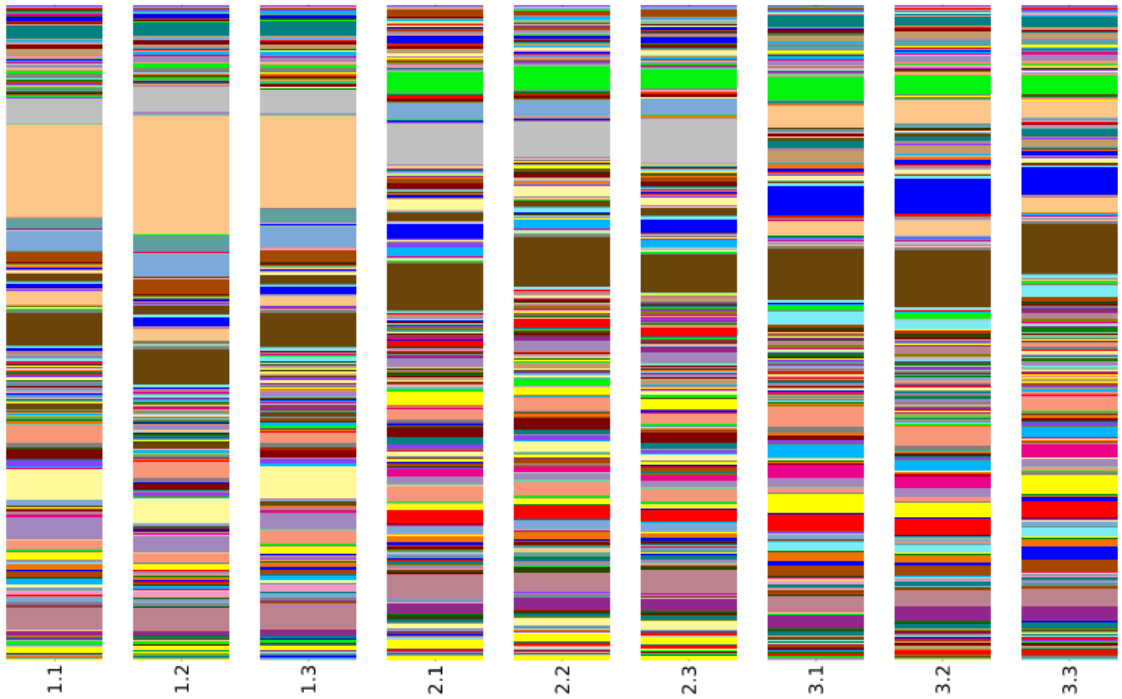


Figure 15: Taxonomic summary of three activated sludge samples extracted using three DNA extraction methods. Columns labelled Z.X where Z = the sample and X = the extraction method. Sample 1 was Tudhoe Mill nitrifying plant treating domestic sewage and prone to foaming. Sample 2 was Amble non-nitrifying sequencing batch reactor treating domestic sewage. Sample 3 was from Aycliffe nitrifying plant treating domestic and light industrial sewage. Method 1 was the manufacturer's protocol. Method 2 increased bead beating duration to 160 seconds. Method 3 included a final incubation of DNA binding matrix at 55 degrees centigrade. Each colour band represents a taxon and the height of the band represents its relative abundance.

5.8.1 Comparison of DNA extraction and recovery methods

DNA extraction and recovery is not completely efficient largely due to difficulty in lysing resistant cell types. Therefore methods must be optimised to yield DNA that accurately represents the original sample. Given that it is difficult to establish the true abundance and diversity of the microbial community in a sample, a useful proxy of extraction and recovery efficiency is the concentration or mass of DNA extracted from a sample. It can be assumed that where a greater mass of DNA is extracted, a greater number and therefore a larger proportion of the total cells in the sample have been lysed. It follows that it is advantageous to extract and recover the maximum possible concentration/mass of DNA from a sample. Comparison of DNA concentrations extracted from different samples using three different methods gave an indication of which was the most efficient.

The results indicate that method 3 extracts significantly more DNA than methods 1 or 2 (ANOVA, Tukey's method, $p = 0.04$, $p = 0.02$ respectively; Figure 16), and thereby may provide a better representation of the original sample community. Methods 1 and 2 yielded similar DNA concentrations (ANOVA, Tukey's method, $p = 0.37$). The concentration of DNA extracted by method 2 varied more than methods 1 and 3; coefficient of variation of 63.0, 20.4, and 13.4 % respectively.

Methods 1 and 2 differ in the vigour and duration of the cell lysis step; whereas method 3 differs from method 2 in that the DNA binding matrix is incubated at 55 degrees as opposed to room temperature. Given that there was no statistically significant difference in the concentration of DNA extracted by methods 1 and 2, it appears that increased vigour of cell lysis does not significantly affect the amount of DNA extracted. However as the DNA extract concentration varied more in method 2, it may be that while a greater fraction of cells in samples were lysed a proportion of this additional DNA remained bound to the binding matrix. This bound DNA would be lost and therefore not be represented in the final data.

It is thought that the bond strength between single stranded DNA molecules and the silica binding matrix is determined by the number of intermolecular hydrogen bonds (Melzak et al. 1996). It should be noted that the proprietary binding matrix is known to contain guanidine isothiocyanate which denatures DNA molecules. Given that guanine and cytosine each have one more hydrogen atom than adenine or thymine, it follows that DNA strands with high GC content will bind more strongly to the matrix. As GC content is associated with taxonomic identity, samples with different community structures may have varying overall GC content and thereby varying DNA recovery. Given that the original samples were from three plants treating different sewage types, and thereby could be expected to harbour different microbial communities, this may explain the greater variation in DNA extract concentration using method 2. Method 3, which is identical to method 2 excepting the final incubation step, yielded twice the DNA with less variability. The incubation step increases enthalpy in the silica-DNA complex thereby leading to a higher probability of DNA disassociation where the DNA molecule features a high GC content. The observed increase in mean DNA extract concentration and decrease in variability may be explained by the hypothesised additional recovery of high GC DNA.

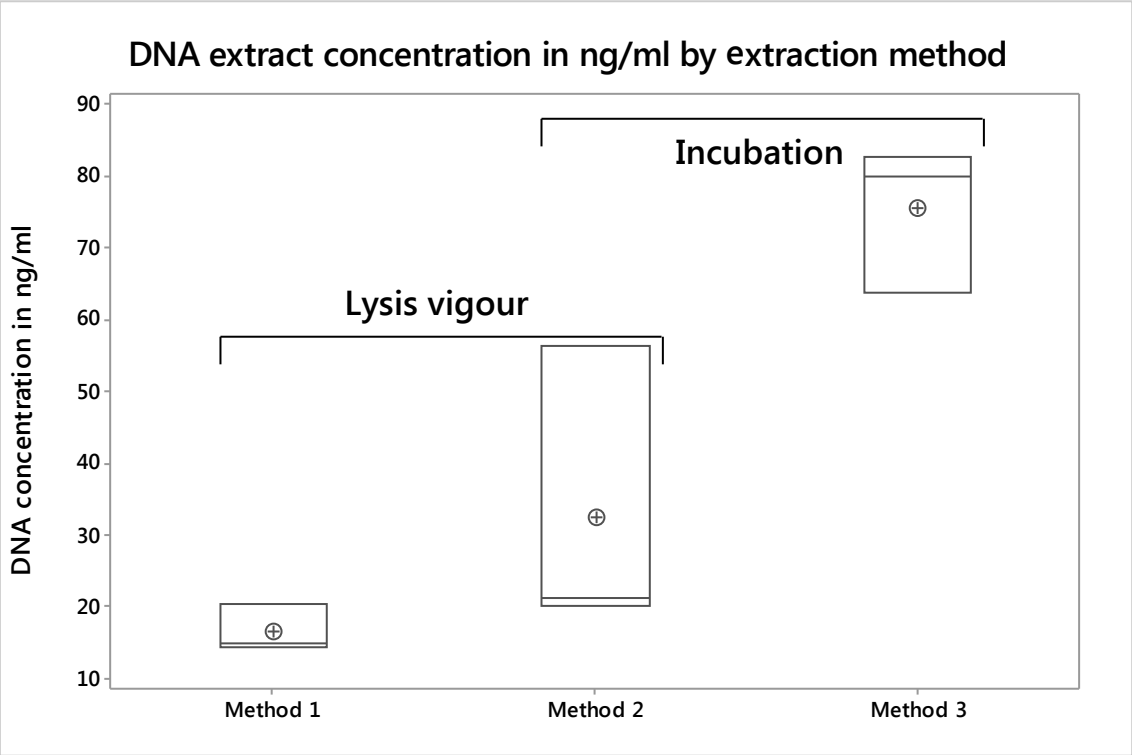


Figure 16: DNA extract concentration in nanograms per microliter where one sample from each of 3 ASPs was extracted using three methods; Method 1: standard manufacturer's protocol, Method 2: standard protocol with increased velocity and duration of bead-beating compared to method 1 (6 ms-1 for 160 secs), Method 3: as Method 2 with the addition of a 5 minute incubation of DNA binding matrix at 55 °c. The sampled sites were Tudhoe Mill nitrifying domestic activated sludge plant (ASP), Amble non-nitrifying domestic sequencing batch reactor ASP, and Aycliffe nitrifying ASP treating influent with some domestic and significant industrial effluent.

5.8.2 Comparison of the abundance of specific taxa-groups across DNA extraction and recovery methods

Given that the taxa of interest (Table 9) are highly phylogenetically divergent it is important to investigate the effect of DNA extraction and recovery method on the observed relative abundance of each functional group.

Foaming bacteria abundance was significantly higher using method 3 (ANOVA p = 0.001, Tukey's test method 1:2 p = 0.827, method 1:3 p = 0.002, method 2:3 p = 0.003, see Figure 17). The coefficient of variation was similar across methods though method 3 produced the most uniform extract concentration (33.01%, 36.97% and 25.27% for methods 1 to 3 respectively). Given that the genera in the foaming group (Table 9) feature high GC content (Ishikawa et al. 2004; Arenskötter et al. 2004; Muller et al. 2012) the observed increase in relative abundance using DNA extraction method 3 fits

the hypothesis that the addition of an incubation step improves recovery of high GC DNA.

NOB abundance was not significantly different between methods 1 and 3, however NOB abundance using method 2 was lower than method 1 ($p = 0.066$) or 3 ($p=0.02$, Figure 20) and had the lowest variation (coefficient of variation of 22.2%, 9.2%, 13.8%, for methods 1, 2 and 3 respectively). There were no significant differences in the abundance of bulking bacteria (Figure 18; ANOVA, $p = 0.10$) or AOB (Figure 19; ANOVA, $p = 0.89$) across the different extraction and recovery methods. The NOB (Ulrich & Zhulin 2010) and AOB (Chain et al. 2003; Kozlowski et al. 2016) do not feature high GC content genomes, providing further support to the DNA recovery hypothesis. The bulking group (Table 9) is quite phylogenetically diverse and features some genera of high GC content e.g. *Microthrix* (Muller et al. 2012) and others of low GC content e.g. *Thiothrix* (Bergey 2005). Therefore the effect of increased recovery of high GC content DNA from the bulking group appears to be insignificant.

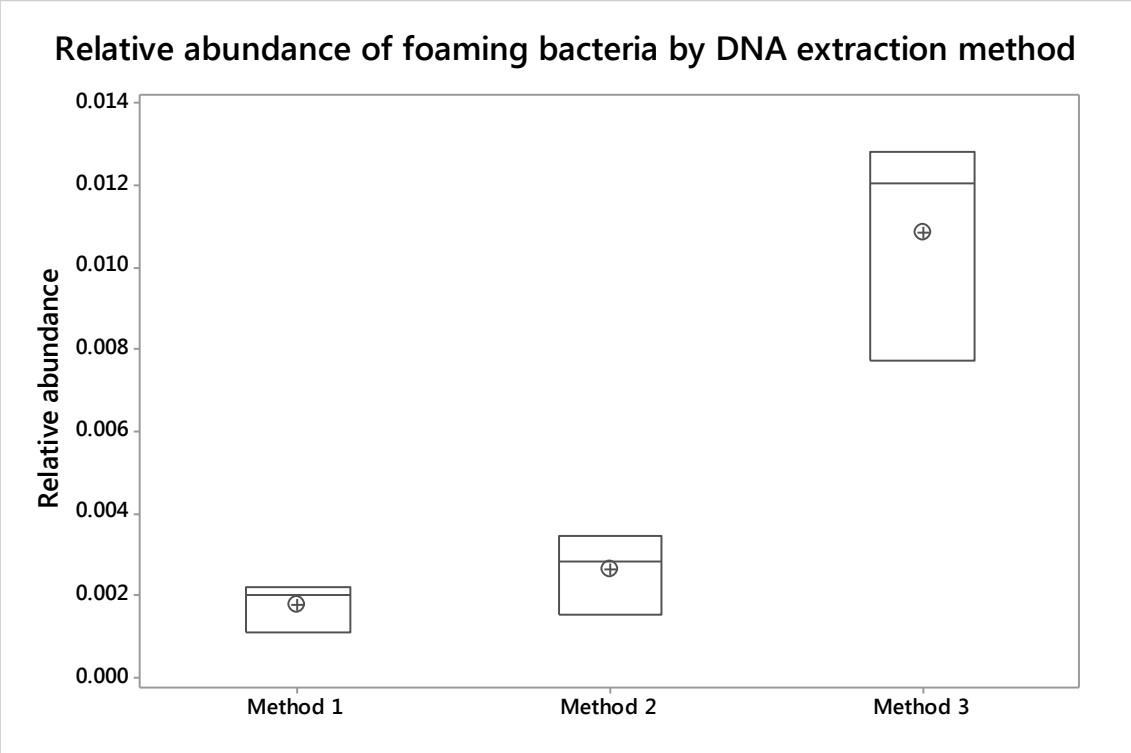


Figure 17 – Foaming bacteria relative abundance – see next page for detail

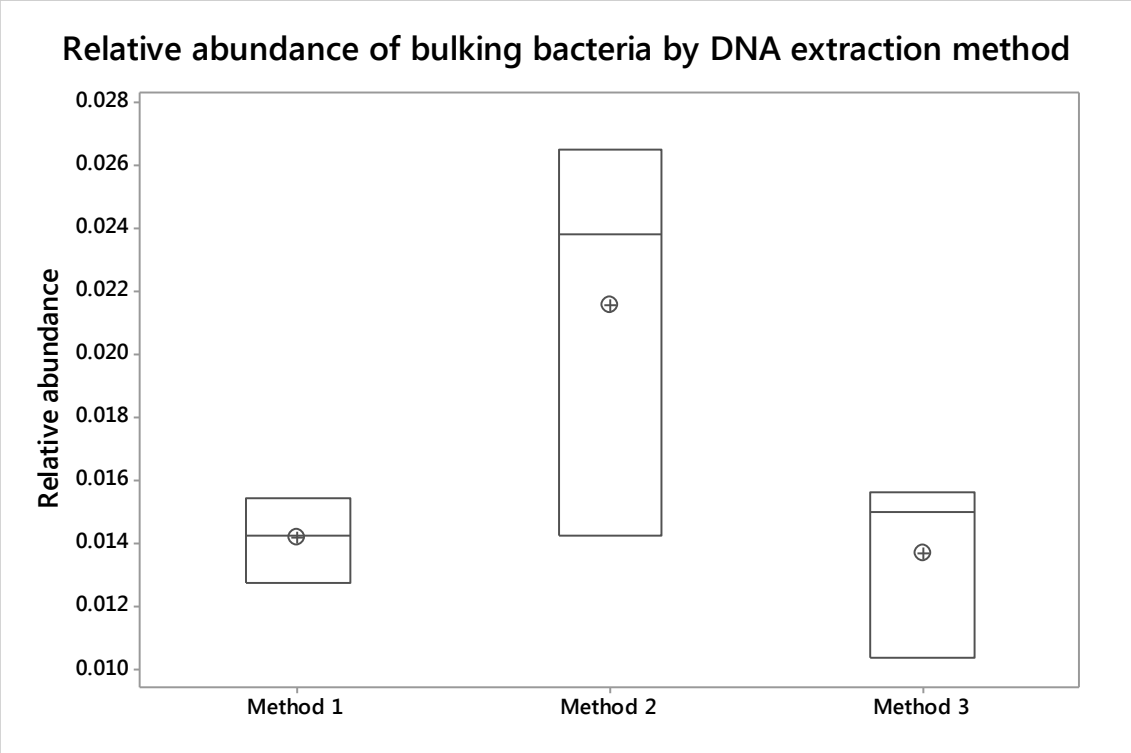


Figure 18: Bulking bacteria relative abundance – see next page for detail

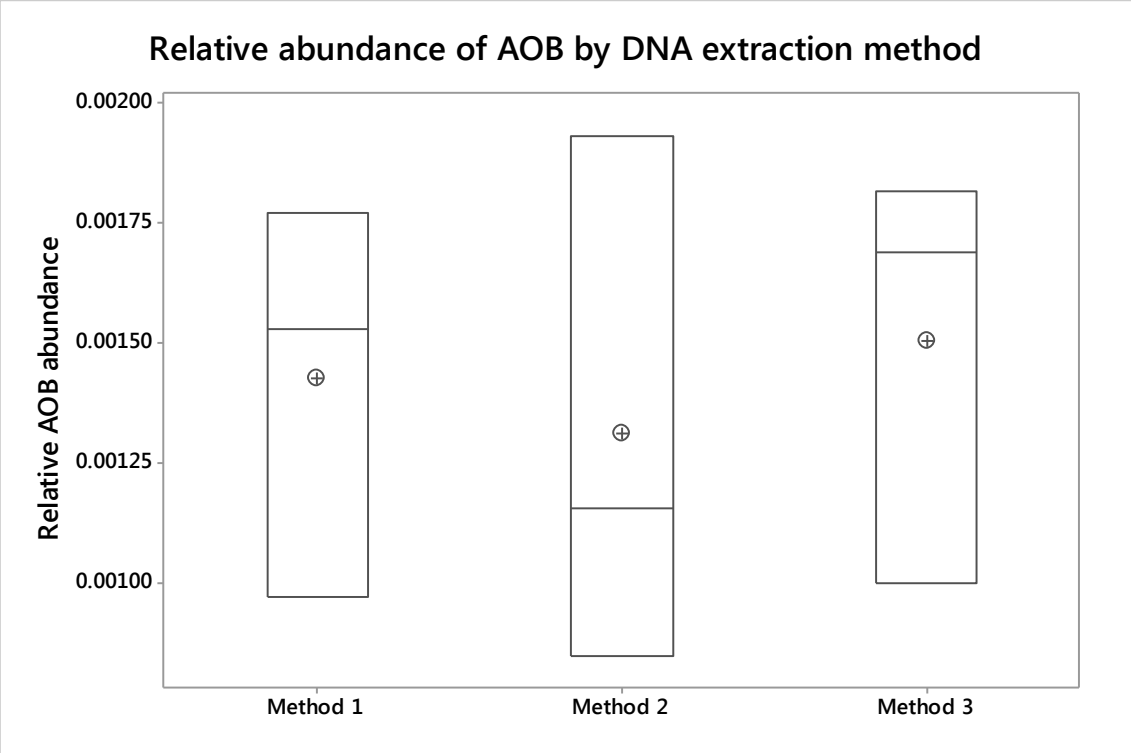


Figure 19: Ammonia oxidising bacteria relative abundance – see below for detail

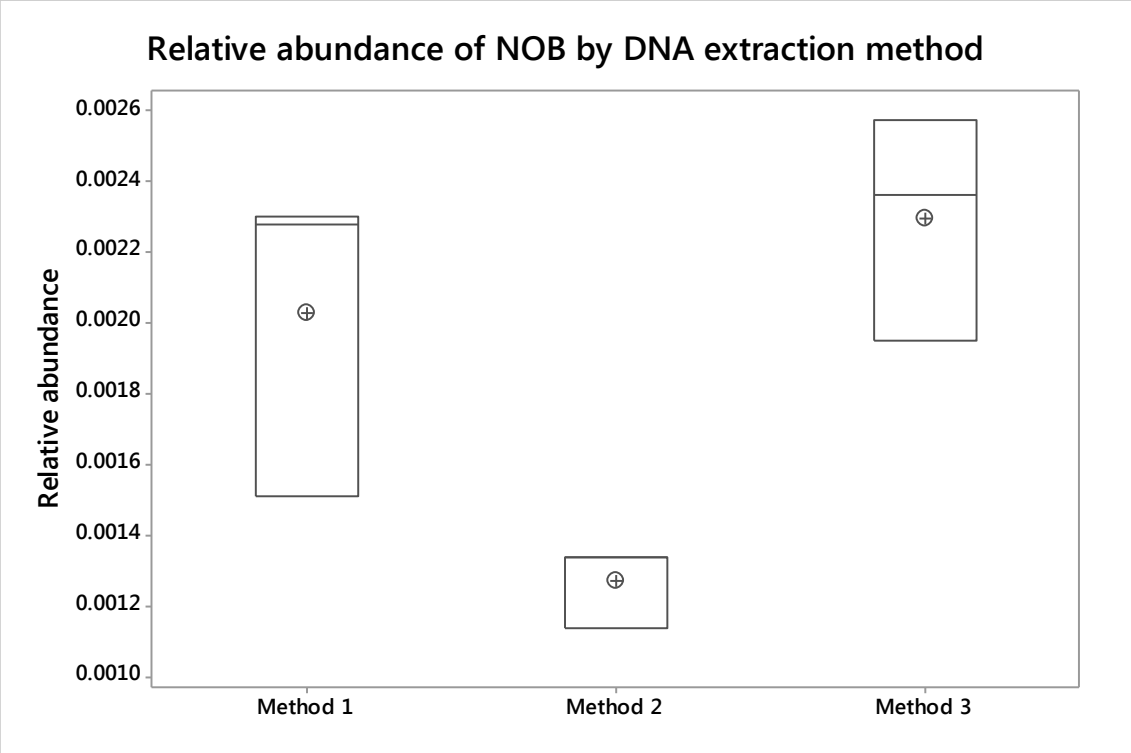


Figure 20: Relative abundance of nitrite oxidising bacteria in samples from 3 activated sludge plants (ASP) using manufacturer's protocol (Method 1), Method 1 with bead beating at 6 ms-1 for 160 seconds (Method 2), and Method 2 with a 5 minute incubation of DNA binding matrix at 55 °c (Method 3). Sample 1 = Tudhoe Mill nitrifying domestic ASP, 2 = Amble non-nitrifying domestic sequencing batch reactor ASP, 3= Aycliffe nitrifying ASP treating influent with some domestic and significant industrial effluent. Each sample was extracted using each method producing 3 groups of 3 DNA extracts. The bottom line of each box is the first quartile, the middle line the median, the top line the third quartile and the cross in a circle is the mean.

The impact of DNA extraction and recovery method on the observed total community structure was assessed using pairwise weighted (taking into account abundances) and unweighted UniFrac distances (see 5.6.1), which quantify the overall phylogenetic distance between communities (Figure 21 and Figure 22 respectively). It is evident that the underlying community diversity had a greater impact on community structure than the DNA extraction and recovery method since the communities cluster according to the sample from which they were taken. Samples one and two are in quite different relative positions when comparing the weighted and unweighted figures, whereas sample three remains in the same relative position. This may be explained by the fact that sample one is from a nitrifying domestic ASP and therefore has a longer sludge age than sample two ASP which is a domestic plant which does not nitrify. Therefore it is suggested that the plants share similar taxa, but the abundance of certain slow growing groups such as the nitrifiers is greater in sample 2 than 1. This would cause the observed shift in community distance when taxa abundance is taken in to account. Sample three does nitrify but receives significant industrial effluent and can therefore be expected to harbour quite different taxa and thereby be dominated by phylogenetic distance as opposed to differential taxa abundance, and therefore the difference between weighted and unweighted metrics is minimal. The statistical significance of the UniFrac distance between samples was calculated using the beta significance approach (see 5.6.2). The mean weighted UniFrac distance between methods 1 and 2 and methods 2 and 3 were not significantly different (2 sample t test, $p = 0.999$) though the test was lacking power due to only 3 observations per factor. The distance and significance data are detailed in Figure 23 and demonstrate that the impact of DNA extraction method varies depending on the microbial community structure being sampled. It is postulated that samples where the community structure was significantly different between methods 1 and 2 feature communities with difficult to lyse cells, whereas samples where the community structure was significantly different between methods 2 and 3 feature a greater proportion of taxa with high GC content genomes.

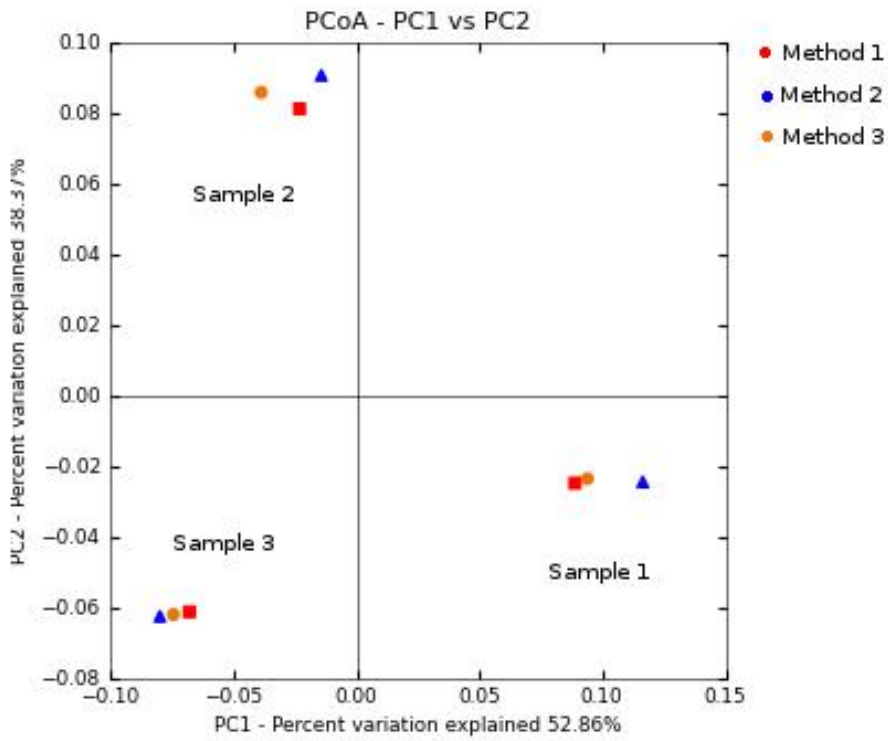


Figure 21: Principal coordinate analysis plot showing weighted UniFrac phylogenetic distance (beta diversity) between samples and DNA extraction methods. Method 1 = manufacturer's protocol,2 = method one with bead beating at 6 ms-1 for 160 seconds, 3 = method 2 plus a 5 minute incubation of the DNA binding matrix at 55 °c

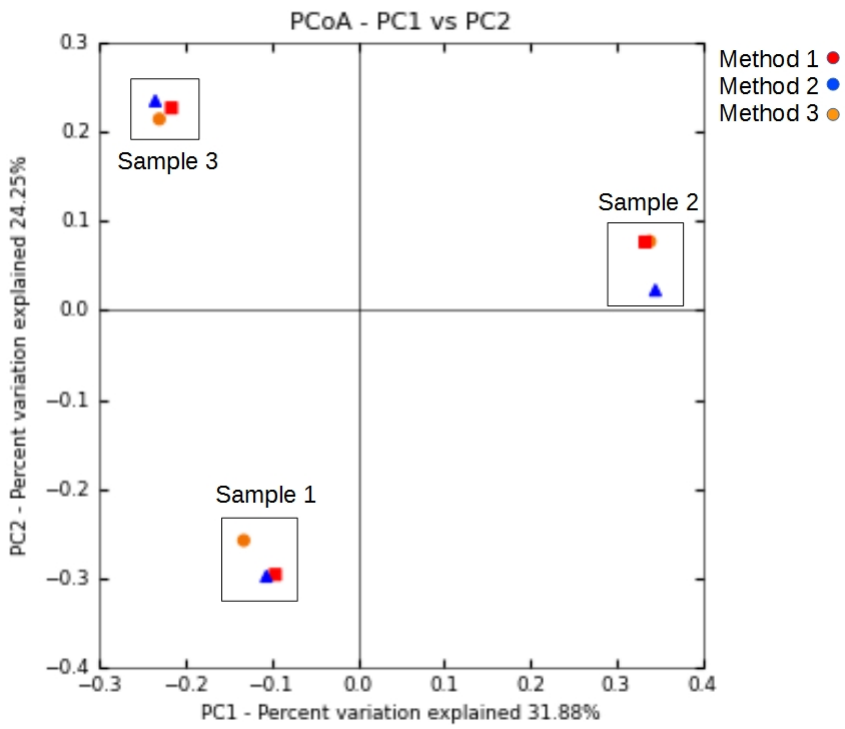


Figure 22: Principal coordinate analysis plot showing unweighted UniFrac phylogenetic distance (beta diversity) between samples and DNA extraction methods. Method 1 = manufacturer's protocol,2 = method one with bead beating at 6 ms-1 for 160 seconds, 3 = method 2 plus a 5 minute incubation of the DNA binding matrix at 55 °c

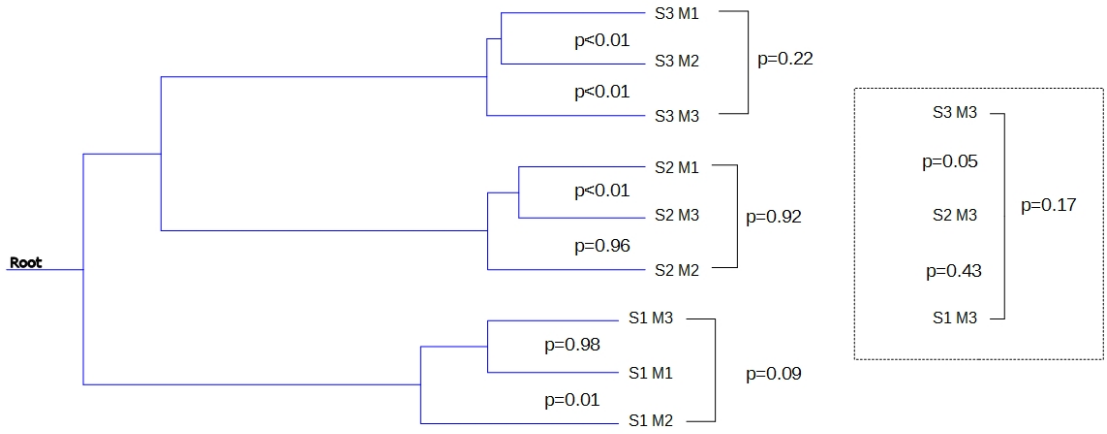


Figure 23: Cluster analysis using weighted UniFrac distance between samples from Tudhoe Mill nitrifying activated sludge plant (ASP) treating domestic sewage (S1), Amble non-nitrifying sequencing batch reactor ASP treating domestic sewage (S2) and Aycliffe ASP treating heavy industrial and domestic sewage (S3). DNA from each sample was extracted according to the standard protocol (method 1, M1), the standard protocol except bead beating at 6 ms-1 for 160 s (method 2, M2) and as method 2 with an additional incubation at 55 °c for 5 minutes (method 3, M3). The length of each branch is proportional to the UniFrac distance between samples. The indicated p values represent the significance of the UniFrac distance determined by 100 Monte Carlo random sequence assignments. Data points S1.M3, S2.M3 and S3.M3 in the right hand box represent samples from each plant where DNA extracted using the same method and were included to demonstrate the distance between samples within extraction method.

5.8.3 DNA extraction and recovery method: Discussion

As demonstrated in the previous section, it was vital to ensure that the experimental methods were able to differentiate between functional groups before attempting to use the method to do so. A key parameter in ensuring NGS results accurately reflect "real" community structure is DNA extraction method. The kit used has been demonstrated to be most suitable for ASP sample extraction but has a number of parameters that can be adjusted. Therefore it was decided to extract three samples using three different sets of extraction parameters (see section 5.7.4 and Table 8). The results showed that method 3 extracted significantly more DNA than methods one and two. There have been no other studies which have investigated the impact of incubation on DNA extraction. It is difficult to assess how much DNA is sufficient given that we cannot quantify the total amount of DNA present in a sample. Instead the key issue is whether or not the DNA present in the extract is representative of the DNA present in the sample both in terms of presence and relative abundance. There was an observed change in community structure between methods 2 and 3 which differed

only in a final incubation of DNA binding matrix. This was unexpected and as discussed in section 5.8.1 appears to be the result of different bond strengths between the silica matrix and DNA strands. This is an important and poorly understood effect which would benefit from further investigations (see 9.1).

5.9 Assessing the use of NGS for the quantification of process-important taxonomic groups

In order to achieve wider uptake of NGS it must be confirmed that the technology is able to accurately measure the abundance of process-important taxonomic groups. NGS derived AOB abundances were therefore compared against quantitative PCR and fluorescence in situ hybridisation (FISH), as trusted "gold standard" technologies, in samples previously analysed by these techniques. The precision of NGS abundance measurements was also assessed by measuring the coefficient of variation in NGS data at different read depths.

5.9.1 Comparing quantification of AOB by NGS against other quantitative molecular tools

Baptista et al (2014) quantified the abundance of ammonia oxidising bacteria (AOB) in samples from full scale (n = 12) and lab scale (n=10) reactors using fluorescence in-situ hybridization (FISH) and quantitative polymerase chain reaction targeting either the 16S rRNA gene (qPCR 16S) or the ammonia monooxygenase gene (qPCR amoA). In the present study the DNA extracts from those 22 samples were each sequenced without replication to compare NGS enumeration of AOB with existing techniques. Linear regression on all three data sets found a statistically significant ($p < 0.001$) positive correlation between AOB relative abundance measured by PGM and all three previously used methods. R squared values of 51.6%, 57.2% and 56.9% demonstrate that the PGM did not always predict the observations using the other techniques accurately (Figure 24, Figure 25, Figure 26 for FISH, amoA and 16S respectively). The PGM versus amoA residuals were randomly distributed and all but one prediction were within the 95% prediction intervals. The exception was reactor C5 which had a large abundance of Nitrosomonas LS79A3 which was not targeted by the FISH or 16S probes used by Baptista et al, leading to the apparent over estimation of AOB by PGM. Similarly the PGM versus 16S and PGM versus FISH regressions showed a single large residual from reactor C5. The residual standard deviations were 0.339 (FISH), 0.328 (16S) and 0.327 (amoA). Therefore the regression analyses show a statistically

significant relationship between AOB abundance measured by PGM and existing techniques, but highlight the discrepancies between FISH and qPCR which used predetermined primers and PGM which utilizes a greater proportion of the 16S gene (section 5.2) alongside a taxonomic database (section 2.5.4) to enumerate a greater proportion of the AOB.

Therefore whilst the PGM produces useful results, this comparison does not in itself establish the suitability of the PGM for ASP prediction and diagnosis. One PGM limitation is the comparison of relative abundance data with absolute abundance data from qPCR and FISH. Therefore cell counts were undertaken on 5 samples in order to convert PGM relative abundance in to absolute abundance. However the small number of samples did not provide sufficient data to draw any firm conclusions. Another possible limitation is the issue of differential taxonomic coverage between PGM data and the taxa targeted by the FISH and qPCR probes used. Data relating to these issues are included in the appendices and further discussed in section 5.9.3.

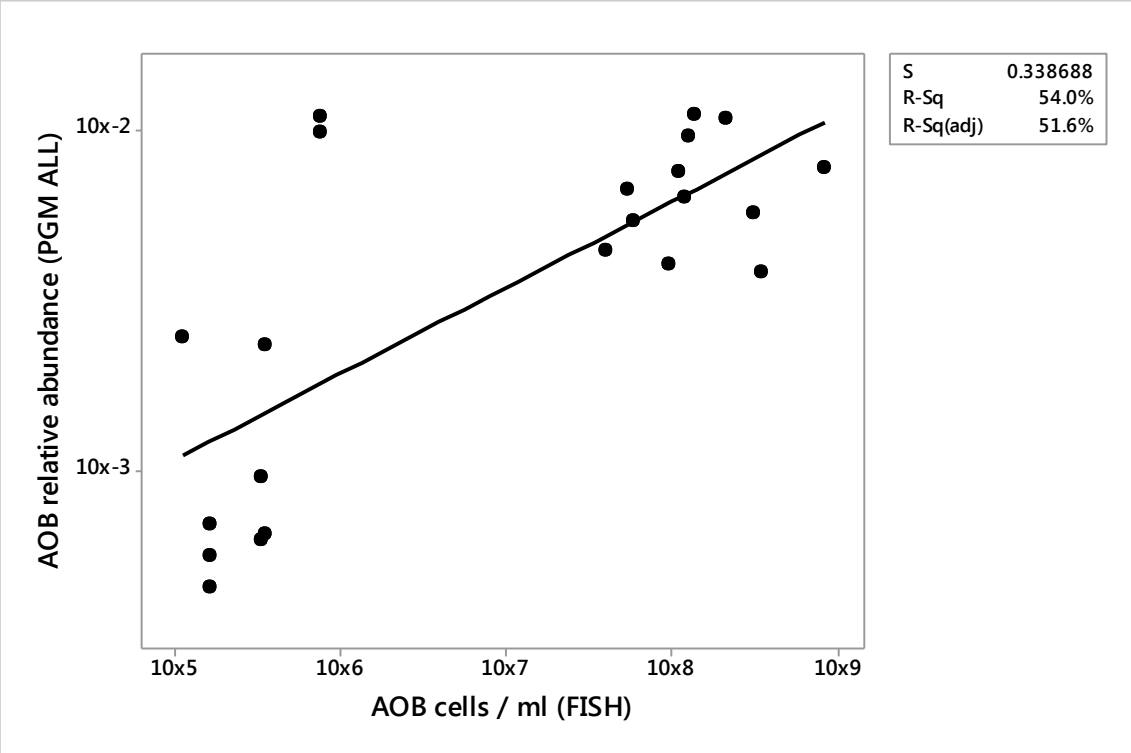


Figure 24: Regression analysis of abundance of ammonia oxidising bacteria (AOB) measured by the Personal Genome Machine (PGM ALL where the values indicate relative abundance) and fluorescence in situ hybridization (FISH where values indicate the number of cells per millilitre) in 10 lab scale and 12 full scale nitrifying activated sludge reactors. Each data point represents the number of AOB cells per ml as measured by FISH and NGS. The number denoted by the letter S is the regression equation standard error. Regression equation: log10(PGM r.a. (all)) = - 4.241 + 0.2541 log10(FISH)

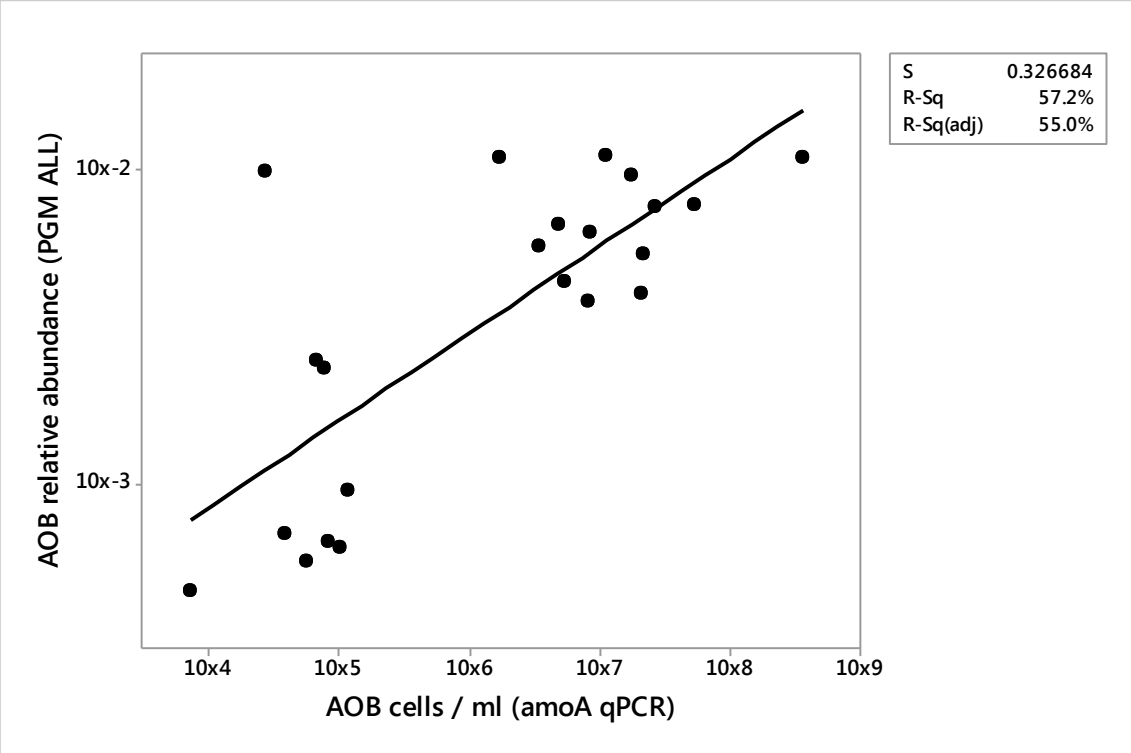


Figure 25: Regression analysis of abundance of ammonia oxidising bacteria (AOB) measured by quantitative polymerase chain reaction (qPCR) targeting the amoA gene and Personal Genome Machine (PGM) in 10 lab scale and 12 full scale nitrifying activated sludge reactors. Each point represents the number of AOB cells per ml measured by FISH and NGS. The letter S denotes the regression standard error. Regression equation: log10(PGM r.a. (all)) = - 4.189 + 0.2781 log10(amoA qPCR).

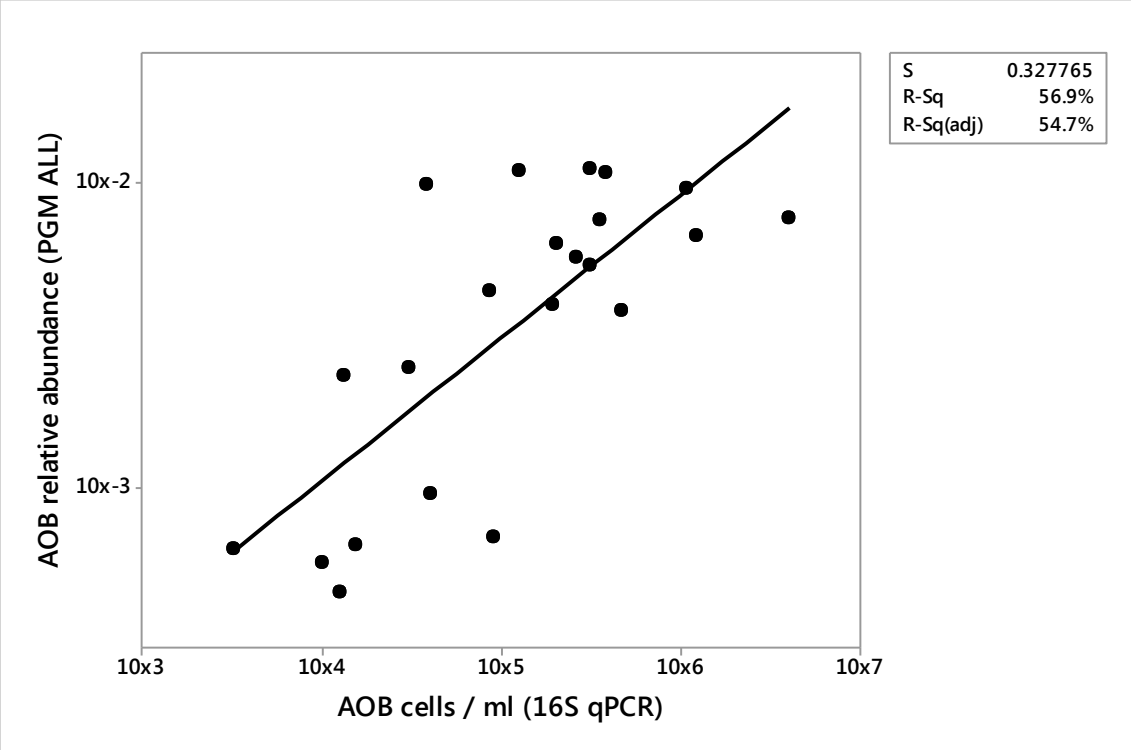


Figure 26:As Figure 25 except qPCR targeted 16S rRNA gene. Regression equation: log10(PGM r.a. (all)) = - 4.859 + 0.4705 log10(16S qPCR).

5.9.2 Evaluating the effect of sampling depth on precision of AOB abundance measurements

It is widely known that the number of observations of individuals in a sub-sample affects the precision of the measurement. Each molecule of DNA in an amplicon pool which is sequenced can be considered an observation of an individual microbe and therefore the greater the number of reads per amplicon pool the greater the precision of taxonomic abundance data. However increased reads come at the cost of sample throughput and therefore sample analysis cost. It is important to establish at what point the number of reads provides an acceptable level of precision as identified by comparison with that of the best existing techniques. One measure of precision is the coefficient of variation, which is the ratio of the standard deviation to the mean, and provides a measure of the dispersion of a frequency distribution. To establish the coefficient of variation in AOB relative abundance observations, one of the samples described in the previous section was sequenced to a depth of 1 million reads. This data was then randomly rarefied with replacement at read depths ranging from 10,000 to 1m in 10,000 read intervals. Each rarefaction was repeated 10 times. There was no significant difference in mean AOB relative abundance across all read depths (mean of 0.4%), however variance between read depths was significant (ANOVA Welch's test $p=0.272$, Bartlett's test statistic 1334.42 $p < 0.001$; Figure 17). Bartlett's test does not provide pairwise p values and so the read depth at which variance in observed AOB abundance was no longer significantly different could not be determined. At only 10,000 reads the NGS coefficient of variation (16%) is substantially lower than that reported for the current gold standard technique - fluorescence in situ hybridization (20 – 50%) (Davenport & Curtis 2004; Coskuner et al. 2005). In addition, 10,000 reads is considered sufficient to differentiate between closely related bacterial communities (Lemos et al. 2011; Caporaso et al. 2011). At this read depth mean relative AOB abundance was 0.04% with relative standard error of 5%. When applied to sample A24 25/07/12 from a stable nitrifying plant, CSAOR is 7.26 +- 0.34. Given that CSAOR is thought to range between 5 and 8 in a stable plant (see section 8.1.1) the precision of AOB enumeration at 10,000 reads is acceptable.

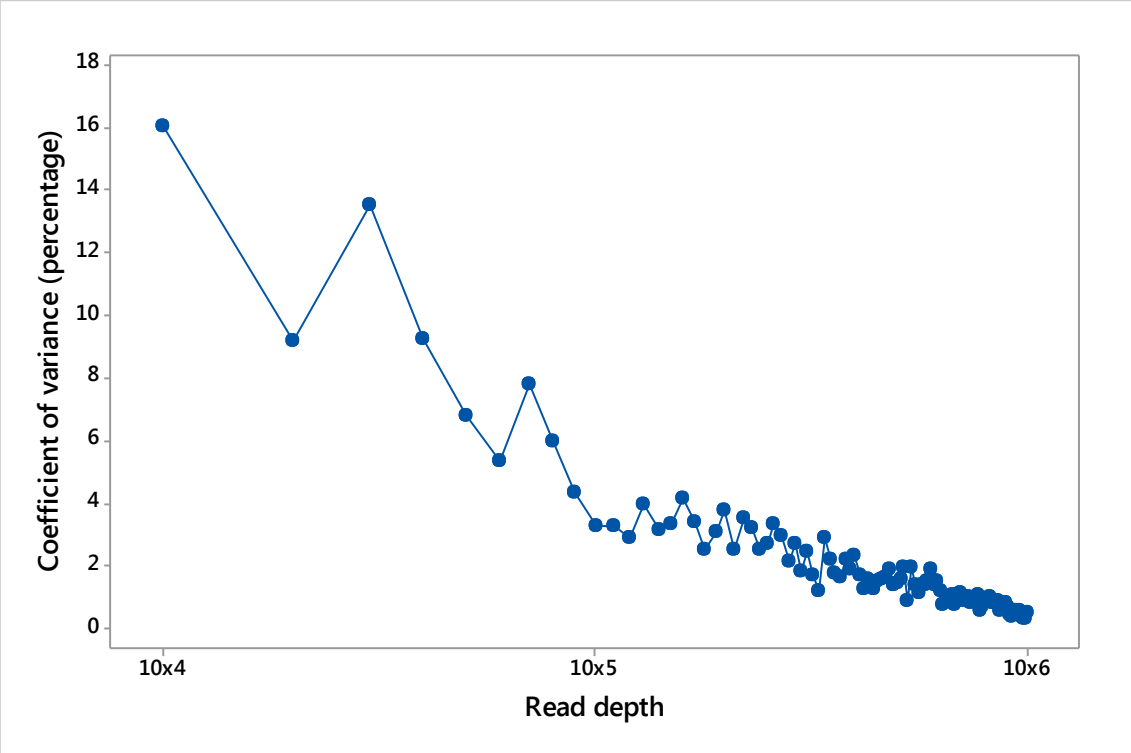


Figure 27: Coefficient of variation of total ammonia oxidising bacteria relative abundance determined by next generation sequencing at multiple read depths where reads were selected by random rarefaction with replacement 10 times at each read depth

5.9.3 NGS AOB enumeration: Discussion

Regression analysis between AOB abundance observed using NGS and qPCR or FISH showed a significant correlation. However the R^2 values of approximately 55% in all cases suggested there was a large proportion of variance which the model did not predict. The sources of the predictive discrepancies may be numerous. The number of 16S rRNA gene copies per microbial genome varies between 1 and 15 (Angly et al. 2014) which may skew the observed relative abundance of AOB. The taxonomic coverage of FISH probes used by Baptista et al (2014) may be different from that of the 16S rRNA V4V5 NGS primers used in this study. As the authors did not state the full taxonomic coverage of their probes, the probe sequences were analysed using ProbeBase (Pruesse et al. 2007). All three probes were found to target non AOB taxa. Of particular concern is the nso1225 probe which targeted a large number of non AOB taxa as well as the NOB genera *Nitrospira*, *Nitrobacter* and *Nitrotoga*. Regular review of the suitability of probes is required and unlike NGS the analysis cannot be rerun in the light of expanded 16S rRNA database sequence content. Also as discussed in section 5.9.2 NGS has a coefficient of variance significantly lower than that of FISH.

Using Figure 27 the NGS AOB abundance coefficient of variation is estimated to be 3 to 16% in comparison with reported FISH coefficient of variation of 55% (Davenport & Curtis 2004). Therefore it may be that the NGS data is more precise than FISH.

5.10 Microbial source tracking method development

5.10.1 NGS enumeration of faecal indicator bacteria: Results

Figure 28 plots the abundance of both Enterococci and *E. coli* in samples taken by the EA, and shows no obvious correlation between the two, despite their both being indicators of faecal contamination. Figure 29 shows EA and NWL culturing results for 5 samples taken contemporaneously. The data was plotted as an indication of the variability observed when using culture based FIB enumeration using the same method on a notionally identical sample, and showed no correlation.

Regression analysis of Enterococci CFUs detected by the EA and Enterococci relative abundance (Figure 30) was undertaken to establish whether there was a correlation between Enterococci enumerated by culturing and by PGM analysis. This was not a core hypothesis in this study, but was thought useful given that the PGM samples were mostly taken contemporaneously with EA samples. As previously discussed, the PGM lower limit of detection was two orders of magnitude greater than the EA's. Enterococci CFU counts showed very weak positive correlation with PGM Enterococci relative abundance, where relative abundance is the proportion of all taxa in the sample identified as Enterococci (R^2 = 16%, Figure 30). In addition Enterococci were detected by the EA in a number of samples yet not by PGM and vice versa.

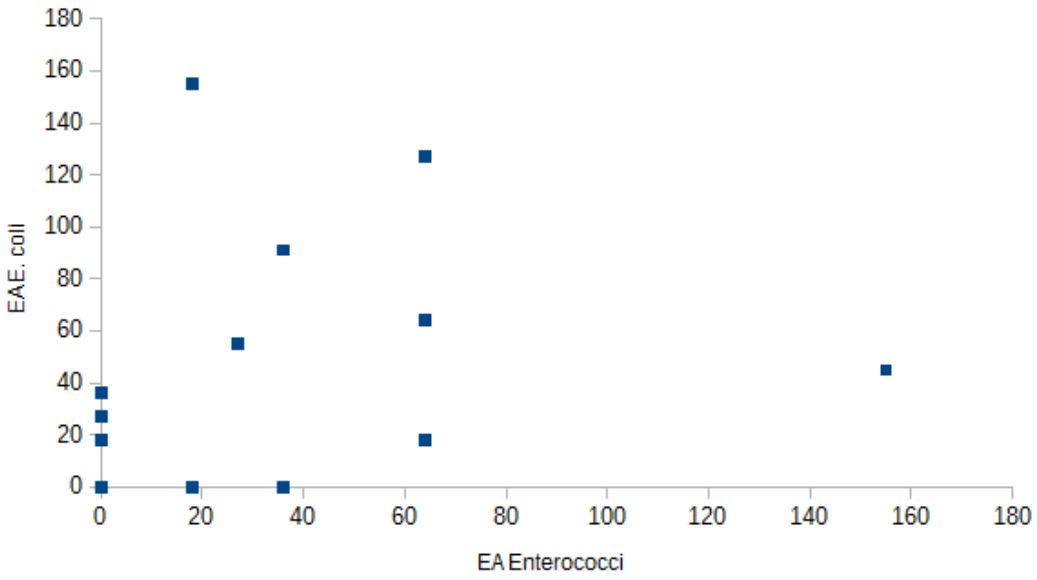


Figure 28: Enterococci and *E. coli* colony forming units (CFU) per 100 ml seawater sample from Saltburn bathing water where analysis was undertaken at the Environment Agency (EA). Samples were taken at the same time and place. n = 23 samples.

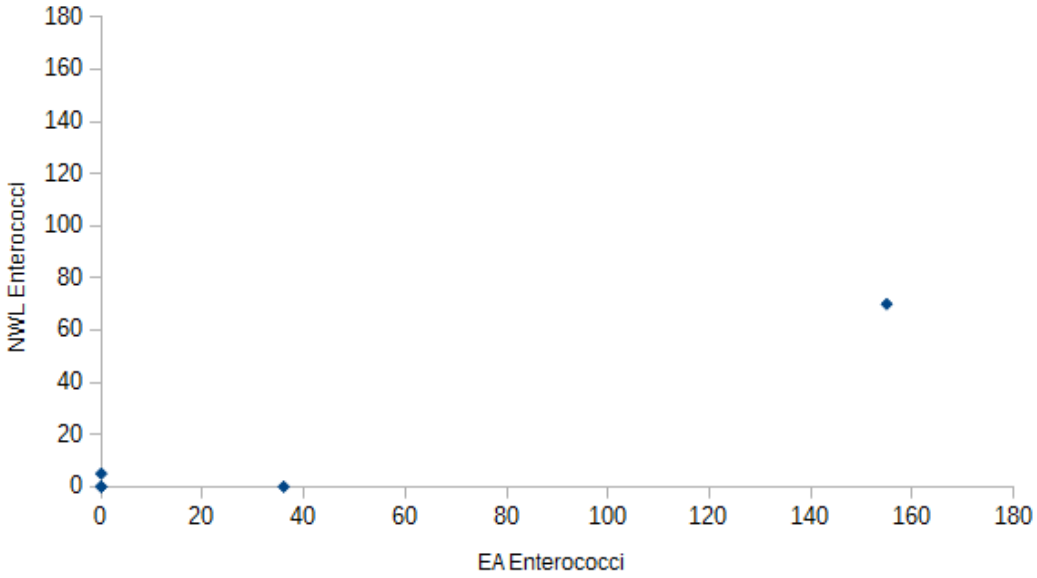


Figure 29: Enterococci colony forming units (CFU) per 100 ml seawater sample from Saltburn bathing water where analysis was undertaken at the Environment Agency (EA) or Northumbrian Water (NWL) labs. Samples (n=5) were taken at the same time and place.

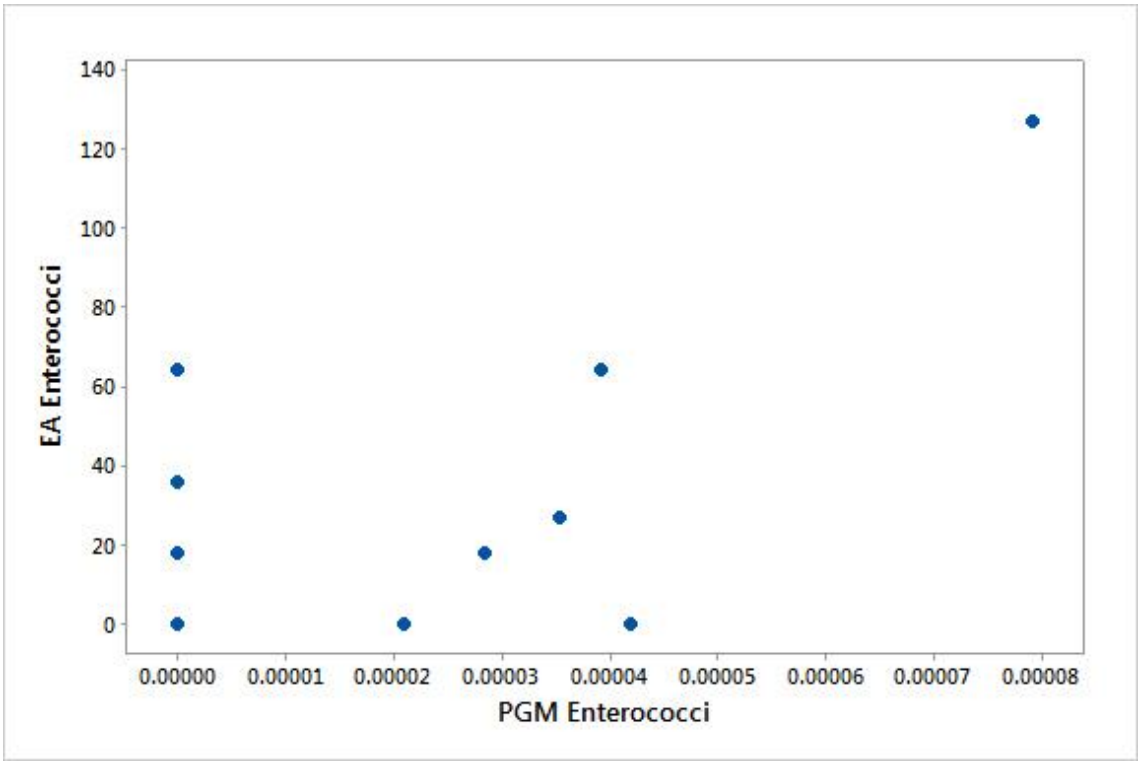


Figure 30: Enterococci colony forming units (CFU) per 100 ml seawater sample from Saltburn bathing water where analysis was undertaken at the Environment Agency (EA) or Newcastle University (NU) laboratories. EA measurements used standard culturing techniques to enumerate Enterococci. NU methods used 16S rRNA gene DNA sequencing by Ion Torrent Personal Genome Machine (PGM) producing relative abundance data. Samples were taken at the same time and place. n = 23 samples

None of the taxa classified to the *Escherichia-Shigella* genus could be further classified to species thereby preventing investigation of correlation between EA *E. coli* and PGM *E. coli* abundance. There was no significant correlation between the abundance of the Escherichia-Shigella genus abundance and EA *E. coli* ($p=0.9$, $r^2=0.1$).

5.11 NGS enumeration of faecal indicator bacteria: Discussion

This was not a core hypothesis as NGS analysis was not under investigation as a replacement for culture based FIB enumeration. Furthermore the NGS lower limit of detection was two orders of magnitude greater than the culture based methods. In addition, it is recognised that a single sample from a bathing water cannot accurately enumerate the FIB in the bathing water, due to heterogeneity of FIB spatial distribution in the water column, the transient nature of faecal plumes, the impact of sediment disturbance and many other factors (European Parliament and Council 2012; Oliver et al. 2014; Figueras et al. 1997; Georgiou & Bateman 2005). However as the

samples were mostly collected contemporaneously with the EA samples, it was felt useful to investigate any possible correlation between the two enumeration methods.

There was weak or no correlation between the faecal indicator bacteria *E. coli* and Enterococci as enumerated by Environment Agency culturing methods and next generation sequencing. There are numerous potential causes throughout the analysis process, from sample collection to extraction of relevant abundance data. The calculated theoretical PGM limit of detection in sea water at 10,000 reads is 10^3 bacteria per 100ml whereas the EA cites detection down to 10 per 100ml, 2 orders of magnitude less than PGM. Given that the lowest number of *E. coli* or Enterococci detected by the EA during the sampling period was 173 per 100ml, PGM limit of detection may be the primary cause of poor correlation. Deep sequencing of samples would mitigate the PGM limit of detection observed in this study. To match the EA limit of detection, 10m reads per sample are required at an approximate cost of £750 in reagents alone (ThermoFisher, Massachusetts USA). This is in contrast to £17 for culture analysis of both *E. coli* and Enterococci (Waites 2012). The limit of detection issue necessitates ongoing high costs and therefore whilst PGM may be capable of accurate and precise measurement of FIB abundance, costs outweigh any potential benefits such as greater taxonomic precision over culture plates growing non-target taxa.

Whitman and Nevers (2004) found that a single sample is only 30% precise when assessing *E. coli* abundance by culture methods, and several authors have reported high spatial and temporal heterogeneity of FIB distribution in bathing waters (Boehm 2007; Bonilla et al. 2007; Lavender & Kinzelman 2009), which suggests that even when samples are taken simultaneously they may not contain the same number of FIB. Therefore the sample itself may have contributed to some of the variance observed despite most of the samples being collected simultaneously with the EA. It is also unclear whether the taxa which grow on the agar media stipulated by the regulatory tests (Environment Agency 2012) are restricted to *E. coli* and *Enterococcus* and therefore the PGM data may not encompass all relevant taxa (Leclerc et al. 1996; Scott et al. 2002). It is possible that whilst the selected primers target the 16S rRNA gene of the majority of Bacteria and Archaea, the gene sequence in the *E. coli* V4-V5 region may not be sufficiently variable and distinct to reliably assign or classify them

taxonomically (Lukjancenko et al. 2010). This may also be the case for some *Enterococcus* (Patel et al. 1998). Other potential sources of error include varying 16S rRNA gene copy number (and thereby PGM overestimation of FIB abundance) (Kembel et al. 2012) and the detection of 16S rRNA genes from extra-cellular DNA (Wagner et al. 2008) or viable but not culturable cells (Trevors 2012).

It is important to note that whilst there was no significant correlation between direct measurement of FIB by EA and PGM, partial least squares analysis of SourceTracker results did establish a significant correlation between EA Enterococci and SourceTracker. Therefore whilst the PGM may not be suitable for direct enumeration of FIB, the platform is useful for elucidating their source.

5.11.1 Faecal host marker sensitivity and specificity: Results

The presence of all faecal host marker taxa identified in the literature search were checked in the SILVA database (Quast et al. 2013) by manually searching for the organism name in SILVA taxonomy files where the marker taxonomy was specified in the original paper. The ruminant marker genus 5-7N15 (Reischer et al. 2006) was not present in SILVA v119. The SILVA tool TestPrime identified the taxa targeted by the Reischer et al (2006) primers as a number of uncultured taxa in the *Bacterioides* genus. It was therefore decided to remove this marker from results as the target taxa were not named and therefore incompatible with the sequencing and taxonomic assignment approach. The same approach was used to identify pig specific PG10 markers (Gómez-Doñate et al. 2012) which identified two target taxa; *Bifidobacterium* sp. SMst02 and *Alloscardovia* sp. OB7196. Therefore the moniker “PG10” was used as a proxy for the detection of either taxa. The presence or absence of each marker in the PGM data was noted for each faecal sample (Table 11). A number of markers were less than 100% specific; the Human 1 marker was only 50% specific. The Cow 1 marker was not detected in the single cow sample. The specificity of the pig markers could not be assessed as a pig faecal sample could not be sourced. Human markers 3 and 5 and chicken marker 2 were found to be 100% sensitive and specific. Chicken marker 1 was detected in raw sewage but not in non-target faecal samples, which may indicate the presence of non-human faeces in sewage. This is presumably due to surface water runoff containing agricultural faeces and highlights the difficulty of delineating faecal bacteria from NWL versus non NWL sources.

Table 11: Sensitivity and specificity of each faecal host marker bacteria in each faecal sample collected. Each sample identifier e.g. B28 represents a single sample and therefore where multiple identifiers are referenced this indicates the number of samples from that host.

Faecal host marker bacteria		**Faecal/Environmental sample**									
Taxonomy	**Host**	**Cow (B28)**	**Seagull (B36,B37)**	**Chicken (B39,B40)**	**Dog (B41)**	**Horse (B42-45)**	**Sheep (B50)**	**Poultry/Horse (B51)**	**Raw sewage (B29,B30)**	**Sensitivity**	**Specificity**
Methanobrevibacter smithii	Human 1	YES	-	-	-	YES	YES	-	YES	100%	50%
Bifidobacterium adolescentis	Human 2	YES	-	YES	-	-	-	-	YES	100%	75%
Bifidobacterium dentium	Human 3	-	-	-	-	-	-	-	YES	100%	100%
Bifidobacterium pseudolongum	Cow 1	-	-	-	-	-	-	-		0%	100%
Bifidobacterium saeculare	Chicken 1	-	-	YES	-	-	-	-	YES	100%	85%
Brevibacterium avium	Chicken 2	-	-	YES	-	-	-	-		100%	100%
Bacteroides caccae	Human 4	-	-	-	-	-	-	YES	YES	100%	92%
Bacteroides fragilis	Human 5	-	-	-	-	-	-	-	YES	100%	100%
Catellicoccus marimammalium	Gull 1	-	YES	YES	-	-	-	-		100%	83%
Lactobacillus amylovorus	Pig 1	-	-	-	-	-	-	-		ND	100%
Bifidobacterium PG10	Pig 2	-	-	-	-	-	-	-		ND	100%

The EA Enterococci abundance and relative abundance of each faecal host marker was converted to binary presence or absence. Binary logistic regression found no significant relationship between detection of faecal host markers and EA detection of Enterococci (p=0.37). Similarly no significant relationship was found between detection of faecal host markers and EA detection of *E. coli* (p=0.72). However multiple regression showed that the presence of Human markers 1 and 2 correlates with the presence of Enterococci (p= 0.012, r^2 35.8%).

5.11.2 Faecal host marker sensitivity and specificity: Discussion

A number of the host markers investigated in this study are not 100% specific or sensitive and therefore should be used in combination to increase overall specificity, reflecting previous results (Jeong et al. 2011; Roslev & Bukh 2011; Johnston et al. 2013). The faecal host markers were selected based on a literature review at the time. However subsequent studies have cast doubt on the specificity of some markers used. *Methanobrevibacter smithii* in particular seems to be a poor candidate for a human faecal marker due to extensive cross reactivity. In this study it was found in cow, horse and sheep faeces which reflects recent findings (Cersosimo et al. 2016; Seedorf et al. 2015). *M. smithii* has also been found in pigs (Federici et al. 2015). Despite these findings *M. smithii* continues to be used as a human faecal marker (Bianco et al. 2015; Liang et al. 2015). The findings in this study in combination with the aforementioned studies suggest *M. smithii* should no longer be used as a human faecal marker.

A number of review papers (Ahmed et al. 2016; Tran et al. 2015; Tan et al. 2015) and journal special issues (Wuertz & Reis 2013; Wuertz & Reis 2010) have attempted to provide an overarching assessment of the current state of knowledge regarding faecal host markers. However there is no online, regularly updated centralised database of faecal host markers containing current knowledge of specificity and sensitivity. It was concerning that the single cow marker in this study *Bifidobacterium pseudolongum* was not detected in the cow faecal sample, despite previous studies suggesting its ubiquity amongst cows (Gómez-Doñate et al. 2012). It has been suggested that *B. pseudolongum* may be of low abundance in bovine faeces (Balleste & Blanch 2011) and therefore the failure of the PGM to detect the marker may be due to limits of detection. Furthermore the number of faecal samples in this study was low, with only

one cow sample. However this should not be an issue if a marker is considered 100% sensitive, which is the ideal case for faecal host markers. Some markers were found to be sufficiently specific and sensitive to be of use in microbial source tracking (see Chapter 7).

5.11.3 Change in CSO microbial communities over time and space: Results

Knights et al (2011) established that SourceTracker is able to distinguish samples where Jensen-Shannon divergence is 0.05 or greater. The sample collected at 14:22 from the Skelton CSO (site B7) was taken before the CSO was flowing according to the NWL monitoring systems. This sample is less similar to the two subsequent Skelton CSO samples taken at 17:58 (0.550) and 21:00 (0.530), than they were to each other (0.195) but still above the SourceTracker threshold of 0.05. The Skelton B7 17:58 and Leyland B9 18:50 samples were quite similar (0.171) yet still above the threshold, indicating that individual CSO microbial community signatures can be distinguished by SourceTracker (see Table 12).

In order to gain an empirical indication of the specificity of point source signatures in microbial community composition, a sample was taken from Marsden beach bathing water (B35, Table 6), which is not directly impacted by human or agricultural faecal inputs. None of the point sources from the Saltburn catchment were apportioned to the geographically and hydrologically disconnected Marsden bathing water sink (Figure 31). There is a significant seabird colony immediately adjacent to the bathing water sampling point, explaining the gull faecal source. However the presence of agricultural faeces was unexpected. Marsden is potentially impacted by faecal contamination present in the Tyne estuary 4km North, as well as any run off from animals present immediately inland, although this has not been verified.

Table 12: Jensen-Shannon divergence matrix for samples from five CSOs in the Saltburn bathing water catchment during a single storm event

	Skelton CSO B7 17:58	Brotton CSO B16 18:15	Guisborough CSO B8 18:44	Leyland CSO B9 18:50	Caravan Park CSO B33 17:15
Skelton CSO B7 17:58	0	0.526	0.183	0.171	0.489
Brotton CSO B16 18:15		0	0.593	0.546	0.335
Guisborough CSO B8 18:44			0	0.189	0.551
Leyland CSO B9 18:50				0	0.504
Caravan Park CSO B33 17:15					0

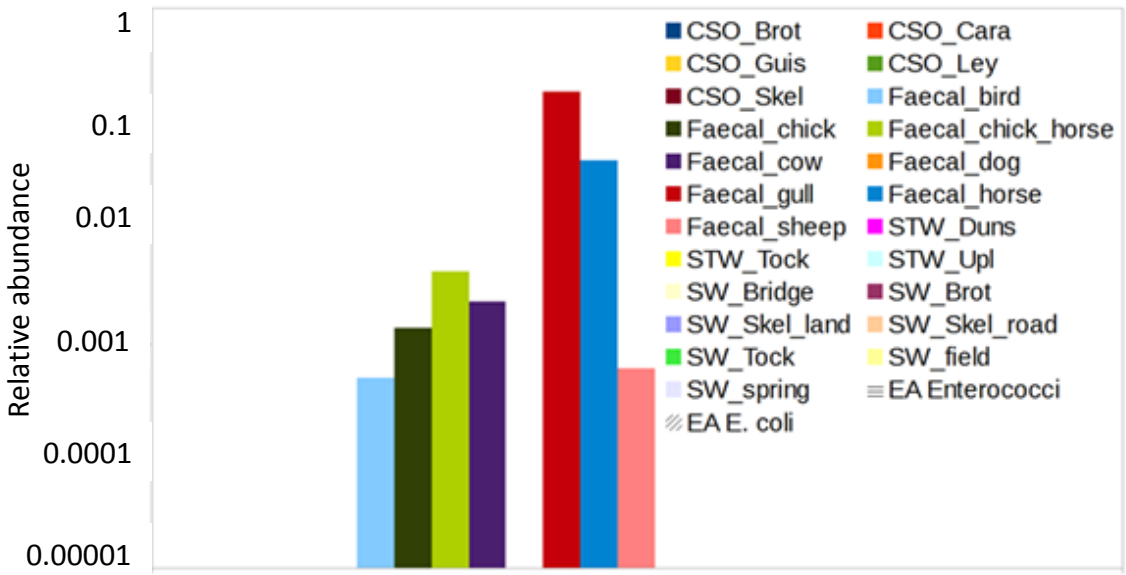


Figure 31: SourceTracker results showing the relative contribution of microbial communities from different sources (identified in the key which reads left to right) to the bathing water (sink) at Marsden Bay. STW = Sewage Treatment Works, Surface = Surface water drains. CSO = Combined Sewer Overflow.

5.11.4 Change in CSO microbial communities over time and space: Discussion

Empirical observation suggests that point sources that contribute to bathing water quality can be distinguished using the whole community signature approach (SourceTracker). The results from the storm of August 10th show the absence of CSO signatures when they were not flowing, and the subsequent appearance of the signatures several hours after the CSO spill began. In addition, CSO signatures were not detected in the catchment when the CSOs had not recently operated. This is the first time that a storm event and individual CSOs have been sampled and analysed by NGS for MST. Neither the CSO nor the sewage treatment work signatures from Saltburn were detected in the Marsden bathing water sample when used as a sink, suggesting that the signatures are sufficiently specific to distinguish individual point sources contributing to a particular sink. This finding is in accordance with other studies (Shanks et al. 2013). However, it is not borne out by the community structure results in section 5.11.3, which showed that samples from the same CSO within the same storm can differ more than samples from different CSOs.

Chapter 6. Activated Sludge Process Monitoring Results

Using the optimised protocols from the Methods section, samples from 32 activated sludge plants (ASPs) were sequenced and the taxa present grouped at "species" level, defined as operational taxonomic units (OTUs) grouped at 97% similarity cut-off. These groupings were then identified to the highest taxonomic level possible.

6.1 Overview of taxa in activated sludge plants

Samples from 32 ASPs treating domestic and industrial sewerage were analysed using the previously described protocols. 31 plants were sampled once, and one plant sampled monthly for 12 months. The microbial communities varied between ASP and across time within a single ASP (Figure 32). There were 42 OTUs observed in all samples, the majority of which are not thought to be vital to plant performance. This paradox between universal presence yet undertaking no known ASP function has been noted in previous studies and is suggested to be largely due to lack of ASP process understanding (Saunders et al. 2015). The relationship between the abundance of known functional groups and plant stability is further investigated below.

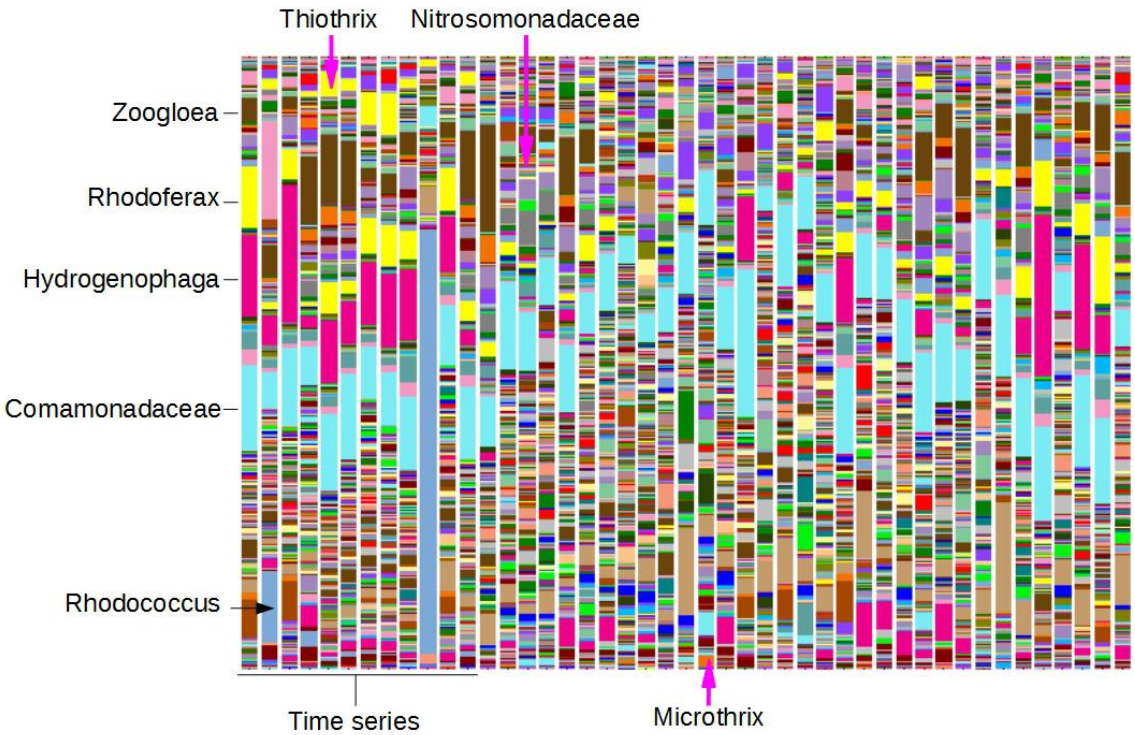


Figure 32: Taxa summary of activated sludge plant samples where each colour represents an OTU. Selected taxa are identified to illustrate the large abundance of bacteria with unknown function e.g. *Hydrogenophaga* and the comparatively low abundance of vital bacteria e.g. Nitrosomonadaceae. All but one plant was sampled once – the time series plant data is indicated.

6.2 Nitrification and the abundance of ammonia oxidising bacteria (AOB)

Single grab samples from full scale activated sludge wastewater treatment works (Table 4) were analysed using NGS and cell specific ammonia oxidation rate (CSAOR) calculated (Equation 5) using ammonia removal data collected contemporaneously as part of NWL's Urban Wastewater Treatment Directive sampling regime (see section 5.1). Figure 33 shows that AOB abundance can vary by 3 orders of magnitude, with an apparent CSAOR threshold of approximately 8, above which instability may be observed which is similar to previous findings (Pickering 2008; Coskuner et al. 2005).

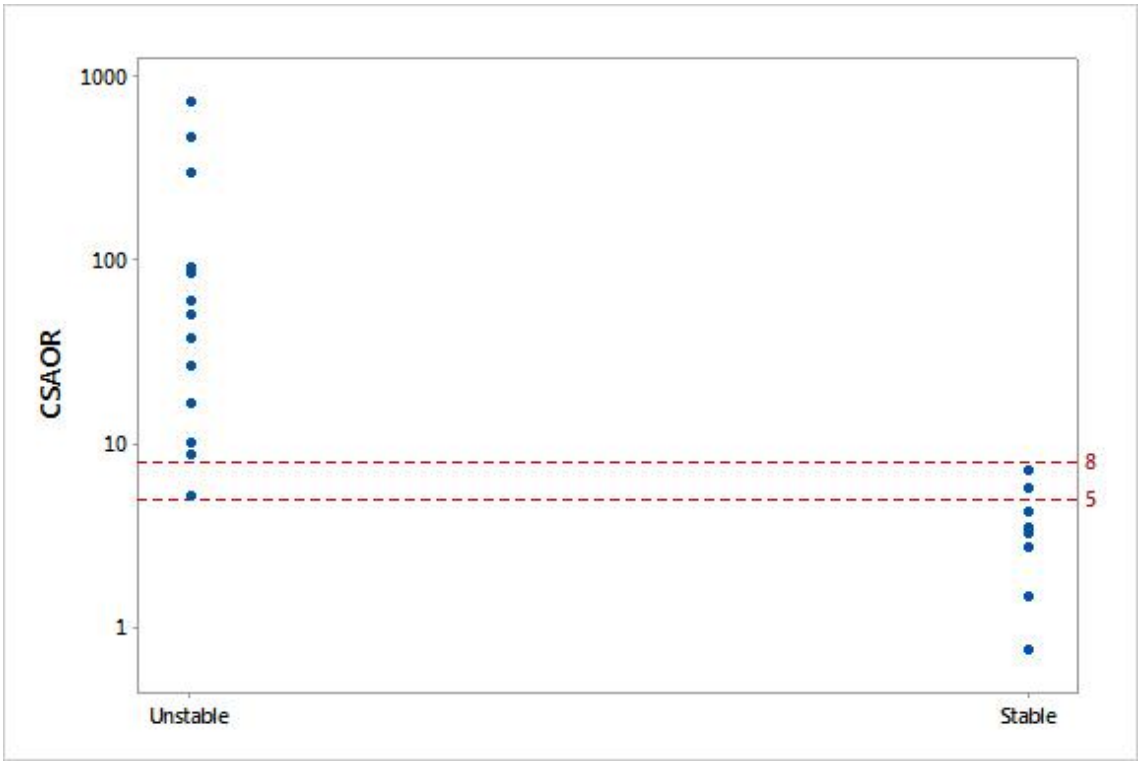


Figure 33: Cell Specific Ammonia Oxidation Rate (CSAOR) for 25 activated sludge plants. A plant was considered stable if it had not breached its ammonia effluent consent in the previous 12 months.

6.3 The abundance of foam causing bacteria and the incidence of foaming

When ASPs were sampled (see 5.1) their operational status according to the operator was noted, including whether they were currently (CF), occasionally (OF, foaming at least once in the 12 months prior to sampling) or non-foaming (NF, no foaming in the 12 months prior to sampling). After extracting the abundance of all foaming bacteria from the NGS data, the data sets were grouped according to their operational status. Foaming bacteria abundance was significantly higher in OF plants than NF plants (ANOVA $p < 0.001$, Tukey's test $p < 0.001$). Only one plant was sampled whilst

currently foaming and thus whilst the difference in foaming bacteria abundance was significant further investigation is required (Tukey's test $p < 0.001$). Overall the data suggests that NGS can be used to predict foaming.

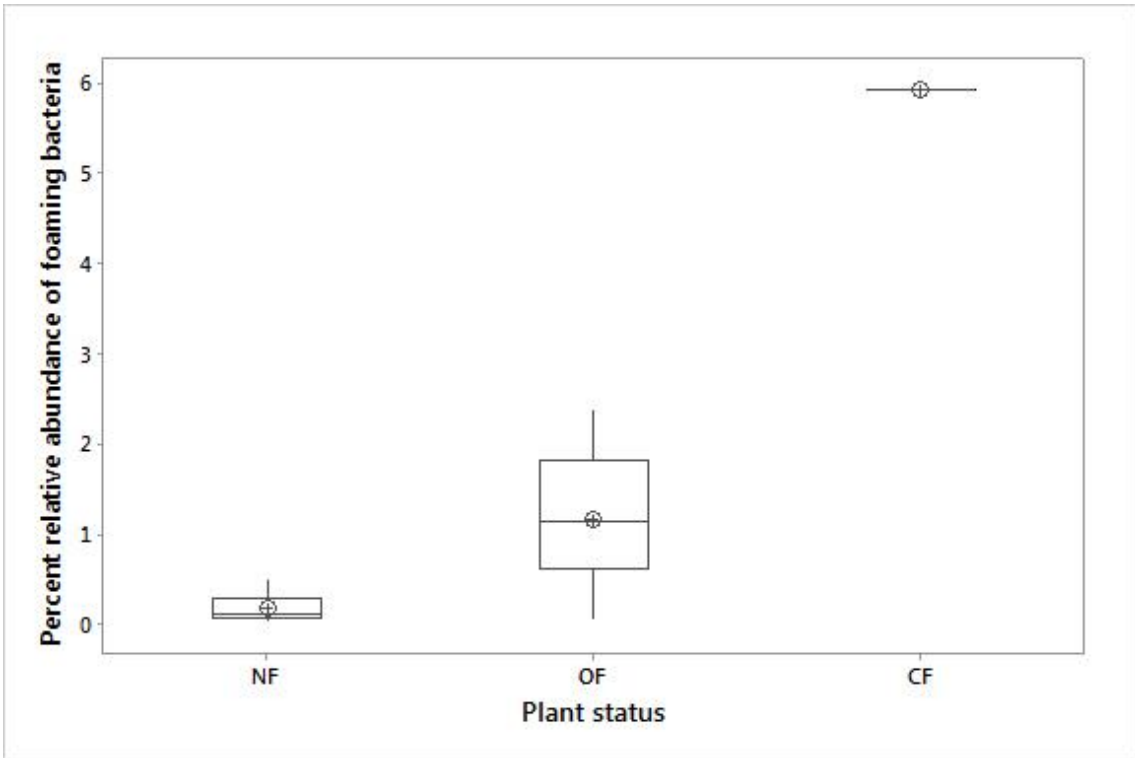


Figure 34: Relative abundance of foaming bacteria in activated sludge plants with samples grouped according to whether they do not foam (NF, n=20), occasionally foam (OF, n=20) or are currently foaming (CF, n=1). Note that only one plant was sampled when currently foaming.

6.4 The abundance of bacteria associated with bulking and the incidence of bulking.

When ASPs were sampled their bulking status according to the operator was noted, including whether they were currently (CB), occasionally (OB, plant has bulked at least once in the 12 months prior to sampling) or non-bulking (NB, plant has not bulked in the 12 months prior to sampling). Four samples were taken during confirmed bulking events with the remainder taken when floc was settling normally. After extracting the abundance of all bulking bacteria from the NGS data, the data sets were grouped according to their operational status. No significant different was found between samples grouped by operational status (ANOVA, $p = 0.335$). I suggest this may be due partially to a poor understanding of which taxa are implicated in bulking events, and partially due to poor recording of bulking incidents since plants may pass consent based on percent removal of solids regardless of whether the system is bulking.

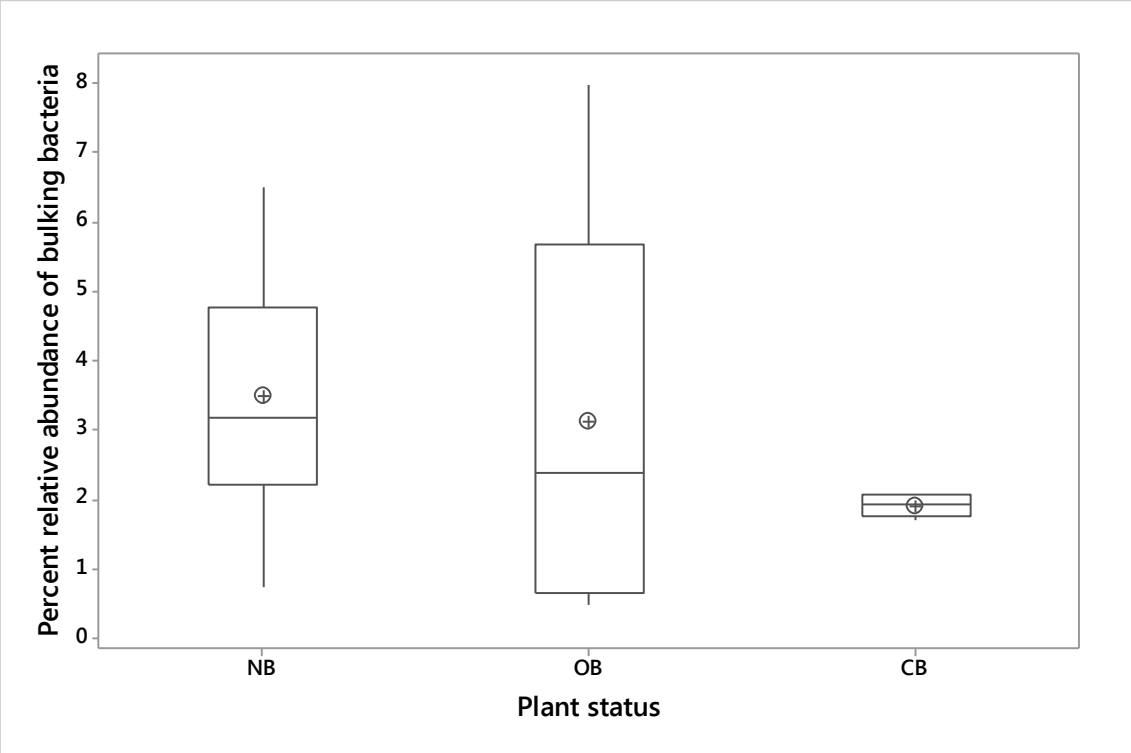


Figure 35: Boxplot showing the percent relative abundance of bulking bacteria from 32 activated sludge plant samples. Plants were grouped by operational status based on reports from plant operators. NB = Not bulking, OB = occasional bulking, CB = currently bulking.

Final effluent suspended solids may be an indicator of a bulking event due to poor floc entrainment in the secondary clarifier. An exploratory investigation of the correlation between bulking bacteria abundance in all ASP samples and final effluent suspended solids data collected from NWL SCADA systems showed found no significant correlation (Pearson correlation -0.136, p=0.423, Figure 35). Note that Stirred Sludge Volume Index was not available in the NWL system. Further investigation is required in order to understand whether taxa abundance can be used to predict bulking and if so which taxa comprise the bulking group (see 9.1).

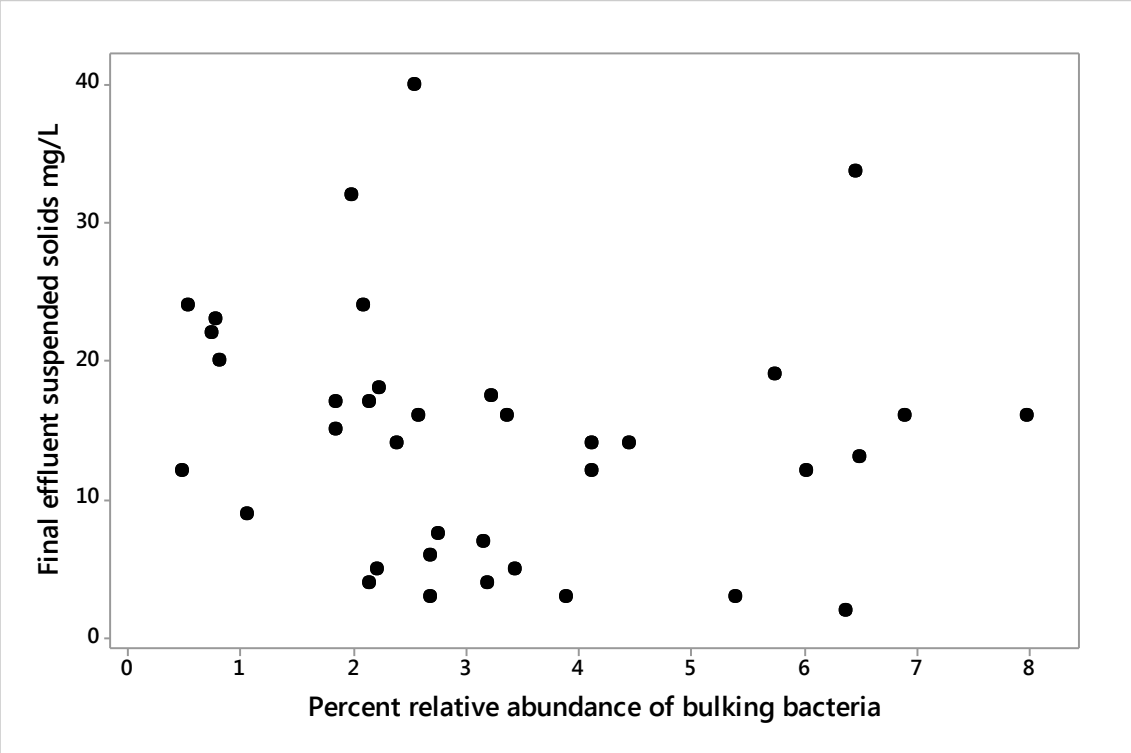


Figure 36: Percent relative abundance of bulking bacteria in samples from 32 activated sludge plants versus effluent suspended solids concentration in mg/L.

Chapter 7. Microbial source tracking results

7.1 Faecal indicator and host marker bacteria

Bathing water samples were collected at the same time and place as the Environment Agency (EA) regulatory samples, to enable interpretation of NGS results in the context of FIB abundance reported by the EA. In addition, 4 of these planned bathing water samples and 1 unplanned storm event sample were analysed by Northumbrian Water laboratories following the EA protocol. This enabled comparison of results using the same method between laboratories to investigate the precision of FIB enumeration by culturing.

7.1.1 Correlation of faecal host markers across the catchment with respect to bathing water faecal indicator bacteria

The abundance of faecal markers was converted into binary presence absence data as was EA Enterococci and *E. coli* abundance in order to test for correlation between detection of faecal contamination by PGM and by the EA. Comparison of the presence of any faecal marker in either Pit Hills Stell (sample point B2) or the Saltburn Gill / Skelton Beck confluence (sample point B11) found no significant correlation (Fisher's exact test, p=1). When considered individually, no significant correlation was found between Pit Hills Stell (Fisher's exact test, p=1) or Saltburn Gill / Skelton Beck confluence (Fisher's exact test, p=0.49) and EA data. This may highlight a lack of correlation between faecal marker abundance and faecal indicator bacteria abundance.

A number of sampling points were upstream of sewage treatment works (STWs) and therefore samples taken other than when CSOs were operating could provide an indication of environmental faecal sources. The Pit Hills Stell subcatchment is not impacted by any STWs and therefore samples from B2 "Pit Hills Stell outfall" would not be expected to contain human faecal markers other than under storm conditions. However on four occasions two or more human faecal markers were detected in non-storm samples (more than 48 hours after the most recent recorded CSO spill), indicating undocumented human faecal contamination. Similarly Saltburn Gill has no continuously discharging NWL assets yet harboured two or more faecal markers on 3 occasions. Enterococci were present in 11 of 15 Saltburn Gill non-storm samples and eight out of nine Pit Hills Stell samples. Four out of six samples from upstream of the allotments at Pit Hills Stell (B15) contained one or more human markers, yet there is

no sewage treatment works or CSO outfall upstream of this point. There are however a number of properties not connected to the sewage network and therefore septic tanks, which may explain the unexpected faecal contamination.

7.2 SourceTracker

The SourceTracker tool (Knights et al. 2011) was used to identify the proportion of bacteria in the bathing water from each source (Table 6). This tool utilises the relative abundance as well as presence or absence of all OTUs in a sample as opposed to individual taxa, obviating the need to discover unique host associated bacteria. This method also allows for the apportionment of faecal contamination to point sources, such as identifying the contribution to contamination from individual CSOs. There was a significant correlation ($p < 0.01$, Figure 37) between Enterococci abundance and the cumulative percentage of bacteria in the bathing water sample from faecal sources. However the majority of samples showed the presence of faecal signatures but no Enterococci, and vice versa. Further investigation showed that the residuals were normally distributed (Ryan-Joiner test, $p>0.1$), but there was an outlier (faecal ST = 55.25%, EA = 155, Dixon's r22 ratio, $p=0.007$). This outlier was not removed as it was taken after the 10th August storm event and featured significant CSO input and therefore would be expected to feature unusually high concentrations of CSO SourceTracker signature and EA Enterococci. In contrast, there was no significant relationship between *E. coli* abundance and the cumulative percentage of bacteria in the bathing water from faecal sources ($p=0.57$ Figure 38).

There was a significant correlation between EA Enterococci abundance and cumulative contribution to bacteria in the bathing water from faecal sources (partial least squares analysis, PLS; $p<0.001$, $r^2=87\%$, Figure 39, see Table 6 for faecal sources). There was no significant correlation between EA *E. coli* abundance total faecal contribution derived from Source Tracker (PLS; $P=0.574$, $r^2=45\%$, Figure 40).

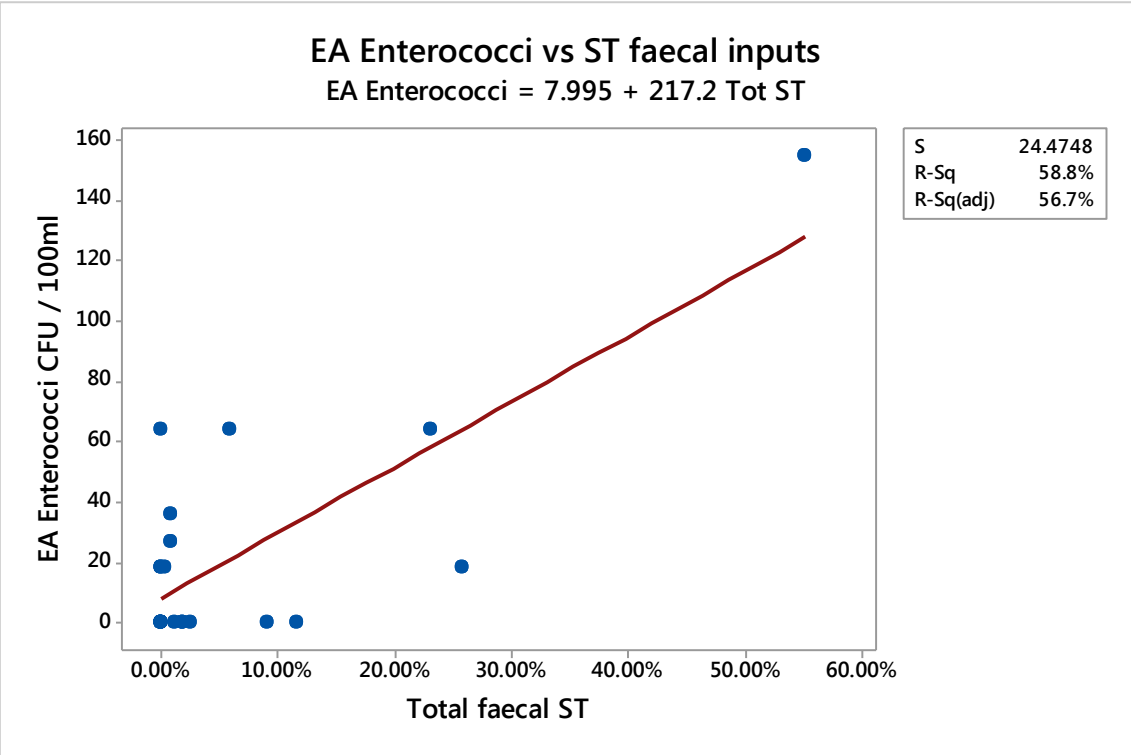


Figure 37: Linear regression analysis between number of Enterococci colony forming units (CFU) measured by the Environment Agency (EA) and total percentage of faecal source in sink determined by the SourceTracker tool. n = 23 samples

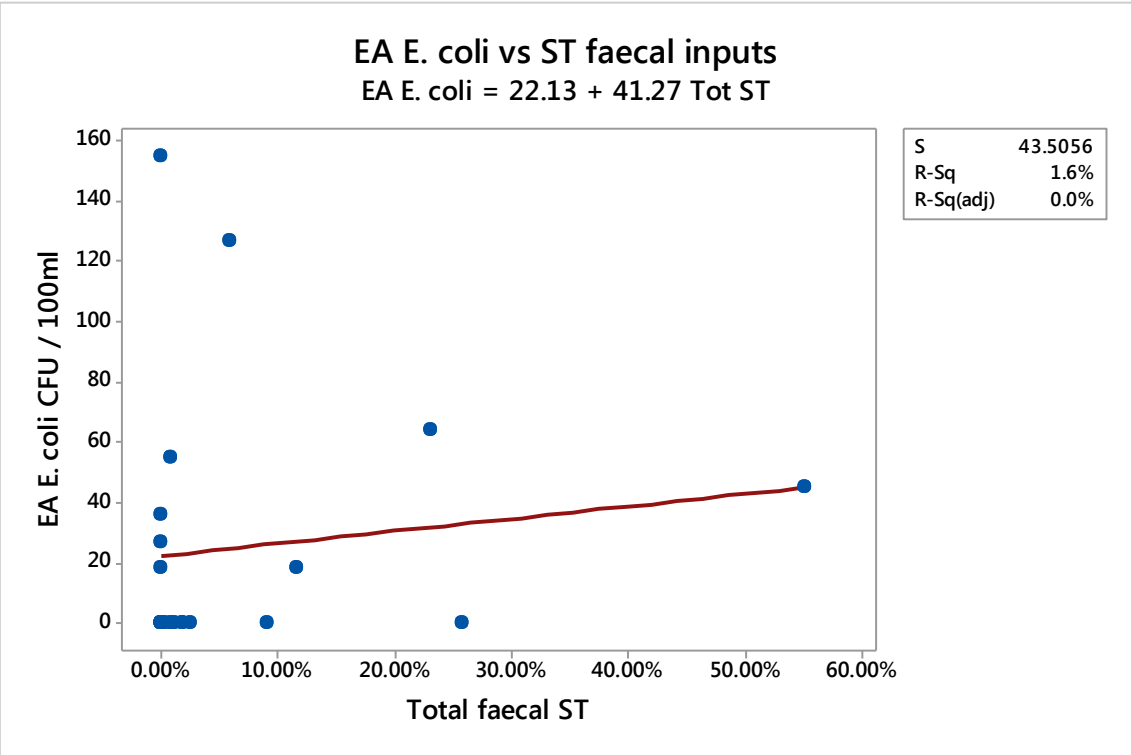


Figure 38: Linear regression analysis between number of *E. coli* colony forming units (CFU) measured by the Environment Agency (EA) and total percentage of faecal source in sink determined by the SourceTracker tool n = 23 samples

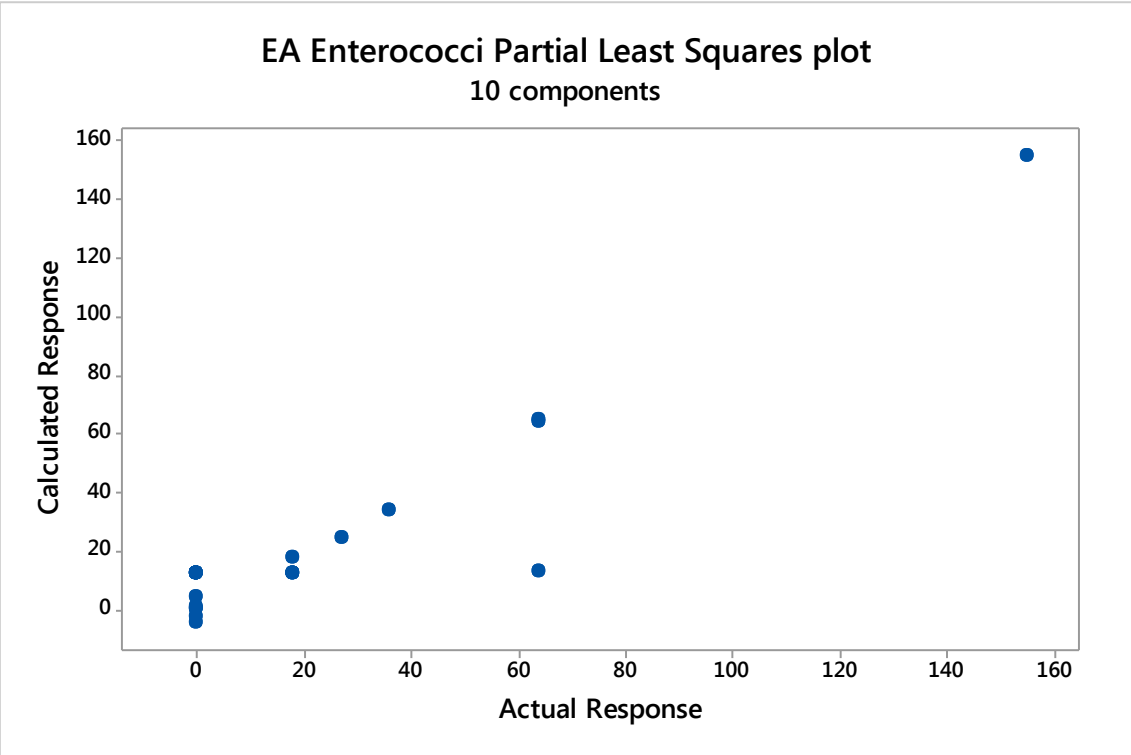


Figure 39: Partial least squares analysis between the number of Enterococci colony forming units measured by the Environment Agency (EA) and total percentage of faecal source in sink determined by the SourceTracker tool. n = 23 samples.

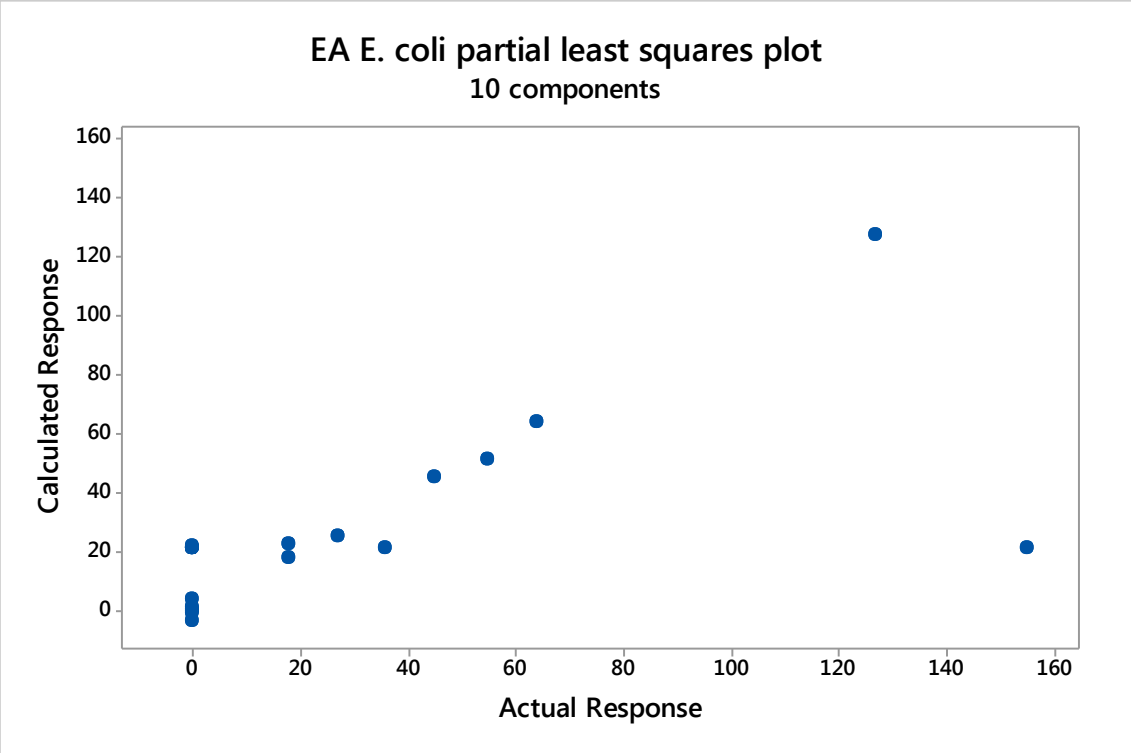


Figure 40: Partial least squares analysis between the number of E. coli colony forming units measured by the Environment Agency (EA) and total percentage of faecal source in sink determined by the SourceTracker tool. n = 23 samples.

7.3 Faecal contamination sources

In the 2014 bathing season, faecal pollution was detected by the EA in 14 of 20 sampling events. Of these, samples taken on June 5th, July 9th and August 13th were within 48 hours of CSO operation, and were therefore considered a storm impacted EA sampling event. The remaining 11 sampling events are considered non-storm samples, and therefore data for CSO faecal contribution was excluded to avoid skewing human faecal source data. The selection of a 48 hour time period to delineate storm impacted events is partially validated by the observation that the CSOs were in operation from 07:10 until 09:01 on July 8th, yet the EA sample taken at 10:00 on July 9th contained no *E. coli* and only 18 Enterococci. The June 5th sample contained 64 *E. coli* and 64 Enterococci however the SourceTracker data indicates significant gull faecal contamination in that sample, and the storm was ongoing with a further CSO spill on June 5th from 15:00 – 15:47.

Samples were taken during a storm on August 10th, which was not an EA sampling day. The nearest rain gauge is situated at Lockwood Beck Reservoir, approximately 1km to the west of the Saltburn catchment at 54°30'56.7"N, 0°57'58.4"W. This rain gauge recorded daily rainfall of 19.6mm on August 10th, the highest of the year (Met Office 2006). Table 13 shows the presence of host faecal markers and SourceTracker human versus non-human faecal signatures in samples taken on EA days as well as the number of *E. coli* and Enterococci detected by the EA.

There was no significant correlation between human and non-human faecal sources and EA Enterococci (PLS, p=0.70) or *E. coli* (PLS, p=0.93) in the non-storm data. There was perfect correlation between human and both *E. coli* and Enterococci in the storm data set (PLS, R^2=1). Standardised coefficients identified 5 important predictors; Brotton CSO (0.19), Caravan Park CSO (0.20), Leyland CSO (0.19), Skelton CSO (0.18) and Upleatham STW (0.19).

The data suggest that human faecal sources are the primary source of contamination only during storm conditions. Of all the host markers investigated, only 3 markers were detected in the bathing water samples, all of which were human markers. Of the six samples in which the EA did not detect faecal contamination, the PGM data found no markers in four samples. The PGM data showed human marker 5 but no

Enterococci on August 21st. On August 7th the PGM data showed human marker 1 and Enterococci (Table 14).

During the storm event of August 10th, NWL assets contributed 43.7% of bacteria during CSO flow, and only 0.02% before and 1.27% after. Surface waters contributed 11.6% of bacteria, with the remainder being from unknown, possibly marine sources. There was a small component of 0.02% from the Caravan Park CSO before the CSOs were flowing according to NWL systems (Figure 41). This may have been due to "trickling" of the CSO as the flow monitors require a minimum depth of water to trigger. The CSO is 600m from the bathing water so dilution and transit time are minimal. This may also be a false positive due to the low contribution (Henry et al. 2016). The only non-human faecal source detected during the storm was sheep faeces, contributing 0.03% during CSO flow. The contamination from each point source was notably reduced 6 hours after CSO flow ceased, though contamination from those CSOs further upstream remained present with 0.9% from Guisborough, 0.32% from Leyland and 0.02% from Skelton. This may be indicative of the transit time from these CSOs which are in the upper catchment. Gull faeces was also detected at 0.04% after the storm, though gull faeces was detected in 55% of bathing water samples and therefore this may not be related to the storm or CSO flow.

Table 13: Faecal sources in samples where the Environment Agency detected *E. coli* or Enterococci. EA = Environment Agency. ST = SourceTracker. ND = No data

Sample	EA Enterococci	EA *E. coli*	Human 1	Human 2	Human 4	ST % Human	ST % Non-human
May 8th	18	0	-	-	-	ND	ND
May 15th	18	27	-	-	-	ND	ND
May 21st	18	155	-	-	-	0	0
May 27th	27	55	Yes	-	-	0.95	0
June 5th	64	64	-	-	-	0	23.14
June 11th	0	27	-	-	-	0.01	0.02
June 17th	36	91	-	Yes	-	ND	ND
July 1st	0	36	-	-	-	0	0

Sample	EA Enterococci	EA *E. coli*	Human 1	Human 2	Human 4	ST % Human	ST % Non-human
July 9th	18	0	-	Yes	Yes	0.38	0
July 19th	173	27	-	-	-	ND	ND
July 25th	64	18	-	-	-	0.04	0
August 13th	155	45	Yes	Yes	-	51.9	0
August 28th	18	0	-	-	-	0.02	25.76
September 12th	0	18	-	-	-	2.42	9.28

Table 14: Faecal sources in samples where the Environment Agency did not detect *E. coli* or Enterococci. EA = Environment Agency. ST = SourceTracker. ND = No data

Sample	EA Ent.	EA *E. coli*	Human 1	Human 2	Human 4	Human 5	ST % Human	ST % Non-human
June 24th	0	0	-	-	-	-	0.01	0
July 31st	0	0	-	-	-	-	0	0.02
August 7th	0	0	Yes	-	-	-	0	1.21
August 21st	0	0	-	-	-	Yes	0	2.64
Sept 6th	0	0	-	-	-	-	0	9.1
Sept 18th	0	0	-	-	-	-	0.01	1.93

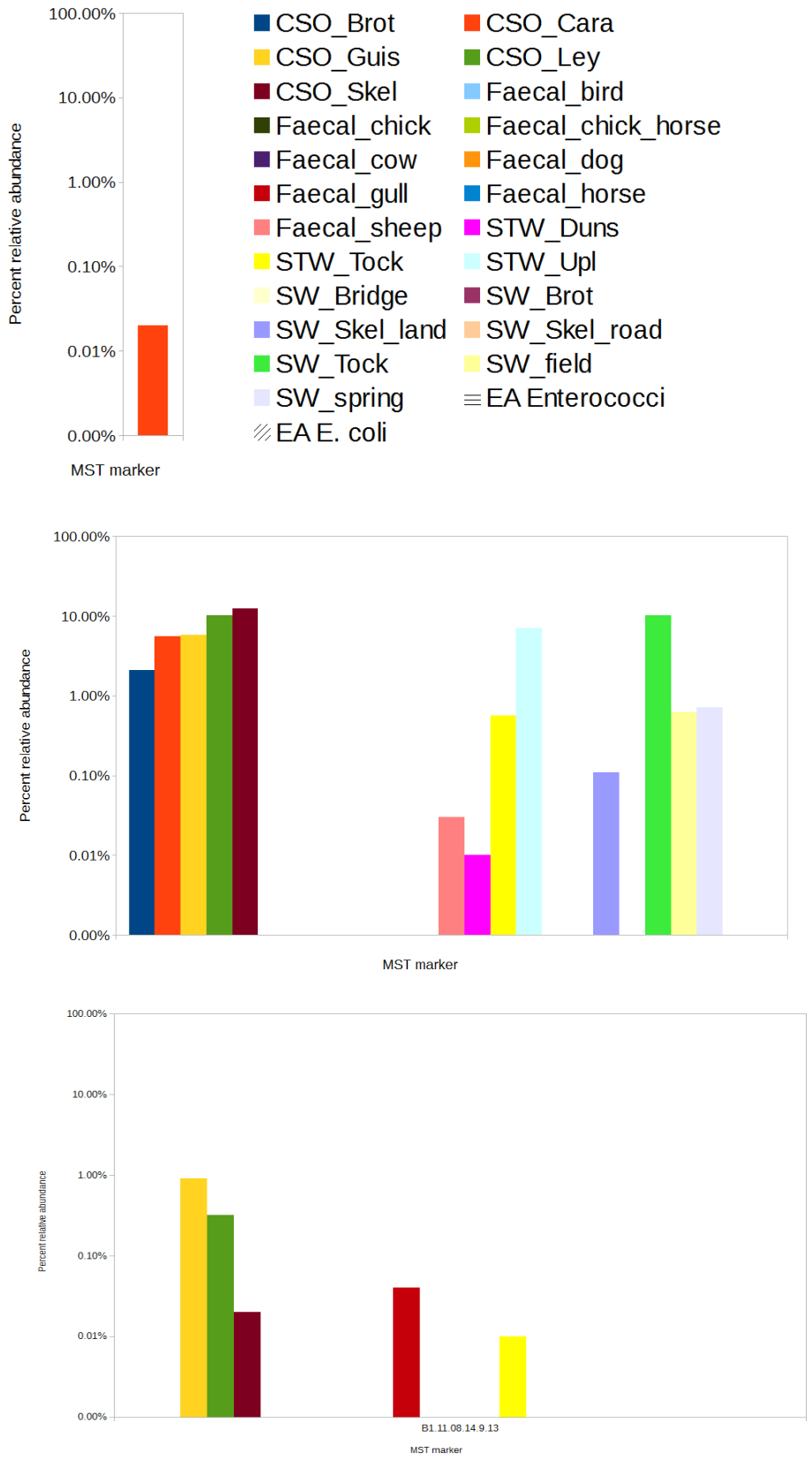


Figure 41:SourceTracker results; the relative contribution of microbial communities from different sources (identified in the key which reads left to right) to the bathing water (sink) microbial community at Saltburn: Top, during 10th August 2014 storm before combined sewer overflow (CSO) operation; middle, during storm and CSO operation; Bottom, on 11th August, 6 hours after CSO cessation. STW = Sewage Treatment Works, Surface = Surface water drains

Chapter 8. Discussion

8.1 Using NGS to diagnose and predict activated sludge operational issues

8.1.1 Nitrification monitoring by NGS AOB enumeration

It is vital that a plant nitrifies in a stable manner in order to protect the environment from the deleterious effects of excessive ammonia in sewage effluent. In terms of nitrification, a plant is considered stable where effluent ammonia concentration has not exceeded the limit set by the Environment Agency in the 12 months prior to sampling. A number of authors have previously described the relationship between the cell specific ammonia oxidation rate (CSAOR) and the stability of nitrification in an ASP using molecular methods such as FISH and qPCR (Coskuner et al. 2005; Pickering 2008; Eyice et al. 2014). However NGS only generates relative abundance data so the calculation of absolute number of cells per millilitre is impossible without additional measurements such as laborious total cell counts. In this study the cell specific oxidation rate was calculated using a proxy total AOB abundance; milligrams of ammonia oxidising bacteria (see section 5.6.5). CSAOR considers both the abundance of AOB and the ammonia load they are exposed to. It was estimated that CSAOR of 8 fmol/cell/hr is the threshold above which nitrification becomes unstable, which is in agreement with previous studies using FISH (Coskuner et al. 2005; Eyice et al. 2014). However it is not clear whether this threshold is universal to all ASPs because it is known that numerous parameters affect AOB metabolic rates, including temperature and nitrite concentration (Koops et al. 2006).

The agreement in ideal CSAOR between FISH and the present study suggests that NGS is at least equal to FISH at monitoring AOB abundance with the advantage of higher throughput and greater precision (see 5.9.3). This is relevant as the routine monitoring of ASP function necessitates the processing of a large number of samples in a timely manner. For example, NWL has 26 ASPs which assuming bi-daily sampling would equate to 78 samples per week. Running a single sample through the FISH protocols can take up to 2 days (Kao et al. 2015), whereas NGS could process all 78 samples in the same time period. In addition the data analysis can be automated with sufficient coding and integration with sequencing instruments, which may make ASP nitrification monitoring by a central water company laboratory feasible. Additional benefits from NGS over FISH may be gained from the ability to differentiate between species present

as opposed to using a general AOB FISH probe. Given the different metabolic rates and substrates utilised by different species this data may prove relevant in determining the resilience of an ASP to fluctuations in temperature or other environmental factors.

NGS monitoring of ASPs could bring about environmental, financial and public relations benefits. For example, with regular sampling nitrification failure could be predicted and circumvented either by adjusting plant operational parameters or instigating emergency ammonia removal measures such as ammonia stripping. This would prevent the discharge of excess ammonia to the receiving water body thereby protecting its ecology. Pollution incidents with concordant fines up to £1m, negative press and complaints by members of public regarding dead fish and other deleterious effects of eutrophication would be reduced. There are also opportunities to reduce energy consumption due to over aeration. The data in section 6.2 suggests that plants with CSAOR lower than 8 are stable, raising the question at what point is CSAOR so low that the ASP is operating inefficiently? Dissolved oxygen partially determines AOB respiration rates (Kowalchuk & Stephen 2001; Limpiyakorn et al. 2005; Park & Noguera 2004) and therefore the growth and ammonia oxidation rate per AOB cell (see Equation 3 and Equation 4). One reason for low CSAOR could be excess dissolved oxygen due to over aeration resulting in excess AOB abundance. It has been established that nitrification can be achieved at low dissolved oxygen levels (Park & Noguera 2004; Gieseke et al. 2003) and therefore optimisation of aeration by CSAOR measurement may be useful (see section 9.1). As described in section 3.1 energy usage is a significant proportion of the expense and carbon footprint associated with activated sludge, and so efforts to reduce excess aeration could bring significant financial and environmental benefits.

8.1.2 Prediction of ASP foaming by NGS

Foaming in ASPs is the occurrence of a stable biological foam which covers some or all of the surface of the aeration basin. The phenomenon is problematic since the foam can carry over in to the final effluent, increasing effluent solids and visual impact, and preventing efficient control of solids retention time by escaping the sludge wastage mechanism. Foaming requires the presence of surfactants, hydrophobic particles, solids and gas bubbles. The bubbles are necessarily present due to aeration, whereas the hydrophobic particles are often present due to over growth of filamentous

bacteria. A proportion of surfactants present may be from extra-cellular polymeric substances (EPS) produced by bacteria under stress. It was hypothesised that foaming incidents only occur when the cumulative abundance of foaming taxa (Table 9) is greater than a fixed threshold (Davenport et al. 2000; Davenport et al. 2008; Narayanan et al. 2010).

The foaming results in section 6.3 suggest that a significant difference in the abundance of foaming bacteria exists between non-foaming and foaming plants, which is in line with the findings of Davenport et al (2000). It is tentatively suggested that a plant is at risk of foaming where total relative abundance of foaming bacteria is between 1 and 2%, with foaming expected at abundances greater than 2%. There is some precedent for this low threshold as Martin et al (2004) suggest filament abundance above 1% can cause bulking. Further analysis is required to validate this threshold, but if proven accurate this enables proactive management of foaming with a sufficient sampling regime and subsequent to feasibility of laboratory arrangements as discussed in section 8.1.1. Few studies have attempted to quantify the abundance of foaming bacteria in foaming ASPs using molecular methods. Marrengene et al (2011) used qPCR to quantify *Gordonia* spp in foaming plants and found that all samples had >10% *Gordonia* abundance. However this was established by combining PCR products from universal primers and Gordonia specific primers at different ratios, which fails to take account of primer non-specificity and 16S rRNA gene copy number variation. Petrovski et al (2011) concluded that a universal threshold concept does not hold due to the addition of surfactants reducing the abundance of filaments required to induce foaming. What is not clear is the role that surfactants present in the influent have (Narayanan et al. 2010; Chauhan & Sharma 2014), as well as those produced by bacteria in the mixed liquor such as Bacillus subtilis (Petrovski et al. 2011; Blackall et al. 1991; Soddell & Seviour 1990). See section 9.2 for suggested further investigation.

Foaming is a global problem (Marrengane et al. 2011; Naidoo et al. 2011; Davenport et al. 2008; Guo & Zhang 2012) which influences the operating cost and environmental impact of ASPs (Blackall et al. 1985). Therefore the development of a new tool which can accurately diagnose or even predict foaming could save the industry and thus the consumer money, as well as reducing environmental impact. It is difficult to quantify

the cost of foaming events since relevant data such as blower energy consumption is neither fully utilised by water companies nor publically available. Similarly foaming events are not routinely catalogued so the number of events per year cannot be quantified. Whilst the impact of foaming on operating costs is not clear, the impact of a pollution event is detailed in the magistrate's guidance for environmental offences (Sentencing Council 2014). The range of fines for a negligent category 3 incident is £35,000 to £150,000 where category 3 is defined as "Minor, localised adverse effect or damage to air or water quality, amenity value, or property". There may also be further financial penalties imposed by the regulator OFWAT if the utility exceeds its target number of pollution incidents. The impact of foam on the environment would be specific to each case, however could include localised death of fish and other fauna as well as the additional carbon footprint from increased aeration. Where the foaming event severely impacts control of sludge age, this may have secondary impacts on effluent quality parameters such as biological oxygen demand and ammonia concentration. Therefore there are clear benefits from predicting and preventing foaming in ASPs.

8.1.3 Prediction of ASP bulking by NGS

Bulking in ASPs occurs when the floc does not settle within the secondary clarifier, leading to excessive sludge blanket depth and carry over of floc in to the final effluent (Sezgin et al. 1978). The subsequent poor control of sludge age and high final effluent solids content can result in breach of effluent consents, deleterious effects on the environment and subsequent fines and PR issues (Jenkins et al. 2003). However the problem is poorly understood from a biological perspective and thought to result from overgrowth of filamentous bacteria (Martins et al. 2004). As the length of filaments in the floc increase the settling velocity of the floc decreases (Martins et al. 2004), and therefore all taxa with filamentous morphology could induce this phenomenon. Some commonly observed taxa in bulking events are listed in Table 9 and the abundance of this group is analysed below.

Much like ASP foaming, bulking is a global issue the impact of which is unclear. One important difference is that foaming events are unlikely to impact the final effluent solids concentration assuming the foam does not overtop scum boards in the final clarifier. This is not the case for bulking since the floc not as buoyant as foam and is

carried with the flow of water (Jenkins et al. 2003). Therefore the likelihood of a bulking incident causing a consent failure or pollution incident is higher. The relative impacts of foam reducing dissolved oxygen in rivers by smothering the water surface and increasing biological demand as opposed to bulking related increase in biological oxygen demand and solids content has not been quantified. However given the increased likelihood of environmental impact and subsequent prosecution, it follows that developing techniques to predict bulking is worthwhile.

It was hypothesised that bulking incidents would only occur where bulking bacteria abundance is greater than a fixed threshold (Asvapathanagul et al. 2015). Unfortunately the results in section 6.4 indicate there is no significant correlation between the abundance of bulking taxa and the occurrence of bulking. This may be due to insufficient sampling, with only one sample taken during an actual bulking episode. This may also be due to the different mechanisms involved in bulking versus foaming. The abundance of mycolata in foaming directly relates to ASP propensity to foam as the number of hydrophobic particles in solution directly impacts micelle formation and stability (Davenport et al. 2008). However in bulking it is thought that the phenomenon is caused by the length of filaments in flocs as opposed to their abundance (Jenkins et al. 2003). It may be that bulking cannot be predicted by abundance alone if the causative factor is filament length, however further NGS surveys of bulking events will shed light on this (see 9.1)

8.2 NGS for microbial source tracking

8.2.1 What does the study tell us about the function of the catchment during dry and wet weather

Overall the data suggests that the dominant contributor to FIB during storm conditions was NWL infrastructure (CSOs and STWs), whilst under dry conditions they are at most equal contributors to faecal contamination, and often contribute significantly less than animal sources.

There are several confounding factors that may contribute to the poor correlation between SourceTracker results and EA FIB results which reflects previous studies (Ahmed et al. 2015). These include SourceTracker apportionment of total bacterial community as opposed to direct apportionment of FIB. For example, 20% of bacteria in the bathing water sample may have arisen from surface water sources and only 2%

from STWs, yet the STWs will presumably contain a significantly greater proportion of FIB than the surface water. This is highlighted by the poor correlation between SourceTracker faecal source appointment and FIB by linear regression in contrast with the PLS regression which weights the sources according to their impact on FIB abundance. In addition, under sampling of faecal sources may contribute to failure to assign OTUs to their faecal source due to high variation in host microbial communities.

In addition, the way in which CSOs are monitored and managed posed additional challenges. It was established during the study that at least one CSO was not triggering an alarm as designed, and another CSO was not triggering the alarm at low flows. The data from NWL indicated only that a CSO somewhere in the catchment was flowing. However, little is known about the catchment hydrology and the transit time of flow from CSOs at different points within the catchment. It may be that for some storm impacted samples only CSOs nearest the bathing water operated whilst for others those up to 25 km upstream operated, which would create a significant difference in time between CSO operation and effect on bathing water quality. Microbial community decay may be significantly greater from CSOs far inland, depending on total riverine flow rates and solar exposure (Whitman et al. 2004; Boehm 2007). Whilst a number of studies have investigated the decay of individual faecal host markers (Bae 2009; Bae & Wuertz 2015; Brown & Boehm 2015; Marti et al. 2011; Solecki et al. 2011), none have investigated the decay of a whole community signature such as that from a STW.

8.2.2 The impact of NWL water company assets versus other human and non-human sources

It is important to note that human faecal contamination from sources other than NWL assets also influenced bathing water quality. Saltburn Gill harboured human faecal markers when no NWL discharge was documented (see 7.1.1). Similarly samples from Pit Hills Stell upstream of allotments (B15) also contained human faecal markers despite there being no NWL assets upstream of the sample point. Investigation of selected isolated properties adjacent to the watercourses show that they are not connected to the sewer system. Such contamination may therefore arise from septic tanks or similar, suggesting that diffuse human faecal contamination is present. However, it is also evident from the faecal marker studies that non-human faeces were

present in raw sewage which is presumed to arise from surface run off in to combined sewer systems (Table 11). Therefore both human and non-human faecal contamination can arise from both diffuse and point sources, confusing the delineation of NWL versus non-NWL contamination. The SourceTracker signatures from both diffuse and point sources are unaffected by this issue since all OTUs present are taken in to account as opposed to individual markers. Therefore the assessment of NWL versus non-NWL contribution to faecal contamination is based on the SourceTracker results.

It is clear that during storm conditions which result in CSO operation that the CSOs are the primary contributors of bacteria from the faecal sources assessed in this study. However SourceTracker does not provide source apportionment at the individual OTU level, and even if it did the taxa which grow on regulatory agar plates are not yet understood. Therefore the SourceTracker results at present provide an overview of the source of bacteria in the community as opposed to predicting the source of *E. coli* or Enterococci. Bathing water samples from May 21st and July 25th contained high levels of FIB yet SourceTracker did not identify any responsible source for May 21st and only 0.04% human sources for July 25th. This suggests that faecal sources in the catchment may be under sampled. One cause may be the lack of a pig faecal signature which are known to inhabit the catchment. Where samples show low abundance of FIB and no SourceTracker sources it may be that the FIB present are of environmental as opposed to faecal origin. This has been shown for both Enterococci (Byappanahalli et al. 2012) and *E. coli* (Ishii & Sadowsky 2008; van Elsas et al. 2011).

Overall the data suggest that NWL assets are at most equal contributors to faecal contamination under non-storm conditions, and often contribute significantly less than animal sources.

Chapter 9. Further Work

9.1 Further NGS method optimisation

The DNA extraction experiment (section 5.3) found that a community structure shift is observed where the DNA binding matrix is incubated at 55 degrees before elution (method 3). On the assumption that DNA which is released with no incubation is not destroyed by incubation at 55 degrees, it can be concluded that method 3 provides the most representative DNA extract of the investigated methods. However whether method 3 produces truly representative DNA extracts cannot be commented upon in the absence of further experimental data. In section 5.3 I postulated that the cause of this putative differential bond strength may be due to varied GC content in the DNA strands. One such study might be the extraction of a mock community of known relative abundance in which the taxa feature a range of GC content from low to high. A range of incubation temperatures could be applied to the bound DNA to quantify the impact of different DNA binding strengths and incubation temperatures on observed community structure.

One of the challenges when comparing enumeration by multiple methods is that the values obtained may be biased by experimental design as opposed to the method itself. For example, if the probes used in a FISH analysis do not capture exactly the same taxa as those used in an NGS analysis then any attempt to correlate their findings is confounded by the fact that they are not necessarily measuring the same thing. Therefore, in order to compare enumeration of AOB or any other group by FISH, NGS and qPCR, the experiment must be carefully designed such that NGS primers, qPCR primers and FISH probes all target the same taxa. By designing such an experiment and analysing samples with a wide range of AOB abundance, it would be possible to gain a greater understanding of each method's precision and accuracy. Undertaking such a study may be key in gaining wider acceptance of NGS ASP monitoring.

9.2 NGS for ASP control

In order to establish whether NGS can predict nitrification failure, we must first establish both the timescale and the manner in which the ASP microbial community changes prior to failure. This can only be achieved by regular sampling of an ASP in anticipation of failure. The time this would take cannot be predicted or mitigated, but costs could be limited by collecting and preserving samples so that only those in e.g.

the 30 days prior to failure are analysed. Once this data has been collected for multiple ASPs and the predictive microbial community changes identified, the method must then be applied in a proactive manner to predict failure. Once this is achieved, the predictive capabilities of NGS ASP monitoring can be demonstrated and the wider acceptance of the approach achieved. It can be presumed that the metric for this predictive capability is the CSAOR. Whilst it is important to establish the upper threshold beyond which failure may occur, it would also be useful to establish the lower threshold below which the plant may be operating inefficiently. This would create a tool that not only troubleshoots but optimises and therefore create more value to the user per sample.

A similar approach is required for both foaming and bulking, given that the bacterial abundance thresholds above which foaming or bulking will occur remains uncertain. Regular sampling in anticipation of a foaming and bulking event as described for nitrification failure will enable the validation of my proposed thresholds and demonstrate their predictive capability. However a further complication with NGS monitoring of bulking is the uncertainty that surrounds the taxa involved. NGS cannot be used to analyse the abundance of a group whose taxa are not defined, and therefore further studies investigating the taxa which are involved in ASP bulking are required. NGS is uniquely able to correlate the abundance of all OTUs detected with sample categories such as foaming and non-foaming, which may be of great assistance in uncovering the causative taxa in bulking. In addition, cutting edge methods such as micromanipulation and DNA extraction from single cells may shed light on bulking taxonomy.

It is widely accepted that ASPs as currently operated in the UK cannot achieve sufficient removal of pollutants such as phosphorus to meet proposed consents (Rostron & Holt 2017). At present a typical phosphorus effluent consent is 1-2 mg/L (EEC Council 1991) which is often achieved by chemical coagulation via ferric chloride dosing (Metcalf & Eddy. et al. 2003). However phosphorus consents as low as 0.2 mg/L are being proposed in the UK in order to meet the demands of the Water Framework Directive, along with newly created iron limits (European Parliament and Council 2000). Options under investigation to meet this requirement include ion exchange resins and magnetite ballast systems, which are energy intensive and require

expensive chemicals. The proven enhanced biological phosphorus removal (EBPR) process is ideal for this scenario, requiring no chemical input and providing valuable phosphorus rich sludge which can be incorporated in fertiliser products (de-Bashan & Bashan 2004). However EBPR is known to fail unpredictably (Seviour et al. 2010) much like nitrification and therefore may benefit from monitoring by NGS (Crocetti et al. 2000). An example would be the monitoring of *Accumulibacter* abundance, a polyphosphate accumulating organism (PAO) which sequesters polyphosphate within the cell and thereby facilitates the removal of phosphorus through activated sludge wastage (Crocetti et al. 2000). One possible benefit from NGS monitoring of ASPs is that all functional groups – AOB, NOB, foamers, PAO - can be monitored by a single analysis of one sample, as opposed to e.g. multiple FISH analyses requiring various probes and sample preps.

Another emerging area of research is the removal of micropollutants by ASPs. At present it is not known which chemicals will be regulated in wastewater effluents, and a UK wide investigation program is under way to identify the current concentration of priority substances listed by the EU (European Parliament and Council 2012). It is thought that current sewage treatment processes will not be sufficient to meet regulatory consents for priority substances though data is lacking in this area. However it has been shown that bacteria in ASPs can degrade a wide range of compounds including the herbicide Anilofos (Zhang et al. 2011), oestrone E1 (Zang et al. 2008) and pharmaceuticals such as atenolol (Helbling et al. 2012) and ibuprofen (Pomiès et al. 2015). NGS can be used to survey plants for bacteria capable metabolising micropollutants, and even predict the total metabolic capability of the community present (Ye et al. 2012). NGS has a vital part to play in improving our understanding and operational control of NGS to meet present and future challenges in wastewater treatment.

9.3 NGS for microbial source tracking

If SourceTracker is to be used to apportion specific taxa such as *E. coli* or Enterococci, the exact taxonomy as established by 16S rRNA gene sequencing of those bacteria cultured by the standard methods must be established. This study would enable to accurate extraction of relevant taxa from SourceTracker results, and may also provide further insight in to the public health risk associated with culture based results. For

example where a growth media is thought selective for *E. coli* but in fact harbours other taxa such as *E. hermannii*, the results from CFU counts must be treated with caution in terms of planning mitigation measures.

It is also unclear to what extent the abundance of FIB varies spatially and temporally. There is evidence that a single sample from a bathing water is not sufficient to classify the public health risk from swimming in it (European Parliament and Council 2012), and the results in this study have shown that FIB abundance in a sample taken a the same time and place can vary by an order of magnitude (section 7.1). Therefore further work to establish the number of replicates both at a single sampling event and also across numerous events required to accurately characterise the bathing water are essential. Finally, there is a growing body of evidence that FIB can grow outside the intestine and therefore a method of distinguishing "environmental" from "faecal" FIB is required if the true risk to public health is to be established.

A continued effort to identify 100% specific and sensitive markers is warranted given that their detection enables a survey of faecal contamination at minimal cost and sampling effort and enables a targeted sampling regime for additional SourceTracker analysis of the catchment. This work should include establishing the abundance of markers in faeces in comparison to FIB, since the abundance of markers should be at least equal to FIB in order to ensure that detection of one infers presence of the other. Furthermore detection of individual taxa may be possible at near real time in the field using methods such as LAMP. Therefore new faecal host markers may provide for real time protection of public health, alerting bathers to current water conditions rather than simply providing a broad brush historical assessment. The abundance of markers must also be established in order to design NGS MST studies with a lower limit of detection sufficient to capture the markers. For example in this study the cow marker was not detected in the cow faecal sample, suggesting it is of relatively low abundance. This data is vital in order to limit false negatives in faecal marker studies.

The ability of SourceTracker to distinguish sources quantitatively has yet to be established. Further experiments consisting of artificially constructed "pooled" samples from a number of point sources would elucidate point source resolution. Henry et al (2016) constructed artificial pooled samples for human and duck faeces,

but did not consider multiple point sources such as artificially mixing three CSO samples. The inclusion of the Henry et al Relative Standard Deviation approach may also assist in excluding false positives. An important component of this study could be the measurement of community signature decay. For example it is not clear what impact exposure to UV and oxic conditions may have on the microbial community signature from a CSO far upstream of the bathing water.

Chapter 10. Business case

In order to assess the business case for the use of NGS in the water industry, the cost of analysis per sample must be compared to potential savings and other indirect benefits, such as positive public relations and the protection of coastal economies. As the applications of NGS described in the preceding case studies are novel, cost estimates must be developed based upon existing commercial services.

10.1 NGS monitoring of ASPs

Activated sludge process monitoring and troubleshooting using NGS uses the same sample processing and initial data analysis steps as other NGS applications. The commercial cost per sample for DNA extraction, PCR, sequencing and basic bioinformatics is approximately £211 ex VAT (GATC Biotech, Konstanz Germany) when processing batches of 96 samples. It should be noted that costs increase at lower throughput due to fixed reagent costs per sequencing run. Additional bioinformatics is required to extract operationally meaningful data from the sequencing output, such as cell specific ammonia oxidation rates (CSAOR). Assuming an average of one hour per sample, this would cost an additional £65 per sample (Bioinformatics Support Unit, Newcastle UK), giving a total analysis cost of £276 per sample.

Northumbrian Water has 26 activated sludge plants (excepting small rural descriptive works), of which nine are nitrifying plants. As no one has attempted to predict nitrification failure in ASPs using molecular microbial methods and in particular CSAOR, it is difficult to determine a suitable sampling regime for the purposes of this business case. Data from Graham et al (2007) shows that a decline in AOB and NOB abundance may occur over as little as 5 days with a concomitant increase in ammonia in the effluent. Therefore I will assume that samples are required every 48 hours to predict nitrification failure, though that is subject to verification in the field. For the nine NWL ASPs, monitoring costs would therefore be £ 453,330 per annum, or £ 50,370 per ASP monitored. These costs do not take account of developing the relevant bioinformatics pipelines or databases.

Utilising NGS to monitor CSAOR could reduce NWL's aeration costs by up to £0.6m as suggested in section 3.1. If CSAOR monitoring is shown to predict nitrification failure, we can also consider the potential savings in terms of fines avoided. Whilst historically such a consent failure has not led to financial penalties, changes in the approach of the

courts in 2014 has led to a significant increase in potential fines (Sentencing Council 2014). These new guidelines have only recently come in to force and to date only one ammonia related fine has been issued to South West Water in the amount of £142k. Therefore if one fine per year were avoided and the potential aeration savings realised the instigation of CSAOR monitoring would bring about a profit of approximately £288,670 per annum. This does not take account of intangibles such as the PR benefits from reduced CO2 emissions, better river status due to lower ammonia load and avoidance of negative news stories from fines.

The Water Framework Directive requires that all rivers be of "Good" status by 2020 (European Parliament and Council 2000). One determinand receiving new scrutiny is phosphorus, with the UK government suggesting that sewage effluent consent limits as low as 0.1 mg/L may be required (Parliamentary Office of Science and Technology 2014). A number of expensive, energy intensive or chemical consuming technologies are being investigated in order to achieve this (Severn Trent 2016). However enhanced biological phosphorus removal is a well-studied form of the activated sludge process, and by utilising this approach it is possible that savings could be achieved by reducing chemical usage. In countries where EBPR is well established such as Denmark, NGS is already used to monitor and troubleshoot reactors (Nielsen 2017). Per Nielsen and his group in Aalborg University has already seen success in this area, having published a number of papers discussing the microbial community structure in EBPR plants (Mielczarek et al. 2012; Cokro et al. 2017; McIlroy et al. 2014) and using NGS to identify novel taxa which form the PAO functional group (Kristiansen et al. 2012). Therefore NGS analysis may assist in the design and optimisation of EBPR plants in the UK.

Whilst it is not clear that NGS can directly assist the UK water industry in avoiding bulking episodes, this study has produced evidence that foaming may be predicted using NGS. Data concerning industry prosecutions is scant, with the majority of information available from news publications and thereby focussing mostly on large fines. Instead I have used the Magistrate's Sentencing Guidelines (Sentencing Council 2014) for the offence of "Organisations: Unauthorised or harmful deposit, treatment or disposal etc. of waste / Illegal discharges to air, land and water" to determine likely fines. I have supposed that the offence would either be deemed Negligent and given

that foam entering a watercourse is not as deleterious as raw sewage entering the water course have presumed a Category 3 offence. This is based on R. v Thames Water October 2014 in which Thames Water was fined £250,000 for allowing untreated sewage to enter a brook through a nature reserve. The judge remarked that this case was at the higher end of a Negligent Category 3 case. The fine range for such a case against a "Large" company with turnover of £50m p.a. is £35,000 to £150,000. As Thames Water's turnover is in the regions of £2bn the judge presumable increased the fine due to the severity of the case. Therefore I estimate that a foaming incident fine would be around £100,000 for a company the size of NWL with turnover of £730m p.a.

This study was not able to identify the frequency of sampling required to predict a foaming event, and a literature search highlighted that no author has yet achieved this using any enumeration method. Therefore it is difficult to assess the economic benefits of monitoring ASPs for foaming. Effectively this could be done for "free" when the ASP is being monitored for CSAOR or EPBR since sequencing can identify all bacteria in a sample. 48 hour sampling and analysis of NWL's remaining 16 ASPs would cost £805,920 p.a., equating to approximately 8 foaming related fines avoided per year. This is unlikely to occur given the sparsity of foaming related fines at present. In addition, these non-nitrifying plants are at lower risk of foaming due to lower sludge age than nitrifying plants. In combination these factors do not suggest that NGS monitoring of foaming in non-nitrifying ASPs would be cost effective.

10.2 NGS for microbial source tracking

Due to pressure from the public as well as regulators, wastewater companies are expected to invest in their infrastructure in order to improve bathing water quality. These interventions can take numerous forms, from storm tanks to reduce the number of CSO spills to ultra-violet disinfection of final effluent. Such measures can be costly both in terms of capital and operational expenditure. For example, United Utilities estimated total life costs of £17m for storm tanks and £15.9m for UV disinfection, with both estimates based on reducing the impact of storm flows from an 8,800 population equivalent sewage treatment works (Palfrey 2015). Therefore it is important to quantify the impact these interventions will have on bathing water quality. However to date there has been no proven method for differentiating between human and non-human faecal contamination, and no attempt to quantify the proportion of faecal

contamination from individual point sources such as CSOs. Instead the industry relies upon hydrological modelling of catchments, with assumptions of microbial quality of sources, microbial decay and hydraulic transport processes. This leads to a "wait and see" approach where the intervention is designed based upon the available knowledge and the final impact on bathing water quality is uncertain. Microbial source tracking (MST) by NGS can accurately apportion the contribution to faecal indicator bacteria from individual point sources as well as the overall contribution from diffuse sources, enabling direct assessment of the potential impact from proposed interventions. However, data is not available on individual project spend versus bathing water quality improvement, so the potential savings cannot be quantified. Therefore I compared the cost of NGS microbial source tracking in a bathing water catchment versus hydraulic modelling.

Given that the bathing water is classified on the basis of samples taken and paid for by the Environment Agency, it is logical when investigating the cause of poor bathing water quality to take investigatory samples contemporaneously with the Environment Agency samples. In this manner, the findings of a project investigating the source of faecal contamination can be directly related to the regulatory results. These regulatory samples are collected once per week over the 20 week bathing water sampling period in May to September of each year and are analysed at the Environment Agency's central laboratories.

Using the Saltburn bathing water catchment as an example, taking seven samples (bathing water, Pit Hills Stell, Skelton Beck, Saltburn Gill, Dunsdale, Upleatham, Tocket's Mill) each of the 20 EA sampling events would cost approximately £42,000 assuming £300 per sample due to additional bioinformatics analysis required. A storm event could be fully captured by sampling all 26 CSOs as well as the main rivers and bathing water, totalling 30 samples. The total cost for sampling and analysing a full bathing season including one storm event would therefore be £51,000.

This is in line with previous hydrological model costs for this catchment (Waites 2012) yet unlike hydrological modelling, MST has the clear advantage of establishing which point sources are contributing most to faecal indicator abundance, as well as highlighting any unexpected sources via faecal host markers. This would enable a

water company to identify which of their assets are contributing the most microbial contamination as opposed to which assets contribute the most water. Furthermore the water company would understand what proportion of overall contamination arises from their assets and therefore what impact an intervention is likely to have on the bathing water itself. For example, a single storm tank was built at the Guisborough CSO at a cost of £6m which would reduce the number of spills to 3 per season which was used as a "rule of thumb" to achieve "Good" bathing water status. This was entirely based on water flow as opposed to investigating the true source of bacteria in the bathing water. Therefore an MST study costing less than 1% of the total project budget would provide greater insight in to the proportion of sewage contamination of the bathing water from each point source, as well as quantifying the overall contribution to faecal contamination from sewage. This could result in avoiding investment in assets which would not sufficiently improve bathing water quality, but conversely could increase the evidence base for proposed interventions and thereby increase the likelihood of implementation.

In terms of market demand, 167 bathing waters were classified below "Excellent" in the UK in 2015, while EU wide there were 1926 such beaches. Other opportunities for NGS MST include tracking the source of faecal bacteria entering potable water treatment works or contamination sources in the potable networks. Therefore NGS has an important role to play in the water and wastewater industry and should be further investigated in order to understand how best to integrate these analyses in to existing practice.

Chapter 11. Conclusions

The case studies presented in this investigation have demonstrated the power of next generation sequencing for gaining new insights in to the microbiology so vital to the water and wastewater industry. It is clear that NGS is a useful, precise and accurate method of enumerating and identifying bacteria, and therefore the challenge is to identify new applications for this data.

The first case study identified some potential uses of NGS in the monitoring and control of activated sludge, however the cost benefit analysis did not provide a compelling case for the regular analysis of ASP samples by NGS. Instead it may be that NGS can be used to inform the design of ASPs by gaining a better understanding of the interactions and function of bacteria, both in terms of existing and emerging contaminants. NGS is also of use when investigating "one off" issues, such as unexplained nitrification failure or intractable foaming. In this way, NGS can bring about cost savings in the construction and operation of ASPs, as well as play an important part in meeting new challenges such as the reduction of micropollutants.

The second case study highlighted the significant impact NGS can have on managing bathing water quality, but also the many knowledge gaps which exist. This is no doubt due to the novelty of this application of NGS, both in terms of identification of faecal markers and whole community analysis. However in contrast with the first study, the cost benefit analysis for bathing waters shows a clear case for further use of NGS in identifying the sources of faecal contamination. NGS is able to match existing approaches in terms of cost, yet provides significantly greater insight in to the real world impact of diffuse as well as individual point sources. However much work has yet to be done in terms of identifying and understanding the role individual faecal markers have to play, as well as the resolution achievable with whole community analysis.

These case studies have identified the many areas which require optimisation in NGS analysis. These begin at the point of sampling – how we can ensure samples contain comparable microbial communities? – and end with converting bacterial taxonomy and abundance in to data with operational meaning – when and why will an ASP have a bulking episode? This effort is further complicated by the ever shifting landscape of

NGS capabilities and the software tools available for sequence data interpretation. The benefits from utilising new algorithms and sequencing chemistries must be balanced against the need for protocol validation and industry trust.

This investigation set out to develop and validate the protocols, and data analysis approaches for the use of NGS in the water industry. In both case studies, I have demonstrated that the methods detailed can provide meaningful insight in to important microbial processes, and in particular with the activated sludge work have demonstrated that NGS results are comparable with existing approaches. The microbial source tracking case study has highlighted the fact that we do not truly understand that taxa which are measured by existing approaches, and therefore whilst direct comparisons could not be made, future studies may be able to elucidate this link.

More generally, next generation sequencing opens new avenues of exploration. Questions which were previously unanswerable, such as the metabolic potential of an activated sludge plant, can now be addressed. Furthermore, the use of NGS in areas other than 16S rRNA gene analysis has yet to be investigated. Whole genome analysis of colonies on a membrane filter could potentially track faecal indicator bacteria directly to a host animal, or even to an individual herd. Metagenomics will enable the discovery of new species and direct interrogation of their metabolic potential by analysis of gene homology. There exists much to learn about the bacteria which serve us so well, and we are just at the beginning of this journey.

Chapter 12. References

Adam, J. & Walmsley, R., 2001. Bathing Water Sampling Procedure ES005. , p.8.

Ahmed, W. et al., 2015. Toolbox Approaches Using Molecular Markers and 16S rRNA Gene Amplicon Data Sets for Identification of Fecal Pollution in Surface Water C. A. Elkins, ed. *Applied and Environmental Microbiology*, 81(20), pp.7067–7077. Available at: http://aem.asm.org/lookup/doi/10.1128/AEM.02032-15 [Accessed September 8, 2016].

Ahmed, W., Hughes, B. & Harwood, V., 2016. Current Status of Marker Genes of Bacteroides and Related Taxa for Identifying Sewage Pollution in Environmental Waters. *Water*, 8(6), p.231. Available at: http://www.mdpi.com/2073-4441/8/6/231 [Accessed September 24, 2016].

Albertsen, M. et al., 2015. Back to basics - The influence of DNA extraction and primer choice on phylogenetic analysis of activated sludge communities. *PLoS ONE*, 10(7).

Allen, L.A., 1944. The bacteriology of activated sludge. *The Journal of hygiene*, 43(6), pp.424–431. Available at: <Go to ISI>://MEDLINE:20475707.

Altschul, S.F. et al., 1990. Basic local alignment search tool. *Journal of molecular biology*, 215(3), pp.403–410.

Amand, L. & Carlsson, B., 2012. Optimal aeration control in a nitrifying activated sludge process. *Water Research*, 46(7), pp.2101–2110.

Amann, R.I., Ludwig, W. & Schleifer, K.H., 1995. PHYLOGENETIC IDENTIFICATION AND IN-SITU DETECTION OF INDIVIDUAL MICROBIAL-CELLS WITHOUT CULTIVATION. *Microbiological Reviews*, 59(1), pp.143–169. Available at: <Go to ISI>://WOS:A1995QK98600007.

Angiuoli, S. V et al., 2011. CloVR: A virtual machine for automated and portable sequence analysis from the desktop using cloud computing. *BMC Bioinformatics*, 12(1), p.356. Available at: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-356 [Accessed September 19, 2016].

Angly, F.E. et al., 2014. CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. *Microbiome*, 2, p.11. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4021573&tool=pmcentrez&rendertype=abstract [Accessed February 18, 2016].

Ardern, E. & Lockett, W.T., 1914. Experiments on the oxidation of sewage without the aid of filters. *Journal of the Society of Chemical Industry*, 33(10), pp.523–539. Available at: http://doi.wiley.com/10.1002/jctb.5000331005 [Accessed June 10, 2015].

Arenskötter, M. et al., 2004. Biology of the metabolically diverse genus Gordonia. *Applied and Environmental Microbiology*, 70(6), pp.3195–3204. Available at: <Go to ISI>://000221981100002 [Accessed February 3, 2016].

Armougom, F. & Raoult, D., 2009. Exploring Microbial Diversity Using 16S rRNA High-Throughput Methods. *Journal of Computer Science & Systems Biology*, 2(1), pp.74–92.

Asvapathanagul, P. et al., 2015. Identification and quantification of Thiothrix eikelboomii using qPCR for early detection of bulking incidents in a full-scale water reclamation plant. *Applied Microbiology and Biotechnology*, 99(9), pp.4045–4057. Available at: http://link.springer.com/10.1007/s00253-014-6230-5.

Bae, S., 2009. Rapid decay of host-specific fecal Bacteroidales cells in seawater as measured by quantitative PCR with propidium monoazide. *Water Research*, 43(19), pp.4850–4859.

Bae, S. & Wuertz, S., 2015. Decay of host-associated Bacteroidales cells and DNA in

continuous-flow freshwater and seawater microcosms of identical experimental design and temperature as measured by PMA-qPCR and qPCR. *Water Research*, 70, pp.205–213. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0043135414007301.

Balakrishnan, S. & Eckenfelder, W.W., 1969. NITROGEN RELATIONSHIPS IN BIOLOGICAL TREATMENT PROCESSES I NITRIFICATION IN THE ACTIVATED SLUDGE PROCESS. *Water Research*, 3(1), pp.73–81. Available at: <Go to ISI>://BCI:BCI196950094275.

Balleste, E. & Blanch, A.R., 2011. Bifidobacterial Diversity and the Development of New Microbial Source Tracking Indicators. *Applied and Environmental Microbiology*, 77(10), pp.3518–3525. Available at: http://aem.asm.org/cgi/doi/10.1128/AEM.02198-10 [Accessed September 7, 2016].

Baptista, J.D.C. et al., 2014. Agreement between amoA gene-specific quantitative PCR and fluorescence in situ hybridization in the measurement of ammonia-oxidizing bacteria in activated sludge. *Applied and environmental microbiology*, 80(19), pp.5901–10. Available at: http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=1&SID=U22oqJbv1wru44uycwb&page=1&doc=1 [Accessed May 26, 2015].

Bartlett, J.M.S. & Stirling, D., 2003. *Methods in molecular biology: PCR protocols*,

Behrendt, L. et al., 2012. Microbial diversity of biofilm communities in microniches associated with the didemnid ascidian Lissoclinum patella. *The ISME Journal*, 6(6), pp.1222–1237.

Bellucci, M. & Curtis, T.P., 2011. AMMONIA-OXIDIZING BACTERIA IN WASTEWATER. In M. G. Klotz & L. Y. Stein, eds. *Methods in Enzymology, Vol 46: Research on Nitrification and Related Processes, Pt B*. San Diego: Elsevier Academic Press Inc, pp. 269–286. Available at: <Go to ISI>://WOS:000290776300011.

Bergey, D.H., 2005. *Bergey's Manual of Systematic Bacteriology Volume Two Part B The Gammaproteobacteri* D. J. Brenner, ed.,

Bianco, K. et al., 2015. Fecal pollution source tracking in waters intended for human supply based on archaeal and bacterial genetic markers. *Journal of Water and Health*, 13(4), pp.985–995. Available at: http://jwh.iwaponline.com/lookup/doi/10.2166/wh.2015.292 [Accessed September 7, 2016].

Blackall, L.. et al., 1985. Biological foam and scum formation in activated sludge treatment systems. In *Australian Water and Wastewater Association 11th convention*.

Blackall, L.L. et al., 1991. ACTIVATED-SLUDGE FOAMS - EFFECTS OF ENVIRONMENTAL VARIABLES ON ORGANISM GROWTH AND FOAM FORMATION. *Environmental Technology*, 12(3), pp.241–248. Available at: <Go to ISI>://WOS:A1991FM40600005.

Blackburne, R. et al., 2007. Kinetic characterisation of an enriched Nitrospira culture with comparison to Nitrobacter. *Water research*, 41(14), pp.3033–42. Available at: http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=8&SID=P2SKMkNOKSzLTmBnkxV&page=1&doc=10 [Accessed February 12, 2016].

Blanch, A.R. et al., 2006. Integrated Analysis of Established and Novel Microbial and Chemical Methods for Microbial Source Tracking. *Applied and Environmental Microbiology*, 72(9), pp.5915–5926. Available at: http://aem.asm.org/cgi/doi/10.1128/AEM.02453-05 [Accessed July 25, 2016].

Boehm, A.B., 2007. Enterococci Concentrations in Diverse Coastal Environments Exhibit Extreme Variability.

Boehm, A.B. et al., 2013. Performance of forty-one microbial source tracking methods: A

twenty-seven lab evaluation study. *Water Research*, 47(18), pp.6812–6828.

Bofill-Mas, S. et al., 2011. Cost-effective Method for Microbial Source Tracking Using Specific Human and Animal Viruses. *Journal of Visualized Experiments*, (58). Available at: http://www.jove.com/details.php?id=2820 [Accessed August 2, 2016].

Bofill-Mas, S. et al., 2010. Quantification of Human Adenoviruses in European Recreational Waters. *Food and Environmental Virology*, 2(2), pp.101–109. Available at: http://link.springer.com/10.1007/s12560-010-9035-4 [Accessed August 2, 2016].

Bond, P.L. et al., 1995. Bacterial community structures of phosphate-removing and non-phosphate-removing activated sludges from sequencing batch reactors. *Applied and environmental microbiology*, 61(5), pp.1910–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7544094 [Accessed September 19, 2016].

Bonilla, T.D. et al., 2007. Prevalence and distribution of fecal indicator organisms in South Florida beach sand and preliminary assessment of health effects associated with beach sand exposure. *Marine Pollution Bulletin*, 54(9), pp.1472–1482.

Boon, B. & Laudelout, H., 1962. Kinetics of nitrite oxidation by Nitrobacter winogradskyi. *Biochemical Journal*, 85(3), pp.440–447. Available at: http://www.biochemj.org/content/85/3/440.abstract [Accessed February 16, 2016].

Boon, N. et al., 2000. Bioaugmentation of activated sludge by an indigenous 3-chloroaniline-degrading Comamonas testosteroni strain, I2gfp. *Applied and Environmental Microbiology*, 66(7), pp.2906–2913. Available at: <Go to ISI>://WOS:000088057600029.

Bouchez, T. et al., 2000. Ecological study of a bioaugmentation failure. *Environmental Microbiology*, 2(2), pp.179–190. Available at: <Go to ISI>://WOS:000087106600006.

Brown, J., 1964. THE PUBLIC HEALTH HAZARDS OF BATHING. *Medical services journal, Canada*, 20, pp.135–147. Available at: http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=1&SID=X2HM79YkaGKOKWnjiZd&page=1&doc=6 [Accessed June 15, 2015].

Brown, K.I. & Boehm, A.B., 2015. Comparative decay of Catellicoccus marimmalium and enterococci in beach sand and seawater. *Water Research*, 83, pp.377–384. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0043135415300543.

Byappanahalli, M.N. et al., 2012. Enterococci in the environment. *Microbiology and molecular biology reviews : MMBR*, 76(4), pp.685–706. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3510518&tool=pmcentrez&rendertype=abstract.

Cao, Y. et al., 2013. Evaluation of molecular community analysis methods for discerning fecal sources and human waste. *Water Research*, 47(18), pp.6862–6872.

Caporaso, J.G. et al., 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences of the United States of America*, 108 Suppl(Supplement_1), pp.4516–22. Available at: http://www.pnas.org/content/108/Supplement_1/4516.full [Accessed May 11, 2015].

Caporaso, J.G. et al., 2010. PyNAST: A flexible tool for aligning sequences to a template alignment. *Bioinformatics*, 26(2), pp.266–267.

Caporaso, J.G. et al., 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5), pp.335–336.

Cersosimo, L.M. et al., 2016. Influence of periparturient and postpartum diets on rumen methanogen communities in three breeds of primiparous dairy cows. *BMC Microbiology*,

16(1), p.78. Available at: http://bmcmicrobiol.biomedcentral.com/articles/10.1186/s12866-016-0694-7 [Accessed September 7, 2016].

Chain, P. et al., 2003. Complete genome sequence of the ammonia-oxidizing bacterium and obligate chemolithoautotroph Nitrosomonas europaea. *Journal of bacteriology*, 185(9), pp.2759–73. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=154410&tool=pmcentrez&rendertype=abstract [Accessed February 3, 2016].

Chauhan, S. & Sharma, K., 2014. Effect of temperature and additives on the critical micelle concentration and thermodynamics of micelle formation of sodium dodecyl benzene sulfonate and dodecyltrimethylammonium bromide in aqueous solution: A conductometric study. *The Journal of Chemical Thermodynamics*, 71, pp.205–211. Available at: http://www.sciencedirect.com/science/article/pii/S0021961413004849 [Accessed December 10, 2015].

Chudoba, J., Grau, P. & Ottova, V., 1973. CONTROL OF ACTIVATED-SLUDGE FILAMENTOUS BULKING .2. SELECTION OF MICROORGANISMS BY MEANS OF A SELECTOR. *Water Research*, 7(10), p.1389-. Available at: <Go to ISI>://WOS:A1973Q774800001.

Claesson, M.J. et al., 2009. Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. *PLoS ONE*, 4(8).

Claesson, M.J. et al., 2010. Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Research*, 38(22), p.13. Available at: <Go to ISI>://WOS:000285420300004 [Accessed June 8, 2015].

Clara, M. et al., 2005. Removal of selected pharmaceuticals, fragrances and endocrine disrupting compounds in a membrane bioreactor and conventional wastewater treatment plants. *Water Research*, 39(19), pp.4797–4807. Available at: <Go to ISI>://WOS:000233812000021.

Cokro, A.A. et al., 2017. Non-denitrifying polyphosphate accumulating organisms obviate requirement for anaerobic condition. *Water Research*, 111, pp.393–403.

Converse, R.R. et al., 2012. Dramatic Improvements in Beach Water Quality Following Gull Removal.

Coskuner, G. et al., 2005. Agreement between theory and measurement in quantification of ammonia-oxidizing bacteria. *Applied and Environmental Microbiology*, 71(10), pp.6325–6334. Available at: http://aem.asm.org/content/71/10/6325.full [Accessed September 7, 2015].

Coskuner, G. & Curtis, T.P.P., 2002. In situ characterization of nitrifiers in an activated sludge plant: detection of Nitrobacter Spp. *Journal of Applied Microbiology*, 93(3), pp.431–437. Available at: <Go to ISI>://WOS:000177382500009 [Accessed February 12, 2016].

Coskuner, G. & Jassim, M.S., 2008. Development of a correlation to study parameters affecting nitrification in a domestic wastewater treatment plant. *Journal of Chemical Technology and Biotechnology*, 83(3), pp.299–308. Available at: <Go to ISI>://WOS:000253688500012.

Court of Justice of the European Union, 2003. *Commission of the European Communities v Kingdom of Spain C-278/01*,

Crocetti, G.R. et al., 2000. Identification of polyphosphate-accumulating organisms and design

of 16S rRNA-directed probes for their detection and quantitation. *Applied and Environmental Microbiology*, 66(3), pp.1175–1182. Available at: <Go to ISI>://000085604800044 [Accessed February 19, 2016].

D'Erchia, A.M. et al., 2015. Tissue-specific mtDNA abundance from exome data and its correlation with mitochondrial transcription, mass and respiratory activity. *Mitochondrion*, 20, pp.13–21.

Daims, H., Ramsing, N.B., et al., 2001. Cultivation-independent, semiautomatic determination of absolute bacterial cell numbers in environmental samples by fluorescence in situ hybridization. *Applied and environmental microbiology*, 67(12), pp.5810–5818. Available at: http://aem.asm.org/content/67/12/5810.full [Accessed September 7, 2015].

Daims, H., Nielsen, J.L., et al., 2001. In situ characterization of Nitrospira-like nitrite-oxidizing bacteria active in wastewater treatment plants. *Applied and environmental microbiology*, 67(11), pp.5273–84. Available at: http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=2&SID=P2SKMkNOKSzLTmBnkxV&page=1&doc=4 [Accessed February 16, 2016].

Davenport, R.J. et al., 2008. A universal threshold concept for hydrophobic mycolata in activated sludge foaming. *Water Research*, 42(13), pp.3446–3454. Available at: http://www.sciencedirect.com/science/article/pii/S0043135408001115 [Accessed February 3, 2016].

Davenport, R.J. et al., 2000. Quantitative use of fluorescent in situ hybridization to examine relationships between mycolic acid-containing actinomycetes and foaming in activated sludge plants. *Applied and Environmental Microbiology*, 66(3), pp.1158–1166. Available at: <Go to ISI>://WOS:000085604800042.

Davenport, R.J. et al., 2005. Selective inhibition of mycolata to control foam while maintaining nitrification in activated sludge plants. In S. J. Hall & L. L. Blackall, eds. *Microbial Population Dynamics in Biological Wastewater Treatment (ASPD4)*. Gold Coast, Australia: IWA Publishing.

Davenport, R.J. & Curtis, T.P., 2004. Quantitative fluorescence in situ hybridisation (FISH): statistical methods for valid cell counting. In *Molecular Microbial Ecology Manual*. pp. 1487–1516.

de-Bashan, L.E. & Bashan, Y., 2004. Recent advances in removing phosphorus from wastewater and its future use as fertilizer (1997-2003). *Water research*, 38(19), pp.4222–46. Available at: http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=4&SID=S2wj766ttzi9aAqdmBp&page=1&doc=3 [Accessed May 15, 2015].

Department for Environment Food and Agriculture, 2015. England's bathing water results 2015. Available at: https://www.gov.uk/government/news/englands-bathing-water-results-2015 [Accessed July 22, 2016].

Derrien, M. et al., 2011. Extreme Variability of Steroid Profiles in Cow Feces and Pig Slurries at the Regional Scale: Implications for the Use of Steroids to Specify Fecal Pollution Sources in Waters. *Journal of Agricultural and Food Chemistry*, 59(13), pp.7294–7302. Available at: http://pubs.acs.org/doi/abs/10.1021/jf201040v [Accessed July 25, 2016].

DeSantis, T.Z. et al., 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, 72(7), pp.5069–5072.

Dias, F.F. & Bhat, J. V, 1964. MICROBIAL ECOLOGY OF ACTIVATED SLUDGE. I. DOMINANT

BACTERIA. *Applied microbiology*, 12, pp.412–417. Available at: <Go to ISI>://MEDLINE:14215970.

Dickenson, J.A. & Sansalone, J.J., 2012. Distribution and disinfection of bacterial loadings associated with particulate matter fractions transported in urban wet weather flows. *Water research*, 46(20), pp.6704–14. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22244969 [Accessed July 22, 2016].

Diston, D. et al., 2015. Evaluation of Molecular- and Culture-Dependent MST Markers To Detect Fecal Contamination and Indicate Viral Presence in Good Quality Groundwater. *Environmental science & technology*, 49(12), pp.7142–51. Available at: http://dx.doi.org/10.1021/acs.est.5b00515 [Accessed July 17, 2015].

Diston, D., Ebdon, J.E. & Taylor, H.D., 2014. Inactivation of Bacteriophage Infecting *Bacteroides* Strain GB124 Using UV-B Radiation. *Photochemistry and Photobiology*, 90(3), pp.622–627. Available at: http://doi.wiley.com/10.1111/php.12223 [Accessed June 23, 2016].

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), p.654. Available at: http://jwh.iwaponline.com/cgi/doi/10.2166/wh.2014.199 [Accessed August 2, 2016].

Donald M. Stoeckel, *,† et al., 2004. Comparison of Seven Protocols To Identify Fecal Contamination Sources Using Escherichia coli.

Drancourt, M. et al., 2000. 16S Ribosomal DNA Sequence Analysis of a Large Collection of Environmental and Clinical Unidentifable Bacterial Isolates. *Journal of Clinical Microbiology*, 38(10), pp.3623–3630.

Drechsel, P., Qadir, M. & Wichelns, D. eds., 2015. *Wastewater* 1st ed., Springer Netherlands. Available at: http://www.springer.com/us/book/9789401795449.

Eaton, A.D. & Franson, M.A.H., 2005. *Standard methods for the examination of water & wastewater*, Amer Public Health Assn.

Ebentier, D.L. et al., 2013. Evaluation of the repeatability and reproducibility of a suite of qPCR-based microbial source tracking methods. *Water research*, 47(18), pp.6839–48. Available at: http://www.sciencedirect.com/science/article/pii/S0043135413005484 [Accessed June 17, 2015].

Edgar, R.C., 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*, 5, p.113.

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

Edwards, A.C., Watson, H.A. & Cook, Y.E.M., 2012. Source strengths, transport pathways and delivery mechanisms of nutrients, suspended solids and coliforms within a small agricultural headwater catchment. *Science of The Total Environment*, 434, pp.123–129.

EEC, 1976. *Bathing Water Directive*, EEC.

EEC Council, 1991. 91/271/EEC of 21 May 1991 concerning urban waste-water treatment. *EEC Council Directive*, (L), p.10.

Ehlers, G.A.C., Wagachchi, D. & Turner, S.J., 2012. Nutrient conditions and reactor configuration influence floc size distribution and settling properties. *Water Science and Technology*, 65(1), pp.156–163. Available at: <Go to ISI>://WOS:000299163100020.

van Elsas, J.D. et al., 2011. Survival of Escherichia coli in the environment: fundamental and public health aspects. *The ISME journal*, 5(2), pp.173–83. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/20574458 [Accessed September 25, 2016].

Environment Agency, 2009. *Limiting Climate Change - Water industry carbon reduction*,

Environment Agency, 2012. *The Microbiology of Drinking Water (2012) - Part 5 - Methods for the Isolation and enumeration of enterococci*,

Eren, A.M. et al., 2013. Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data R. Freckleton, ed. *Methods in Ecology and Evolution*, 4(12), pp.1111–1119. Available at: http://doi.wiley.com/10.1111/2041-210X.12114 [Accessed September 19, 2016].

Erhart, R. et al., 1997. Development and use of fluorescent in situ hybridization probes for the detection and identification of Microthrix parvicella'' in activated sludge. *Systematic and Applied Microbiology*, 20(2), pp.310–318. Available at: <Go to ISI>://WOS:A1997XJ63600016.

European Commission, 2016. Bathing water quality. Available at: http://ec.europa.eu/environment/water/water-bathing/summary.html [Accessed May 6, 2016].

European Commission, 2008. Regulation (EC) No 1272/2008 of the European Parliament and of the Council of 16 December 2008 on classification, labelling and packaging of substances and mixtures, amending and repealing Directives 67/548/EEC and 1999/45/EC, and amending Regulation (EC). *Official Journal of the European Union*, 353, pp.1–1355.

European Environment Agency, 2015. *European bathing water quality in 2014*, Available at: http://www.eea.europa.eu/publications/european-bathing-water-quality-in-2014#tab-data-visualisations [Accessed May 6, 2016].

European Parliament, 2006. Directive 2006/7/EC of the European Parliament and of the Council of 15 February 2006 concerning the management of bathing water quality and repealing Directive 76/160/EEC. *Official Journal of the European Union, L series*, 64, pp.37–51.

European Parliament and Council, 2000. *Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy*, EU.

European Parliament and Council, 2012. *Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL amending Directives 2000/60/EC and 2008/105/EC as regards priority substances in the field of water policy*, EU. Available at: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52011PC0876&from=EN.

Eyice, O. et al., 2007. Identification of nitrifiers in a full-scale biological treatment system using fluorescent in situ hybridization. *Journal of Environmental Science and Health Part a-Toxic/Hazardous Substances & Environmental Engineering*, 42(4), pp.517–523. Available at: <Go to ISI>://WOS:000245405000015.

Eyice, Ö., Ince, O. & Ince, B.K., 2014. Monitoring the abundance and the activity of ammonia-oxidizing bacteria in a full-scale nitrifying activated sludge reactor. *Environmental Science and Pollution Research*, 22(3), pp.2328–2334. Available at: http://link.springer.com/10.1007/s11356-014-3519-y.

Falas, P. et al., 2012. Impact of solid retention time and nitrification capacity on the ability of activated sludge to remove pharmaceuticals. *Environmental Technology*, 33(8), pp.865–872. Available at: <Go to ISI>://WOS:000303581600004.

Federici, S. et al., 2015. Archaeal microbiota population in piglet feces shifts in response to weaning: Methanobrevibacter smithii is replaced with Methanobrevibacter boviskoreani.

FEMS Microbiology Letters, 362(10), p.fnv064-fnv064. Available at: http://femsle.oxfordjournals.org/cgi/doi/10.1093/femsle/fnv064 [Accessed September 7, 2016].

Figueras, M.J. et al., 1997. Past, present and future perspectives of the EU Bathing Water Directive. *Marine Pollution Bulletin*, 34(3), pp.148–156. Available at: http://www.sciencedirect.com/science/article/pii/S0025326X97001070 [Accessed May 6, 2016].

Fitzgerald, C.M. et al., 2015. Ammonia-oxidizing microbial communities in reactors with efficient nitrification at low-dissolved oxygen. *Water Research*, 70, pp.38–51. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0043135414008203 [Accessed February 12, 2016].

Fleisher, J.M. et al., 1996. The enterovirus test in the assessment of recreational water-associated gastroenteritis. *Water Research*, 30(10), pp.2341–2346. Available at: http://linkinghub.elsevier.com/retrieve/pii/0043135496001261 [Accessed August 2, 2016].

Flynn, R.M. et al., 2016. Using microbiological tracers to assess the impact of winter land use restrictions on the quality of stream headwaters in a small catchment. *Science of The Total Environment*, 541, pp.949–956. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0048969715307312 [Accessed July 25, 2016].

Fox, G.E. et al., 1977. Classification of methanogenic bacteria by 16S ribosomal RNA characterization. *Proceedings of the National Academy of Sciences*, 74(10), pp.4537–4541. Available at: http://www.pnas.org/content/74/10/4537 [Accessed February 10, 2016].

Fox, G.E., Wisotzkey, J.D. & JR., P.J., 1992. How Close Is Close: 16S rRNA Sequence Identity May Not Be Sufficient To Guarantee Species Identity. *International Journal of Systematic and Evolutionary Microbiology*, 42(1), pp.166–170.

Fu, L.-L. et al., 2011. Temporal genetic variability and host sources of Escherichia coli associated with fecal pollution from domesticated animals in the shellfish culture environment of Xiangshan Bay, East China Sea. *Environmental pollution (Barking, Essex : 1987)*, 159(10), pp.2808–2814. Available at: <Go to ISI>://MEDLINE:21645948.

Furtula, V. et al., 2012. Sewage treatment plants efficiencies in removal of sterols and sterol ratios as indicators of fecal contamination sources. *Water, Air, and Soil Pollution*, 223(3), pp.1017–1031. Available at: http://link.springer.com/10.1007/s11270-011-0920-8 [Accessed June 16, 2015].

Gammie, A., 1997. Does hepatitis a pose a significant health risk to recreational water users? *Water Science and Technology*, 35(11–12), pp.171–177. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0273122397002540 [Accessed August 2, 2016].

Geladi, P. & Kowalski, B.R., 1986. Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, 185(C), pp.1–17.

Georges, K., Thornton, A. & Sadler, R., 2009. Evidence: Transforming wastewater treatment to reduce carbon emissions E. Agency, ed. *Environment Agency Webpage, UK Government*. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/291633/scho1209brnz-e-e.pdf.

Georgiou, S. & Bateman, I.J., 2005. Revision of the EU Bathing Water Directive: economic costs

and benefits. *Marine Pollution Bulletin*, 50(4), pp.430–438.

Gibbons, N.E. & Murray, R.G.E., 1978. Proposals Concerning the Higher Taxa of Bacteria. *International Journal of Systematic Bacteriology*, 28(1), pp.1–6. Available at: http://ijs.microbiologyresearch.org/content/journal/ijsem/10.1099/00207713-28-1-1 [Accessed September 18, 2016].

Gieseke, A. et al., 2003. Structure and activity of multiple nitrifying bacterial populations co-existing in a biofilm. *Environmental Microbiology*, 5(5), pp.355–369. Available at: http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=16&SID=P2SKMkNOKSzLTmBnkxV&page=1&doc=9 [Accessed January 22, 2016].

Gilpin, B. et al., 2003. The use of chemical and molecular microbial indicators for faecal source identification. *Water Science and Technology*, 47(3), pp.39–43. Available at: <Go to ISI>://WOS:000181785000008.

Gilpin, B.J., Gregor, J.E. & Savill, M.G., 2002. Identification of the source of faecal pollution in contaminated rivers. *Water Science and Technology*, 46(3), pp.9–15. Available at: <Go to ISI>://WOS:000178025500002.

Girones, R. et al., 2010. Molecular detection of pathogens in water - The pros and cons of molecular techniques. *Water Research*, 44(15), pp.4325–4339. Available at: http://www.sciencedirect.com/science/article/pii/S0043135410004173 [Accessed February 12, 2016].

Glenn, T.C., 2011. Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, 11(5), pp.759–769. Available at: <Go to ISI>://000293915800001.

Gómez-Doñate, M. et al., 2012. New molecular quantitative pcr assay for detection of host-specific bifidobacteriaceae suitable for microbial source tracking. *Applied and environmental microbiology*, 78(16), pp.5788–5795. Available at: http://aem.asm.org/content/78/16/5788.full [Accessed May 27, 2015].

Gomi, R. et al., 2014. Fecal Source Tracking in Water by Next-Generation Sequencing Technologies Using Host-Specific Escherichia coli Genetic Markers. *Environmental Science & Technology*, 48(16), pp.9616–9623. Available at: <Go to ISI>://000340701800081.

Gower, J.C., 1975. Generalized procrustes analysis. *Psychometrika*, 40(1), pp.33–51. Available at: http://link.springer.com/10.1007/BF02291478 [Accessed December 29, 2014].

Graham, D.W. et al., 2007. Experimental demonstration of chaotic instability in biological nitrification. *The ISME journal*, 1(5), pp.385–93. Available at: http://dx.doi.org/10.1038/ismej.2007.45 [Accessed February 12, 2016].

Guo, F. et al., 2013. Taxonomic Precision of Different Hypervariable Regions of 16S rRNA Gene and Annotation Methods for Functional Bacterial Groups in Biological Wastewater Treatment. *PLoS ONE*, 8(10).

Guo, F. & Zhang, T., 2013. Biases during DNA extraction of activated sludge samples revealed by high throughput sequencing. *Applied Microbiology and Biotechnology*, 97(10), pp.4607–4616.

Guo, F. & Zhang, T., 2012. Profiling bulking and foaming bacteria in activated sludge by high throughput sequencing. *Water Research*, 46(8), pp.2772–2782. Available at: <Go to ISI>://MEDLINE:22406202.

Haas, B.J. et al., 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research*, 21(3), pp.494–504.

Haeckel, E.H., 1866. *Generelle Morphologie der Organismen allgemeine Grundzuge der organischen Formen-Wissenschaft, mechanisch begrundet durch die von Charles Darwin reformirte Descendenz-Theorie von Ernst Haeckel: Allgemeine Entwickelungsgeschichte der Organismen kritische G*, Verlag von Georg Reimer.

Hagedorn, C. et al., 1999. Determining sources of fecal pollution in a rural Virginia watershed with antibiotic resistance patterns in fecal streptococci. *Applied and environmental microbiology*, 65(12), pp.5522–31. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10584013 [Accessed July 26, 2016].

Hale, V.L. et al., 2015. Effect of preservation method on spider monkey (Ateles geoffroyi) fecal microbiota over 8weeks. *Journal of Microbiological Methods*, 113, pp.16–26. Available at: http://www.sciencedirect.com/science/article/pii/S0167701215001104 [Accessed January 27, 2016].

Hamady, M. & Knight, R., 2009. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Research*, 19(7), pp.1141–1152.

Harms, G. et al., 2003. Real-Time PCR Quantification of Nitrifying Bacteria in a Municipal Wastewater Treatment Plant. *Environmental Science & Technology*, 37(2), pp.343–351. Available at: http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=14&SID=Y2uO2WvxMuwqrjMRdpM&page=2&doc=13 [Accessed February 12, 2016].

Harwood, V.J. et al., 2013. Performance of viruses and bacteriophages for fecal source determination in a multi-laboratory, comparative study. *Water Research*, 47(18), pp.6929–6943.

Hayes, T.B. et al., 2002. Hermaphroditic, demasculinized frogs after exposure to the herbicide atrazine at low ecologically relevant doses. *Proceedings of the National Academy of Sciences of the United States of America*, 99(8), pp.5476–5480. Available at: <Go to ISI>://WOS:000175087000135.

Helbling, D.E. et al., 2012. Micropollutant biotransformation kinetics associate with WWTP process parameters and microbial community characteristics. *Environmental science & technology*, 46(19), pp.10579–88. Available at: http://pubs.acs.org.libproxy.ncl.ac.uk/doi/abs/10.1021/es3019012 [Accessed January 17, 2016].

Henry, R. et al., 2016. Into the deep: Evaluation of SourceTracker for assessment of faecal contamination of coastal waters. *Water Research*, 93, pp.242–253. Available at: http://www.sciencedirect.com/science/article/pii/S0043135416300884 [Accessed February 26, 2016].

Henze, M. et al., 2000. Activated Sludge Models ASM1, ASM2, ASM2d and ASM3. *IWA Publishing*, p.121. Available at: http://books.google.com/books?hl=en&lr=&id=W0NTq3LUfOUC&oi=fnd&pg=PT7&dq=ACTIVATED+SLUDGE+MODELS+ASM1,+ASM2,+ASM2d+AND+ASM3&ots=km38yqNbuv&sig=2845cSyyl2jo_Kf708i6MfmdE0w%5Cnhttp://test.iwaponline.com/wio/2006/10/wio200610RF1900222248.htm%5Cnhttp://books.

Higuchi, R. et al., 1993. Kinetic PCR Analysis: Real-time Monitoring of DNA Amplification Reactions. *Bio/Technology*, 11(9), pp.1026–1030. Available at: http://www.nature.com/doifinder/10.1038/nbt0993-1026 [Accessed September 18, 2016].

Hildebrand, F. et al., 2014. LotuS: an efficient and user-friendly OTU processing pipeline.

Microbiome, 2(1), p.30. Available at: http://www.microbiomejournal.com/content/2/1/30 [Accessed September 19, 2016].

Hindson, B.J. et al., 2011. High-Throughput Droplet Digital PCR System for Absolute Quantitation of DNA Copy Number. *Analytical Chemistry*, 83(22), pp.8604–8610. Available at: http://pubs.acs.org/doi/abs/10.1021/ac202028g [Accessed July 25, 2016].

Hiorns, W. et al., 1995. AMPLIFICATION OF 16S RIBOSOMAL-RNA GENES OF AUTOTROPHIC AMMONIA-OXIDIZING BACTERIA DEMONSTRATES THE UBIQUITY OF NITROSOSPIRAS IN THE ENVIRONMENT. *MICROBIOLOGY-UK*, 141, pp.2793–2800. Available at: http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=21&SID=Y2uO2WvxMuwqrjMRdpM&page=1&doc=6 [Accessed February 12, 2016].

Hong, S.H. et al., 2009. Polymerase chain reaction primers miss half of rRNA microbial diversity. *Isme Journal*, 3(12), pp.1365–1373. Available at: <Go to ISI>://WOS:000272500800004.

Horan, N.J., 2014. *Activated Sludge Past, Present and Future*, Aqua Enviro Technology Transfer.

Hughes, M.S., Coyle, P. V & Connolly, J.H., 1992. Enteroviruses in recreational waters of Northern Ireland. *Epidemiology and infection*, 108(3), pp.529–36. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1318220 [Accessed August 2, 2016].

Huse, S.M. et al., 2008. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genetics*, 4(11).

Integrated DNA Technologies, 2015. OligoAnalyzer. Available at: http://eu.idtdna.com/calc/analyzer [Accessed May 6, 2015].

International Standards Organisation, 2003. ISO 5667-3 Water quality - Sampling. *Part 3: Guidance on the preservation and handling of water samples*.

International Water Association, 2000. *Activated Sludge Models Scientific and Technical Report No.9*,

Isabelle, M. et al., 2011. Isolation of estrogen-degrading bacteria from an activated sludge bioreactor treating swine waste, including a strain that converts estrone to beta-estradiol. *Canadian Journal of Microbiology*, 57(7), pp.559–568. Available at: <Go to ISI>://WOS:000294932300004.

Ishii, S. & Sadowsky, M.J., 2008. Escherichia coli in the Environment: Implications for Water Quality and Human Health. *Microbes and environments / JSME*, 23(2), pp.101–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21558695 [Accessed September 25, 2016].

Ishikawa, J. et al., 2004. The complete genomic sequence of Nocardia farcinica IFM 10152. *Proceedings of the National Academy of Sciences of the United States of America*, 101(41), pp.14925–30. Available at: http://www.pnas.org/content/101/41/14925.abstract [Accessed November 16, 2015].

Iyengar, V. et al., 1991. Human stools as a source of viable colonic epithelial cells. *FASEB*, 5(13), pp.2856–2859. Available at: http://www.fasebj.org/content/5/13/2856.short [Accessed June 24, 2015].

Janda, J.M. & Abbott, S.L., 2007. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. *Journal of Clinical Microbiology*, 45(9), pp.2761–2764.

Jenkins, D., Richard, M. & Daigger, G., 2003. *Manual on the Causes and Control of Activated*

Sludge Bulking, Foaming and other Solids Separation Problems 3rd ed., IWA Publishing.

Jeong, J.Y. et al., 2011. Microbial Community Analysis and Identification of Alternative Host-Specific Fecal Indicators in Fecal and River Water Samples Using Pyrosequencing. *Journal of Microbiology*, 49(4), pp.585–594. Available at: <Go to ISI>://000294443100009.

Johnston, C. et al., 2013. Probabilistic analysis showing that a combination of Bacteroides and Methanobrevibacter source tracking markers is effective for identifying waters contaminated by human fecal pollution. *Environmental science & technology*, 47(23), pp.13621–8. Available at: http://pubs.acs.org/doi/abs/10.1021/es403753k [Accessed May 27, 2015].

Jones, K. & Obiri-Danso, K., 1998. Non-compliance of beaches with the EU directives of bathing water quality: evidence of non-point sources of pollution in Morecambe Bay. *Journal of applied microbiology*, 85 Suppl 1, p.101S–107S. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21182698 [Accessed May 6, 2016].

Juretschko, S. et al., 1998. Combined molecular and conventional analyses of nitrifying bacterium diversity in activated sludge: Nitrosococcus mobilis and Nitrospira-like bacteria as dominant populations. *Applied and Environmental Microbiology*, 64(8), pp.3042–3051. Available at: <Go to ISI>://WOS:000075127200044.

Juretschko, S. et al., 2002. The microbial community composition of a nitrifying-denitrifying activated sludge from an industrial sewage treatment plant analyzed by the full-cycle rRNA approach. *Systematic and Applied Microbiology*, 25(1), pp.84–99. Available at: <Go to ISI>://WOS:000176285200011.

Kamra, A. et al., 2005. Exfoliated colonic epithelial cells: surrogate targets for evaluation of bioactive food components in cancer prevention. *The Journal of nutrition*, 135(11), pp.2719–22. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16251637 [Accessed July 25, 2016].

Kao, K.J. et al., 2015. A fluorescence in situ hybridization (FISH) microfluidic platform for detection of HER2 amplification in cancer cells. *Biosensors and Bioelectronics*, 69, pp.272–279.

Kapoor, V. et al., 2013. Correlative assessment of fecal indicators using human mitochondrial DNA as a direct marker. *Environmental Science and Technology*, 47(18), pp.10485–10493.

Karst, S.M. et al., 2013. DNA Extraction and primer selection. In *IWA Microbial Ecology for Wastewater Engineering*. International Water Association. Available at: http://www.slideshare.net/sorenkarst/130707-karst-mewe13-dnaextractionnonotes.

Katoh, K. & Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4), pp.772–780.

Kembel, S.W. et al., 2012. Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance C. von Mering, ed. *PLoS Computational Biology*, 8(10), p.e1002743. Available at: http://dx.plos.org/10.1371/journal.pcbi.1002743 [Accessed September 23, 2016].

van Kessel, M.A.H.J. et al., 2015. Complete nitrification by a single microorganism. *Nature*, 528(7583), p.555–+. Available at: http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=4&SID=Y2uO2WvxMuwqrjMRdpM&page=1&doc=1 [Accessed November 26, 2015].

Klindworth, A. et al., 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for

classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research*, 41(1). Available at: <Go to ISI>://WOS:000312889900001.

Knights, D. et al., 2011. Bayesian community-wide culture-independent microbial source tracking. *Nature Methods*, 8(9), pp.761-U107. Available at: <Go to ISI>://000294439100014.

Koops, H. et al., 2006. The Lithoautotrophic Ammonia-Oxidizing Bacteria. *The Prokaryotes*, 5, pp.778–811.

Kopczynski, E.D., Bateson, M.M. & Ward, D.M., 1994. Recognition of chimeric small-subunit ribosomal DNAs composed of genes from uncultivated microorganisms. *Applied and environmental microbiology*, 60(2), pp.746–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8135524 [Accessed September 19, 2016].

Kowalchuk, G.A. & Stephen, J.R., 2001. Ammonia-oxidizing bacteria: A model for molecular microbial ecology. *Annual Review of Microbiology*, 55, pp.485–529. Available at: <Go to ISI>://WOS:000171732600020 [Accessed February 12, 2016].

Kozlowski, J.A., Kits, K.D. & Stein, L.Y., 2016. Genome Sequence of Nitrosomonas communis Strain Nm2, a Mesophilic Ammonia-Oxidizing Bacterium Isolated from Mediterranean Soil. *Genome announcements*, 4(1), pp.e01541-15. Available at: http://genomea.asm.org/content/4/1/e01541-15.full [Accessed February 3, 2016].

Kragelund, C. et al., 2007. Ecophysiology of mycolic acid-containing Actinobacteria (Mycolata) in activated sludge foams. *Fems Microbiology Ecology*, 61(1), pp.174–184. Available at: http://femsec.oxfordjournals.org/content/61/1/174.abstract [Accessed February 3, 2016].

Kragelund, C. et al., 2011. Eikelboom's morphotype 0803 in activated sludge belongs to the genus Caldilinea in the phylum Chloroflexi. *Fems Microbiology Ecology*, 76(3), pp.451–462. Available at: <Go to ISI>://WOS:000290314900005.

Kristiansen, R. et al., 2012. A metabolic model for members of the genus Tetrasphaera involved in enhanced biological phosphorus removal. *The ISME Journal*, 7(10), pp.543–554.

Lane, D.J., 1991. 16S/23S rRNA sequencing. In M. Goodfellow & E. Stackebrandt, eds. *Nucleic Acid Techniques in Bacterial Systematics*. Chichester, UK: John Wiley &sons, pp. 115–175. Available at: http://ci.nii.ac.jp/naid/10018533690/en/.

LaPara, T.M. et al., 2015. The Bacterial Communities of Full-Scale Biologically Active, Granular Activated Carbon Filters Are Stable and Diverse and Potentially Contain Novel Ammonia-Oxidizing Microorganisms. *Applied and environmental microbiology*, 81(19), pp.6864–72. Available at: http://aem.asm.org/content/81/19/6864.short [Accessed February 12, 2016].

Larkin, M. et al., 2007. ClustalW and ClustalX version 2. *Bioinformatics*, 23(21), pp.2947–2948.

Lavender, J.S. & Kinzelman, J.L., 2009. A cross comparison of QPCR to agar-based or defined substrate test methods for the determination of Escherichia coli and enterococci in municipal water quality monitoring programs. *Water Research*, 43(19), pp.4967–4979.

Leclerc, H., Devriese, L.A. & Mossel, D.A.A., 1996. Taxonomical changes in intestinal (faecal) enterococci and streptococci: consequences on their use as indicators of faecal contamination in drinking water. *Journal of Applied Bacteriology*, 81(5), pp.459–466. Available at: http://doi.wiley.com/10.1111/j.1365-2672.1996.tb03533.x [Accessed September 5, 2016].

Lee, C., Marion, J.W. & Lee, J., 2013. Development and application of a quantitative PCR assay

targeting Catellicoccus marimammalium for assessing gull-associated fecal contamination at Lake Erie beaches. *The Science of the total environment*, 454–455, pp.1–8. Available at: http://www.sciencedirect.com/science/article/pii/S0048969713002854 [Accessed April 1, 2016].

Lee, C.S. & Lee, J., 2010. Evaluation of new gyrB-based real-time PCR system for the detection of B. fragilis as an indicator of human-specific fecal contamination. *Journal of microbiological methods*, 82(3), pp.311–8. Available at: http://www.sciencedirect.com/science/article/pii/S0167701210002435 [Accessed May 27, 2015].

Lee, M. et al., 2005. Synthesis of mycolic acid biosurfactants and their physical and surface-active properties. *Journal of the American Oil Chemists' Society*, 82(3), pp.181–188. Available at: http://link.springer.com/10.1007/s11746-005-5170-8 [Accessed February 16, 2016].

Lemos, L.N. et al., 2011. Rethinking microbial diversity analysis in the high throughput sequencing era. *Journal of Microbiological Methods*, 86(1), pp.42–51.

Lemos, P.C. et al., 2008. Microbial characterisation of polyhydroxyalkanoates storing populations selected under different operating conditions using a cell-sorting RT-PCR approach. *Applied microbiology and biotechnology*, 78(2), pp.351–60. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18193421 [Accessed February 18, 2016].

Liang, L. et al., 2015. Alternative Fecal Indicators and Their Empirical Relationships with Enteric Viruses, Salmonella enterica, and Pseudomonas aeruginosa in Surface Waters of a Tropical Urban Catchment C. A. Elkins, ed. *Applied and Environmental Microbiology*, 81(3), pp.850–860. Available at: http://aem.asm.org/lookup/doi/10.1128/AEM.02670-14 [Accessed September 7, 2016].

Liao, C.S. et al., 2010. Bioremediation of endocrine disruptor di-n-butyl phthalate ester by Deinococcus radiodurans and Pseudomonas stutzeri. *Chemosphere*, 78(3), pp.342–346. Available at: <Go to ISI>://WOS:000274222900019.

Life Technologies, 2012. Ion Amplicon Library Preparation (Fusion Method) for use with : Ion Torrent Personal Genome Machine ® System. , (4468326), pp.1–19.

Life Technologies, 2013a. Ion PGM ™ Sequencing 400 Kit. , (Rev 2.0).

Life Technologies, 2013b. Ion PGM ™ Template OT2 400 Kit. , (Rev 2.0).

Life Technologies, 2014. Torrent Suite.

Life Technologies, 2015. Torrent Suite ™ Tech Bulletins and Whitepapers.

Limpiyakorn, T. et al., 2011. Abundance of amoA genes of ammonia-oxidizing archaea and bacteria in activated sludge of full-scale wastewater treatment plants. *Bioresource technology*, 102(4), pp.3694–701. Available at: http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=19&SID=Y1jPZu1kf3aIQywGLUA&page=1&doc=8 [Accessed February 5, 2016].

Limpiyakorn, T. et al., 2005. Communities of ammonia-oxidizing bacteria in activated sludge of various sewage treatment plants in Tokyo. *FEMS microbiology ecology*, 54(2), pp.205–17. Available at: http://femsec.oxfordjournals.org/content/54/2/205.abstract [Accessed February 19, 2016].

Livak, K.J. et al., 1995. Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. *PCR methods and applications*, 4(6), pp.357–62. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7580930 [Accessed September 18, 2016].

de los Reyes, F.L., 2010. Challenges in determining causation in structure-function studies using molecular biological techniques. *Water Research*, 44(17), pp.4948–4957. Available at: <Go to ISI>://WOS:000283910600014.

de los Reyes, F.L. et al., 1997. Group-specific small-subunit rRNA hybridization probes to characterize filamentous foaming in activated sludge systems. *Applied and Environmental Microbiology*, 63(3), pp.1107–1117. Available at: <Go to ISI>://WOS:A1997WL14400047.

Lozupone, C. & Knight, R., 2005. UniFrac: A new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology*, 71(12), pp.8228–35. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1317376&tool=pmcentrez&rendertype=abstract [Accessed January 1, 2015].

Lu, J. et al., 2008. Phylogenetic diversity and molecular detection of bacteria in gull feces. *Applied and environmental microbiology*, 74(13), pp.3969–76. Available at: http://aem.asm.org/content/74/13/3969.abstract [Accessed April 1, 2016].

Lücker, S. et al., 2014. Nitrotoga-like bacteria are previously unrecognized key nitrite oxidizers in full-scale wastewater treatment plants. *The ISME Journal*, 9(3), pp.708–720. Available at: http://www.nature.com/doifinder/10.1038/ismej.2014.158 [Accessed January 7, 2016].

Lukacik, P. et al., 2012. Structural engineering of a phage lysin that targets Gram-negative pathogens. *Proceedings of the National Academy of Sciences of the United States of America*, 109(25), pp.9857–9862. Available at: <Go to ISI>://WOS:000306061400045.

Lukjancenko, O., Wassenaar, T.M. & Ussery, D.W., 2010. Comparison of 61 sequenced Escherichia coli genomes. *Microbial ecology*, 60(4), pp.708–20. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20623278 [Accessed September 23, 2016].

Madigan, M.T. & Brock, T.D., 2012. *Brock biology of microorganisms*, Pearson.

Maixner, F. et al., 2006. Nitrite concentration influences the population structure of Nitrospira-like bacteria. *Environmental microbiology*, 8(8), pp.1487–95. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16872410 [Accessed February 9, 2016].

Mancini, N. et al., 2010. The era of molecular and other non-culture-based methods in diagnosis of sepsis. *Clinical microbiology reviews*, 23(1), pp.235–51. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20065332 [Accessed September 19, 2016].

Marrengane, Z. et al., 2011. Rapid quantification and analysis of genetic diversity among Gordonia populations in foaming activated sludge plants. *Journal of Basic Microbiology*, 51(4), pp.415–423. Available at: <Go to ISI>://WOS:000294183800008.

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), pp.541–8. Available at: http://www.sciencedirect.com/science/article/pii/S0043135404005378 [Accessed May 29, 2015].

Marti, R. et al., 2011. Effect of oxygen and temperature on the dynamic of the dominant bacterial populations of pig manure and on the persistence of pig-associated genetic markers, assessed in river water microcosms. *Journal of applied microbiology*, 111(5), pp.1159–75. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21854514 [Accessed May 27, 2015].

Martins, A.M.P. et al., 2004. Filamentous bulking sludge - a critical review. *Water Research*, 38(4), pp.793–817. Available at: <Go to ISI>://000189084700001.

McAvoy, D.C. et al., 1998. Removal of alcohol ethoxylates, alkyl ethoxylate sulfates, and linear

alkylbenzene sulfonates in wastewater treatment. *Environmental Toxicology and Chemistry*, 17(9), pp.1705–1711. Available at: http://doi.wiley.com/10.1002/etc.5620170909 [Accessed February 17, 2016].

McDonald, D. et al., 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal*, 6(3), pp.610–618.

McIlroy, S.J. et al., 2014. "Candidatus Competibacter"-lineage genomes retrieved from metagenomes reveal functional metabolic diversity. *The ISME Journal*, 8(3), pp.613–624. Available at: http://www.nature.com/doifinder/10.1038/ismej.2013.162.

Melzak, K.A. et al., 1996. Driving Forces for DNA Adsorption to Silica in Perchlorate Solutions. *Journal of Colloid and Interface Science*, 181(2), pp.635–644. Available at: http://www.sciencedirect.com/science/article/pii/S002197979690421X [Accessed October 27, 2015].

Met Office, 2006. MIDAS: UK Daily Rainfall Data. Available at: http://catalogue.ceda.ac.uk/uuid/c732716511d3442f05cdeccbe99b8f90.

Metcalf & Eddy. et al., 2003. *Wastewater engineering : treatment and reuse* 4th /., Boston: McGraw-Hill.

Mielczarek, A.T. et al., 2012. Population dynamics of filamentous bacteria in Danish wastewater treatment plants with nutrient removal. *Water Research*, 46(12), pp.3781–3795. Available at: http://www.sciencedirect.com/science/article/pii/S0043135412002515.

Mizrahi-Man, O., Davenport, E.R. & Gilad, Y., 2013. Taxonomic Classification of Bacterial 16S rRNA Genes Using Short Sequencing Reads: Evaluation of Effective Study Designs. *PLoS ONE*, 8(1).

Mobarry, B. et al., 1996. Phylogenetic probes for analyzing abundance and spatial organization of nitrifying bacteria [published erratum appears in Appl Environ Microbiol 1997 Feb;63(2):815]. *Appl. Envir. Microbiol.*, 62(6), pp.2156–2162. Available at: http://aem.asm.org.libproxy.ncl.ac.uk/content/62/6/2156 [Accessed February 12, 2016].

Monod, J., 1949. The Growth of Bacterial Cultures. *Annual Review of Microbiology*, 3(1), pp.371–394. Available at: http://www.annualreviews.org/doi/10.1146/annurev.mi.03.100149.002103 [Accessed September 25, 2016].

Moore, W.E.C. et al., 1987. Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. *International Journal of Systematic and Evolutionary Microbiology*, 37(4), pp.463–464. Available at: https://www.researchgate.net/publication/243781041_Report_of_the_Ad_Hoc_Committee_on_Reconciliation_of_Approaches_to_Bacterial_Systematics_Int_J_Syst_Bacteriol [Accessed January 18, 2016].

Morales, S.E. et al., 2009. Extensive phylogenetic analysis of a soil bacterial community illustrates extreme taxon evenness and the effects of amplicon length, degree of coverage, and DNA fractionation on classification and ecological parameters. *Applied and environmental microbiology*, 75(3), pp.668–75. Available at: http://aem.asm.org/content/75/3/668.full [Accessed August 24, 2015].

Morales, S.E. & Holben, W.E., 2009. Empirical testing of 16S rRNA gene PCR primer pairs reveals variance in target specificity and efficacy not suggested by in silico analysis. *Applied and environmental microbiology*, 75(9), pp.2677–83. Available at: http://aem.asm.org/content/75/9/2677.full [Accessed August 24, 2015].

Mori, Y. et al., 2001. Detection of Loop-Mediated Isothermal Amplification Reaction by Turbidity Derived from Magnesium Pyrophosphate Formation. *Biochemical and Biophysical Research Communications*, 289(1), pp.150–154. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0006291X01959212 [Accessed July 25, 2016].

Muller, E.E.L. et al., 2012. Genome sequence of "Candidatus Microthrix parvicella" Bio17-1, a long-chain-fatty-acid-accumulating filamentous actinobacterium from a biological wastewater treatment plant. *Journal of bacteriology*, 194(23), pp.6670–1. Available at: http://jb.asm.org/content/194/23/6670.full [Accessed February 3, 2016].

Murphy, M.A. et al., 2002. An evaluation of long-term preservation methods for brown bear (Ursus arctos) faecal DNA samples. *Conservation Genetics*, 3(4), pp.435–440. Available at: http://link.springer.com/article/10.1023/A%3A1020503330767 [Accessed March 31, 2016].

Muscillo, M. et al., 1997. Enteric virus detection in Adriatic seawater by cell culture, polymerase chain reaction and polyacrylamide gel electrophoresis. *Water Research*, 31(8), pp.1980–1984. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0043135496003314 [Accessed August 2, 2016].

Naidoo, D., Kumari, S. & Bux, F., 2011. Characterization of Nocardia farcinica, a Filamentous Bacterium Isolated from Foaming Activated Sludge Samples. *Water Environment Research*, 83(6), pp.527–531. Available at: <Go to ISI>://WOS:000291209800005.

Narayanan, B. et al., 2010. The Role of Dispersed Nocardioform Filaments in Activated Sludge Foaming. *Water Environment Research*, 82(6), pp.483–491. Available at: <Go to ISI>://WOS:000278013600001.

Navacharoen, A. & Vangnai, A.S., 2011. Biodegradation of diethyl phthalate by an organic-solvent-tolerant Bacillus subtilis strain 3C3 and effect of phthalate ester coexistence. *International Biodeterioration & Biodegradation*, 65(6), pp.818–826. Available at: <Go to ISI>://WOS:000295350500011.

Nawrocki, E.P., Kolbe, D.L. & Eddy, S.R., 2009. Infernal 1.0: Inference of RNA alignments. *Bioinformatics*, 25(10), pp.1335–1337.

Newton, R.J. et al., 2013. A Microbial Signature Approach to Identify Fecal Pollution in the Waters Off an Urbanized Coast of Lake Michigan. *Microbial Ecology*, 65(4), pp.1011–1023.

Nielsen, P.H., Kragelund, C., et al., 2009. Identity and ecophysiology of filamentous bacteria in activated sludge. *Fems Microbiology Reviews*, 33(6), pp.969–998. Available at: <Go to ISI>://000270397600001.

Nielsen, P.H., 2017. Microbial biotechnology and circular economy in wastewater treatment. *Microbial Biotechnology*, 10(5), pp.1102–1105.

Nielsen, P.H., Daims, H. & Lemmer, H., 2009. FISH Handbook for Biological Wastewater Treatment.

Nittami, T. et al., 2009. Candidatus Monilibacter spp., common bulking filaments in activated sludge, are members of Cluster III Defluviicoccus. *Systematic and applied microbiology*, 32(7), pp.480–9. Available at: http://www.sciencedirect.com/science/article/pii/S0723202009000873 [Accessed February 18, 2016].

Northumbrian Water, 2011. 2011 June return.

Notomi, T. et al., 2000. Loop-mediated isothermal amplification of DNA. *Nucleic Acids Research*, 28(12). Available at: <Go to ISI>://000207392200007.

Oliver, D.M. et al., 2014. Opportunities and limitations of molecular methods for quantifying microbial compliance parameters in EU bathing waters. *Environment International*, 64, pp.124–128. Available at: http://dx.doi.org/10.1016/j.envint.2013.12.016.

Oshiro, R.K., 2010. Method B : Bacteroidales in Water by TaqMan® Quantitative Polymerase Chain Reaction (qPCR) Assay. , p.111.

Oyafuso, Z.S. et al., 2015. Widespread detection of human- and ruminant-origin *Bacteroidales* markers in subtidal waters of the Salish Sea in Washington State. *Journal of Water and Health*, 13(3), p.827. Available at: http://jwh.iwaponline.com/cgi/doi/10.2166/wh.2015.253 [Accessed July 25, 2016].

Palfrey, R., 2015. UV disinfection - developments and opportunities. *Water and sewerage journal*.

Park, H.-D. & Noguera, D.R., 2004. Evaluating the effect of dissolved oxygen on ammonia-oxidizing bacterial communities in activated sludge. *Water research*, 38(14–15), pp.3275–86. Available at: http://www.sciencedirect.com/science/article/pii/S0043135404002404 [Accessed December 24, 2015].

Parliament, H. of, 2010. Environmental Permitting (England and Wales) Regulations 2010 H. of Parliament, ed. , 675.

Parliamentary Office of Science and Technology, 2014. *Phosphate Resources*,

Patel, R. et al., 1998. Determination of 16S rRNA sequences of enterococci and application to species identification of nonmotile Enterococcus gallinarum isolates. *Journal of clinical microbiology*, 36(11), pp.3399–407. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9774606 [Accessed September 23, 2016].

Petrovski, S. et al., 2011. An examination of the mechanisms for stable foam formation in activated sludge systems. *Water Research*, 45(5), pp.2146–2154. Available at: <Go to ISI>://WOS:000288478700024.

Petrovski, S., Tillett, D. & Seviour, R.J., 2012. Genome Sequences and Characterization of the Related Gordonia Phages GTE5 and GRU1 and Their Use as Potential Biocontrol Agents. *Applied and Environmental Microbiology*, 78(1), pp.42–47. Available at: <Go to ISI>://WOS:000298396700007.

Pickering, R.L., 2008. *How Hard is the Biomass Working?* Thesis. Newcastle upon Tyne, UK: Newcastle University.

Pina, S. et al., 1998. Viral pollution in the environment and in shellfish: human adenovirus detection by PCR as an index of human viruses. *Applied and environmental microbiology*, 64(9), pp.3376–82. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9726885 [Accessed August 2, 2016].

Pinto, A.J. et al., 2016. Metagenomic Evidence for the Presence of Comammox Nitrospira -Like Bacteria in a Drinking Water System S. J. Hallam, ed. *mSphere*, 1(1), pp.e00054-15. Available at: http://msphere.asm.org/content/1/1/e00054-15.abstract [Accessed February 12, 2016].

Pomiès, M. et al., 2015. Lab-scale experimental strategy for determining micropollutant partition coefficient and biodegradation constants in activated sludge. *Environmental science and pollution research international*, 22(6), pp.4383–95. Available at: http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=10&SID=S2wj766ttzi9aAqdmBp&page=1&doc=9 [Accessed February 19, 2016].

Prakasam, T.B. & Dondero, N.C., 1970. AEROBIC HETEROTROPHIC BACTERIAL POPULATIONS OF SEWAGE AND ACTIVATED SLUDGE .5. ANALYSIS OF POPULATION STRUCTURE AND ACTIVITY. *Applied microbiology*, 19(4), p.671-. Available at: <Go to ISI>://WOS:A1970G054100022.

Price, M.N., Dehal, P.S. & Arkin, A.P., 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular biology and evolution*, 26(7), pp.1641–50. Available at: http://mbe.oxfordjournals.org/content/26/7/1641.short [Accessed November 5, 2014].

Price, M.N., Dehal, P.S. & Arkin, A.P., 2010. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE*, 5(3).

Prosser, J.I. & Prosser I., J., 1989. *Autotrophic nitrification in bacteria*, Available at: http://www.ncbi.nlm.nih.gov/pubmed/2700538.

Pruesse, E. et al., 2007. SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, 35(21), pp.7188–7196.

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), pp.10699–10705. Available at: http://pubs.acs.org/doi/abs/10.1021/es202141x [Accessed August 2, 2016].

Quast, C. et al., 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research*, 41(Database issue), pp.D590-6. Available at: http://nar.oxfordjournals.org/content/41/D1/D590.full [Accessed July 10, 2014].

Quince, C. et al., 2011. Removing noise from pyrosequenced amplicons. *BMC bioinformatics*, 12, p.38.

Reischer, G.H. et al., 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), pp.5610–4. Available at: http://aem.asm.org/content/72/8/5610.full [Accessed April 25, 2015].

Ridley, C.M. et al., 2014. Baseline and storm event monitoring of Bacteroidales marker concentrations and enteric pathogen presence in a rural Canadian watershed. *Water Research*, 60, pp.278–288. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0043135414003248 [Accessed July 25, 2016].

Rittmann, B.E. et al., 1999. Molecular and modeling analyses of the structure and function of nitrifying activated sludge. In *Water Science and Technology*. pp. 51–59. Available at: http://www.sciencedirect.com/science/article/pii/S0273122398007756 [Accessed February 12, 2016].

Roesch, L.F.W. et al., 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. *The ISME journal*, 1(4), pp.283–290.

Roslev, P. & Bukh, A.S., 2011. State of the art molecular markers for fecal pollution source tracking in water. *Applied Microbiology and Biotechnology*, 89(5), pp.1341–1355. Available at: <Go to ISI>://000287199900009.

Rossetti, S. et al., 2005. “Microthrix parvicella”, a filamentous bacterium causing bulking and foaming in activated sludge systems: a review of current knowledge. *Fems Microbiology Reviews*, 29(1), pp.49–64. Available at: <Go to ISI>://WOS:000226703200004.

Rostron, W. & Holt, D., 2017. United Utilities CIP2 &P1 Trials. *Water Projects Online*.

Rothberg, J.M. et al., 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356), pp.348–352. Available at: http://dx.doi.org/10.1038/nature10242.

Rotthauwe, J., Witzel, K. & Liesack, W., 1997. The ammonia monooxygenase structural gene amoA as a functional marker: Molecular fine-scale analysis of natural ammonia-oxidizing populations. *APPLIED AND ENVIRONMENTAL MICROBIOLOGY*, 63(12), pp.4704–4712. Available at: http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=21&SID=Y2uO2WvxMuwqrjMRdpM&page=1&doc=1 [Accessed February 12, 2016].

Sang, Y.Y., Xiong, G.M. & Maser, E., 2011. Steroid degradation and two steroid-inducible enzymes in the marine bacterium H5. *Chemico-Biological Interactions*, 191(1–3), pp.89–94. Available at: <Go to ISI>://WOS:000291767200016.

Saunders, A.M. et al., 2015. The activated sludge ecosystem contains a core community of abundant organisms. *The ISME journal*, 10(1), pp.11–20. Available at: http://dx.doi.org/10.1038/ismej.2015.117 [Accessed December 15, 2015].

Schloss, P.D. et al., 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75(23), pp.7537–7541.

Schrader, C. et al., 2012. PCR inhibitors - occurrence, properties and removal. *Journal of Applied Microbiology*, 113(5), pp.1014–1026. Available at: http://doi.wiley.com/10.1111/j.1365-2672.2012.05384.x [Accessed June 17, 2016].

Scott, T.M. et al., 2002. Microbial Source Tracking: Current Methodology and Future Directions. *Applied and Environmental Microbiology*, 68(12), pp.5796–5803. Available at: http://aem.asm.org/cgi/doi/10.1128/AEM.68.12.5796-5803.2002 [Accessed September 5, 2016].

Seedorf, H., Kittelmann, S. & Janssen, P.H., 2015. Few Highly Abundant Operational Taxonomic Units Dominate within Rumen Methanogenic Archaeal Species in New Zealand Sheep and Cattle G. Voordouw, ed. *Applied and Environmental Microbiology*, 81(3), pp.986–995. Available at: http://aem.asm.org/lookup/doi/10.1128/AEM.03018-14 [Accessed September 7, 2016].

Sentencing Council, 2014. Environmental Offences Definitive Guideline. Available at: http://sentencingcouncil.judiciary.gov.uk/docs/Final_Environmental_Offences_Definitive_Guideline_(web).pdf.

Severn Trent, 2008. *How ammonia controls are reducing energy consumption at Severn Trent Water's sewage treatment works*,

Severn Trent, 2016. Severn Trent trials six technologies to reduce phosphorus. *Utility Week*. Available at: http://utilityweek.co.uk/news/severn-trent-trials-six-technologies-to-reduce-phosphorus/1240692#.V90cDPkrLIU [Accessed September 17, 2016].

Seviour, R., Nielsen, P.H. & Robert Seviour, P.H.N., 2010. Microbial ecology of activated sludge.

Seviour, R.J. et al., 2008. Ecophysiology of the Actinobacteria in activated sludge systems. *Antonie Van Leeuwenhoek International Journal of General and Molecular Microbiology*, 94(1), pp.21–33. Available at: <Go to ISI>://WOS:000256754000004.

Seyrig, G. et al., 2011. Detection and Occurrence of Indicator Organisms and Pathogens. *Water Environment Research*, 83(10), pp.900–926. Available at:

http://www.ingentaconnect.com/content/wef/wer/2011/00000083/00000010/art00001?token=00521c12e337643f43a47e41225f40384d2c4b6754766b3c245049576b34272c5f7b3d6d3f4e4b34cc.

Sezgin, M., Jenkins, D. & Parker, D.S., 1978. UNIFIED THEORY OF FILAMENTOUS ACTIVATED-SLUDGE BULKING. *Journal Water Pollution Control Federation*, 50(2), pp.362–381. Available at: <Go to ISI>://WOS:A1978EM13000015.

Shah, V.G. et al., 2007. Evaluating potential applications of faecal sterols in distinguishing sources of faecal contamination from mixed faecal samples. *Water Research*, 41(16), pp.3691–3700.

Shanks, O.C. et al., 2011. Community Structures of Fecal Bacteria in Cattle from Different Animal Feeding Operations. *Applied and Environmental Microbiology*, 77(9), pp.2992–3001. Available at: <Go to ISI>://000289773100022 [Accessed June 21, 2016].

Shanks, O.C. et al., 2013. Comparison of the Microbial Community Structures of Untreated Wastewaters from Different Geographic Locales. *Applied and Environmental Microbiology*, 79(9), pp.2906–2913. Available at: http://aem.asm.org/cgi/doi/10.1128/AEM.03448-12 [Accessed September 7, 2016].

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), pp.53–64.

Siripong, S. & Rittmann, B.E., 2007. Diversity study of nitrifying bacteria in full-scale municipal wastewater treatment plants. *Water Research*, 41(5), pp.1110–1120. Available at: http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=21&SID=Y2uO2WvxMuwqrjMRdpM&page=1&doc=3 [Accessed February 12, 2016].

Smyth, R.P. et al., 2010. Reducing chimera formation during PCR amplification to ensure accurate genotyping. *Gene*, 469(1–2), pp.45–51. Available at: http://www.sciencedirect.com/science/article/pii/S0378111910003616 [Accessed June 11, 2015].

Soares, A., 2008. New flow-sheets for sewage treatment with reduced energy requirements. In *SWIG*.

Soddell, J.A. et al., 2006. Gordonia defluvii sp. nov., an actinomycete isolated from activated sludge foam. *International journal of systematic and evolutionary microbiology*, 56(Pt 10), pp.2265–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17012545 [Accessed February 18, 2016].

Soddell, J.A. & Seviour, R.J., 1990. Microbiology of Foaming in Activated-Sludge Plants. *Journal of Applied Bacteriology*, 69(2), pp.145–176. Available at: <Go to ISI>://A1990DT05800001.

Soergel, D. a W. et al., 2012. Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *The ISME Journal*, 6(7), pp.1440–1444.

Solecki, O. et al., 2011. Persistence of microbial and chemical pig manure markers as compared to faecal indicator bacteria survival in freshwater and seawater microcosms. *Water Research*, 45(15), pp.4623–4633. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0043135411003460 [Accessed July 25, 2016].

Song, S.J. et al., 2013. Cohabiting family members share microbiota with one another and with their dogs. *eLife*, 2, p.e00458. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/23599893 [Accessed June 21, 2016].

Sonnenschein, C. & Soto, A.M., 1998. An updated review of environmental estrogen and androgen mimics and antagonists. *Journal of Steroid Biochemistry and Molecular Biology*, 65(1–6), pp.143–150. Available at: <Go to ISI>://WOS:000074830800014.

Stamatakis, A., 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21), pp.2688–2690.

Stoeckel, D.M. & Harwood, V.J., 2007. Performance, design, and analysis in microbial source tracking studies. *Applied and Environmental Microbiology*, 73(8), pp.2405–2415. Available at: <Go to ISI>://000246542400001.

Tan, B. et al., 2015. Next-generation sequencing (NGS) for assessment of microbial water quality: Current progress, challenges, and future opportunities. *Frontiers in Microbiology*, 6(SEP).

Tanner, N.A., Zhang, Y. & Evans, T.C., 2012. Simultaneous multiple target detection in real-time loop-mediated isothermal amplification. *BioTechniques*, 53(2), pp.81–89.

Tchelet, R. et al., 1999. Population dynamics of an introduced bacterium degrading chlorinated benzenes in a soil column and in sewage sludge. *Biodegradation*, 10(2), pp.113–125. Available at: <Go to ISI>://WOS:000081060000004.

Thompson, C.L., Wang, B. & Holmes, A.J., 2008. The immediate environment during postnatal development has long-term impact on gut community structure in pigs. *The ISME Journal*, 2(7), pp.739–748. Available at: http://www.nature.com/doifinder/10.1038/ismej.2008.29 [Accessed June 21, 2016].

Tolvanen, K.E.S. & Karp, M.T., 2011. Molecular methods for characterizing mixed microbial communities in hydrogen-fermenting systems. *International Journal of Hydrogen Energy*, 36(9), pp.5280–5288. Available at: <Go to ISI>://WOS:000290883100008.

Tran, N.H., Gin, K.Y.-H. & Ngo, H.H., 2015. Fecal pollution source tracking toolbox for identification, evaluation and characterization of fecal contamination in receiving urban surface waters and groundwater. *Science of The Total Environment*, 538, pp.38–57. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0048969715304964 [Accessed September 24, 2016].

Trevors, J.T., 2012. Can dead bacterial cells be defined and are genes expressed after cell death? *Journal of Microbiological Methods*, 90(1), pp.25–28.

Ulrich, L.E. & Zhulin, I.B., 2010. The MiST2 database: a comprehensive genomics resource on microbial signal transduction. *Nucleic acids research*, 38(Database issue), pp.D401-7. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2808908&tool=pmcentrez&rendertype=abstract [Accessed February 3, 2016].

United States Environment Protection agency, 2012. Method 1611: Enterococci in Water by TaqMan ® Quantitative Polymerase Chain Reaction (qPCR) Assay.

Vannelli, T. et al., 1990. Degradation of halogenated aliphatic compounds by the ammonia-oxidizing bacterium Nitrosomonas europaea. *Applied and environmental microbiology*, 56(4), pp.1169–1171.

Versalovic, J. et al., 1994. Genomic fingerprinting of bacteria using repetitive sequence-based polymerase chain reaction. *Methods in molecular and cellular biology*, 5, pp.25–40. Available at: http://www.csa.com/partners/viewrecord.php?requester=gs&collection=ENV&recid=3649417.

Vos, J.G. et al., 2000. Health effects of endocrine-disrupting chemicals on wildlife, with special reference to the European situation. *Critical Reviews in Toxicology*, 30(1), pp.71–133. Available at: <Go to ISI>://WOS:000085738100002.

Wade, T.J. et al., 2003. 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), pp.1102–1109.

Wagner, A.O. et al., 2008. Removal of free extracellular DNA from environmental samples by ethidium monoazide and propidium monoazide. *Applied and environmental microbiology*, 74(8), pp.2537–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18296534 [Accessed September 23, 2016].

Wagner, M. et al., 1996. analysis of nitrifying bacteria in sewage treatment plants. *Water Science and Technology*, 34(1–2), pp.237–244. Available at: http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=5&SID=P2SKMkNOKSzLTmBnkxV&page=2&doc=97 [Accessed February 16, 2016].

Wagner, M. & Loy, A., 2002. Bacterial community composition and function in sewage treatment systems. *Current Opinion in Biotechnology*, 13(3), pp.218–227. Available at: <Go to ISI>://000175903900005.

Waites, D., 2012. Discussions regarding the value to Northumbrian Water Group of NGS.

Wallis, J.L. & Taylor, H.D., 2003. Phenotypic population characteristics of the enterococci in wastewater and animal faeces: implications for the new European directive on the quality of bathing waters. *Water Science and Technology*, 47(3), pp.27–32.

Wang, G.C. & Wang, Y., 1997. Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. *Applied and environmental microbiology*, 63(12), pp.4645–50. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9406382 [Accessed September 19, 2016].

Wang, Q. et al., 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73(16), pp.5261–5267. Available at: <Go to ISI>://WOS:000248825900024.

Water Services Association, 1998. *Water Facts*, London: Water Services Association.

Wei, Y., Liu, J. & Qu, D., 2011. Recent advances in microbial degradation of ibuprofen-A review. *Weishengwu Xuebao*, 51(5), pp.586–594. Available at: <Go to ISI>://BCI:BCI201100367172.

Wells, G.F. et al., 2009. Ammonia-oxidizing communities in a highly aerated full-scale activated sludge bioreactor: betaproteobacterial dynamics and low relative abundance of Crenarchaea. *Environmental microbiology*, 11(9), pp.2310–28. Available at: http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=19&SID=Y1jPZu1kf3aIQywGLUA&page=1&doc=4 [Accessed February 18, 2016].

Werner, J.J. et al., 2012. Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. *The ISME Journal*, 6(1), pp.94–103.

Whitman, R.L. et al., 2004. Solar and Temporal Effects on Escherichia coli Concentration at a Lake Michigan Swimming Beach. *Applied and Environmental Microbiology*, 70(7), pp.4276–4285. Available at: http://aem.asm.org/cgi/doi/10.1128/AEM.70.7.4276-4285.2004 [Accessed September 3, 2016].

Whitman, R.L. & Nevers, M.B., 2004. Escherichia coli Sampling Reliability at a Frequently Closed Chicago Beach: Monitoring and Management Implications. *Environmental Science*

& Technology, 38(16), pp.4241–4246. Available at: http://pubs.acs.org/doi/abs/10.1021/es034978i [Accessed September 5, 2016].

Wiggins, B.A. et al., 2003. Use of Antibiotic Resistance Analysis for Representativeness Testing of Multiwatershed Libraries. *Applied and Environmental Microbiology*, 69(6), pp.3399–3405. Available at: http://aem.asm.org/cgi/doi/10.1128/AEM.69.6.3399-3405.2003 [Accessed July 26, 2016].

William, R.S. & Ronald, W.M., 1982. *Chemical Reaction Equilibrium Analysis: Theory and Algorithms* First., Wiley. Available at: http://www.amazon.com/Chemical-Reaction-Equilibrium-Analysis-William/dp/0471093475%3FSubscriptionId%3D0JRA4J6WAV0RTAZVS6R2%26tag%3Dworldcat-20%26linkCode%3Dxm2%26camp%3D2025%26creative%3D165953%26creativeASIN%3D0471093475 [Accessed May 7, 2015].

Withey, S. et al., 2005. Bacteriophages - potential for application in wastewater treatment processes. *Science of the Total Environment*, 339(1–3), pp.1–18. Available at: <Go to ISI>://000227915700001.

Woese, C.R., 1987. Bacterial evolution. *Microbiological reviews*, 51(2), pp.221–71. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2439888 [Accessed September 17, 2016].

Woese, C.R. et al., 1975. Conservation of primary structure in 16S ribosomal RNA. *Nature*, 254(5495), pp.83–86. Available at: http://dx.doi.org/10.1038/254083a0 [Accessed February 10, 2016].

Woese, C.R. et al., 1984. The phylogeny of purple bacteria: The beta subdivision. *Systematic and Applied Microbiology*, 5(3), pp.327–336.

Woese, C.R., Fox, G.E. & Pechman, K.R., 1977. Comparative Cataloging of 16S Ribosomal Ribonucleic Acid: Molecular Approach to Procaryotic Systematics. *International Journal of Systematic and Evolutionary Microbiology*, 27(1), pp.44–57. Available at: http://www.microbiologyresearch.org/content/journal/ijsem/10.1099/00207713-27-1-44 [Accessed September 17, 2016].

Wuertz, S. & Reis, M., 2010. Editorial for special issue on shifting paradigms in assessment of recreational water quality. *Water Research*, 44(16).

Wuertz, S. & Reis, M.A.M., 2013. Editorial: Microbial source tracking. *Water Research*, 47(18).

Wyer, M.D. et al., 2012. Relationships between human adenoviruses and faecal indicator organisms in European recreational waters. *Water Research*, 46(13), pp.4130–4141.

Ye, L. et al., 2012. Microbial structures, functions, and metabolic pathways in wastewater treatment bioreactors revealed using high-throughput sequencing. *Environmental science & technology*, 46(24), pp.13244–52. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23151157.

Yoon, D.N. et al., 2010. Isolation, characterization, and abundance of filamentous members of Caldilineae in activated sludge. *Journal of Microbiology*, 48(3), pp.275–283. Available at: <Go to ISI>://WOS:000279087800002.

Yu, K. & Zhang, T., 2012. Metagenomic and metatranscriptomic analysis of microbial community structure and gene expression of activated sludge. *PloS one*, 7(5), p.e38183. Available at: http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=21&SID=Y2uO2WvxMuwqrjMRdpM&page=1&doc=14 [Accessed February 12, 2016].

Zang, K. et al., 2008. Analysis of the phylogenetic diversity of estrone-degrading bacteria in activated sewage sludge using microautoradiography-fluorescence in situ hybridization. *Systematic and Applied Microbiology*, 31(3), pp.206–214. Available at: <Go to ISI>://WOS:000258753200005.

Zhang, J. et al., 2011. Rhodanobacter xiangquanii sp. nov., a Novel Anilofos-Degrading Bacterium Isolated from a Wastewater Treating System. *Current Microbiology*, 62(2), pp.645–649. Available at: <Go to ISI>://000286205500048.

Zhang, T., Shao, M.F. & Ye, L., 2012. 454 Pyrosequencing reveals bacterial diversity of activated sludge from 14 sewage treatment plants. *Isme Journal*, 6(6), pp.1137–1147. Available at: <Go to ISI>://WOS:000304047800006.

Ziembinska, A. et al., 2007. Molecular analysis of temporal changes of a bacterial community structure in activated sludge using denaturing gradient gel electrophoresis (DGGE) and fluorescent in situ hybridization (FISH). *Polish Journal of Microbiology*, 56(2), pp.119–127. Available at: <Go to ISI>://BCI:BCI200700565144.

Chapter 13. Appendix

13.1 Protocols

13.1.1 Manufacturer's protocol for Ion Torrent instruments

The Ion Torrent OneTouch 2 protocol used was titled "Ion PGM Templated OT2 400 kit", Catalogue number 4479878, publication number MAN0007218, revision A.0. The Ion Torrent Personal Genome Machine protocol used was titled "Ion PGM Sequencing 400 Kit", Catalogue number 4482002, publication number MAN0007242, revision 2.0

13.1.2 Manufacturer's standard DNA extraction protocol for the FastDNA Spin Kit for Soil

1. Add up to 500 mg of soil sample to a Lysing Matrix E tube.
2. Add 978 µl Sodium Phosphate Buffer to sample in Lysing Matrix E tube.
3. Add 122 µl MT Buffer.
4. Homogenize in the FastPrep® Instrument for 40 seconds at a speed setting of 6.0.
5. Centrifuge at 14,000 x g for 5-10 minutes to pellet debris.

NOTE: Extending centrifugation to 15 minutes can enhance elimination of excessive debris from large samples, or from cells with complex cell walls.

6. Transfer supernatant to a clean 2.0 ml microcentrifuge tube. Add 250 µl PPS (Protein Precipitation Solution) and mix by shaking the tube by hand 10 times.
7. Centrifuge at 14,000 x g for 5 minutes to pellet precipitate. Transfer supernatant to a clean 15 ml tube. NOTE: While a 2.0 ml microcentrifuge tube may be used at this step, better mixing and DNA binding will occur in a larger tube.
8. Resuspend Binding Matrix suspension and add 1.0 ml to supernatant in 15 ml tube.
9. Place on rotator or invert by hand for 2 minutes to allow binding of DNA. Place tube in a rack for 3 minutes to allow settling of silica matrix.
10. Remove and discard 500 µl of supernatant being careful to avoid settled Binding Matrix.
11. Resuspend Binding Matrix in the remaining amount of supernatant. Transfer approximately 600 µl of the mixture to a SPIN™ Filter and centrifuge at 14,000 x g for 1 minute. Empty the catch tube and add the remaining mixture to the SPIN™ Filter and centrifuge as before. Empty the catch tube again.
12. Add 500 µl prepared SEWS-M and gently resuspend the pellet using the force of the
liquid from the pipet tip.

NOTE: Ensure that ethanol has been added to the Concentrated SEWS-M.

13. Centrifuge at 14,000 x g for 1 minute. Empty the catch tube and replace.
14. Without any addition of liquid, centrifuge a second time at 14,000 x g for 2 minutes to "dry" the matrix of residual wash solution. Discard the catch tube and replace with a new, clean catch tube.
15. Air dry the SPIN™ Filter for 5 minutes at room temperature.
16. Gently resuspend Binding Matrix (above the SPIN filter) in 50-100 µl of DES (DNase/ Pyrogen-Free Water).

NOTE: To avoid over-dilution of the purified DNA, use the smallest amount of DES required to resuspend Binding Matrix pellet.

NOTE: Yields may be increased by incubation for 5 minutes at 55˚C in a heat block or water bath.

17. Centrifuge at 14,000 x g for 1 minute to bring eluted DNA into the clean catch tube. Discard the SPIN filter. DNA is now ready for PCR and other downstream applications. Store at -20°C for extended periods or 4°C until use.

13.1.3 Ratio of Ampure XP to PCR products used

Ampure XP (Beckman Coulter, Brea USA) is a suspension of paramagnetic beads used to purify DNA. The volume ratio of Ampure XP to DNA determines the length of DNA fragment which binds to the beads. The manufacturer states a ratio of 1.8 will bind fragments longer than 100bp. Empirical observation of the fragment length bound at different ratios showed that a ratio of 1.1 binds fragments longer than 300bp. Given that the target PCR fragment size for the V4V5 fusion primer pair detailed in Table 7was 476bp the ratio of 1.1 was selected.

13.1.4 Total cell counts

Sample preparation

1) Add 1 µl of sample to 949 µl of filter sterile 1x PBS

2) Add 50 µl of SYBR gold (diluted 100X in filter sterile PBS) to each sample.

3) SYBR gold is light sensitive, so wrap samples in foil and incubate at room temp. for 30 min.

4) Controls:

Ethanol control - if samples were originally harvested into filter sterile ethanol, take 1 ml ethanol and add 50 µl SYBR gold (100 X diluted), wrap in foil and incubate for 30 min.

1 X PBS control – Take 1 ml filter sterile PBS and add 50 µl SYBR gold (100 X diluted), wrap in foil and incubate for 30 min.

Non-SYBR gold stained cells can also be used as a control measure.

Filtration

1) Use sterile Millipore filter holder

2) Unscrew filter holder and, with forceps, aseptically transfer a 13mm membrane filter into base of filter cradle, then reassemble.

3) Sit filter holder on bung of volumetric flask.

4) Switch on vacuum pump, shake sample and transfer into filter holder.

5) Allow sample to be drawn onto the filter, wash with 3 x 1 ml filter sterile water. Once all solutions have been drawn through the filter, leave vacuum pump running to dry filter for 2-3 min.

6) Place a drop of Citifluor onto a microscope slide.

7) Unscrew filter holder, and with forceps, aseptically transfer the membrane filter on to the drop of Citifluor.

8) Place another drop of Citifluor on top of the membrane and place a coverslip over the membrane filter.

9) Place slide in the dark to stop fluorescence from fading.

Microscope and Digital Camera set-up

1) Complete log book and booking sheet before commencing work on the Olympus BX40 Epi-fluorescence microscope.

2) Switch on Olympus U-RFL-T box to right hand side of microscope.

3) Remove dust cover from microscope.

4) Olympus digital camera (E-400), OM adapter (MF-1), battery charger and USB cable stored in locked cupboard within the microbiology lab.

5) Remove the lens from the camera E400 and twist the OM adapter (MF-1) into place.

6) Carefully attach the camera onto the microscope mount so that the controls of the camera are facing you.

7) Switch the camera to the “on” position. Turn the mode dial to the “S” position (shutter speed/exposure time). Set shutter speed to 1” (1 sec) in the first instance, and adjust as necessary to a quicker/slower shutter speed using the dial situated next to the mode dial.

(As a general rule, cells fluorescing brightly will require less exposure time than dull cells)

8) Camera flash should be switched off for fluorescence work.

9) The camera can be set to a 2-second delay to avoid the camera shaking when taking a picture.

10) View slides using 100x oil immersion (UPlanFl 100x lens), under blue light filter (WB) in the dark.

11) Focus the image down the microscope lens, and take a picture.

12) When finished taking images download the pictures onto a PC, using the USB cable.

13) Remember to delete the pictures from the camera after use.

14) Dismantle the camera set-up. Switch off microscope and power supplies. Return camera and all camera equipment back into the locked cupboard; ensure the camera is switched off. If red battery symbol is showing on the camera, recharge the battery for the next user.

15) Remember to put the dust cover over the microscope when finished, be careful not to cover the lamp at the back of the microscope as it will be hot and melt the plastic

covering.

Calculating cells per ml of original sample

1) Calculate average number of cells per field of view, with error

2) Calculate no. of fields of view on the filter membrane e.g.

Diameter of membrane = 9.5 mm, Radius of membrane (r) = 4.75 mm

Therefore, area of membrane = $\pi r2$ = 70.88 mm^2

(NB: each graduation on the stage micrometre is 10 µM)

Area of field of view (calculated using image of stage micrometre) = 100 µM x 120 µM = 0.1 mm x 0.12 mm = 0.012 mm^2

Area of membrane 70.88 mm^2 / Area of fields of view 0.012 mm^2 = 5906.67

3) Multiply average number of cells per field of view by 5906.67

4) Multiply result of 3) by dilution factor (1000X)

5) Multiply result of 4) by original sample dilution factor (samples were originally stored in a 50/50 mix of sample and EtOH – multiply by 2).

6) Result is cells per ml of original sample.

13.2 Additional data from qPCR/FISH/PGM comparison

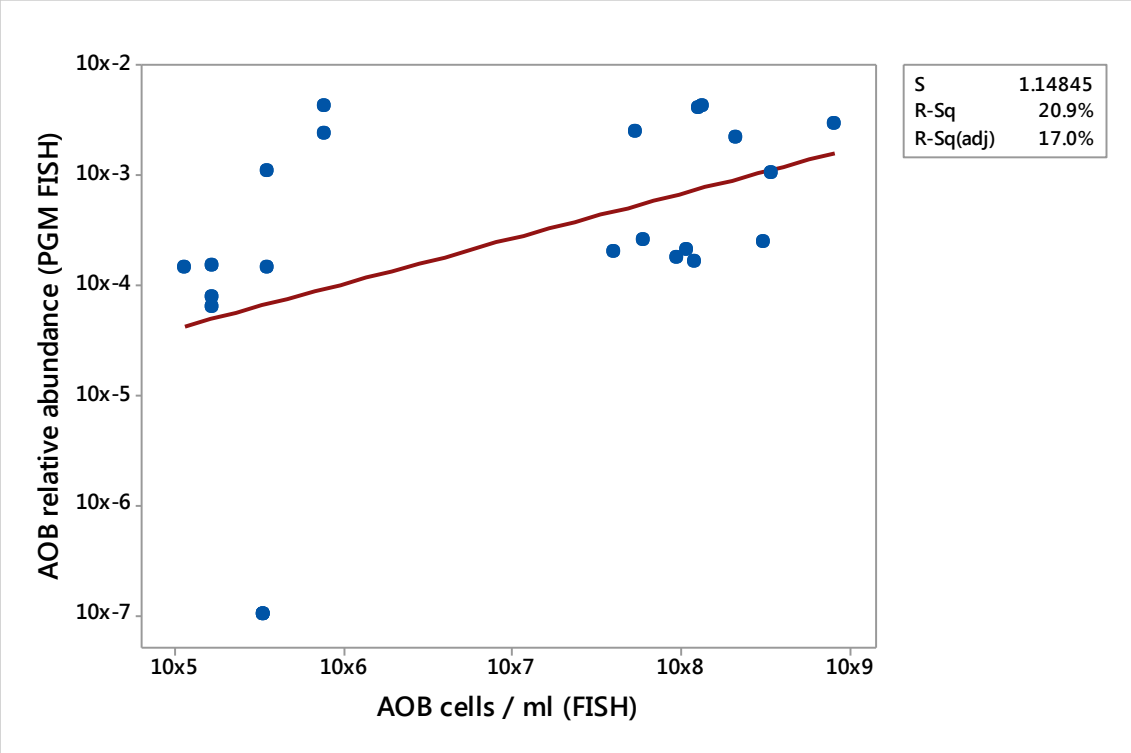


Figure 42: Abundance of ammonia oxidising bacteria measured by fluorescence in situ hybridization (FISH) and Personal Genome Machine (PGM) where only those taxa targeted by FISH probes Neu, NSO1225 and 6a192 were included in PGM data. S denotes the regression standard error. Regression equation: log10(PGM r.a. (FISH)) = -6.436 + 0.4098 log10(FISH)

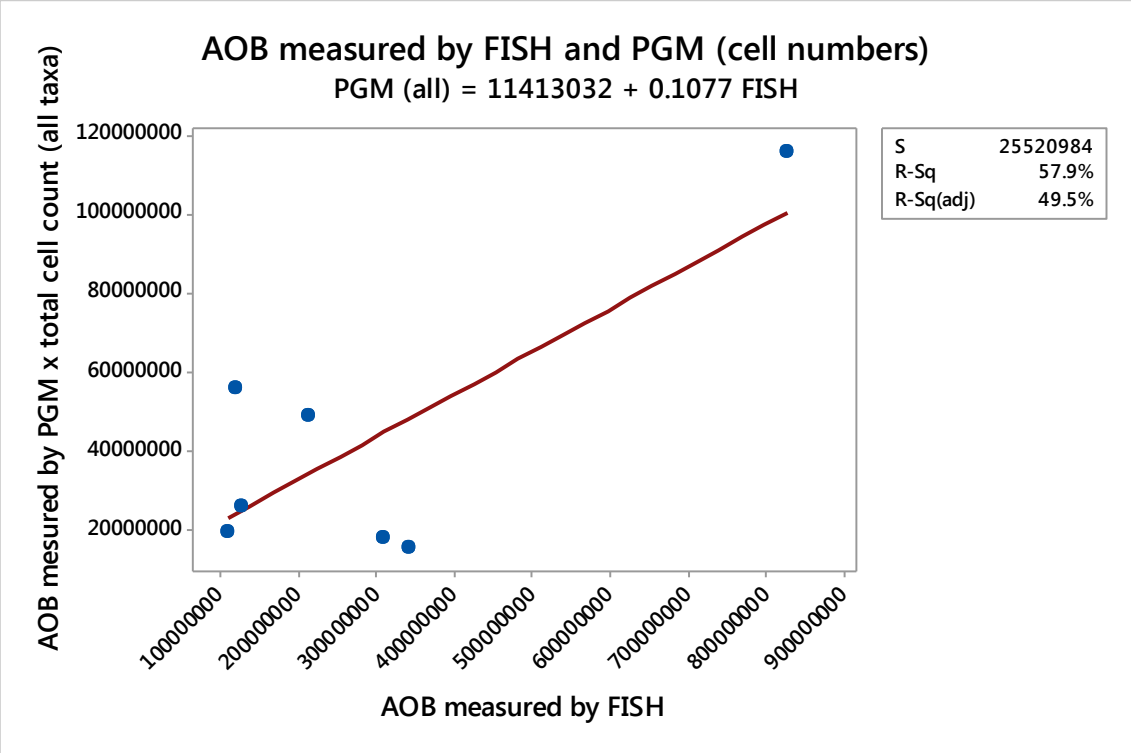


Figure 43: AOB measured by FISH and PGM where PGM data is relative abundance for all AOB taxa detected multiplied by total cell count

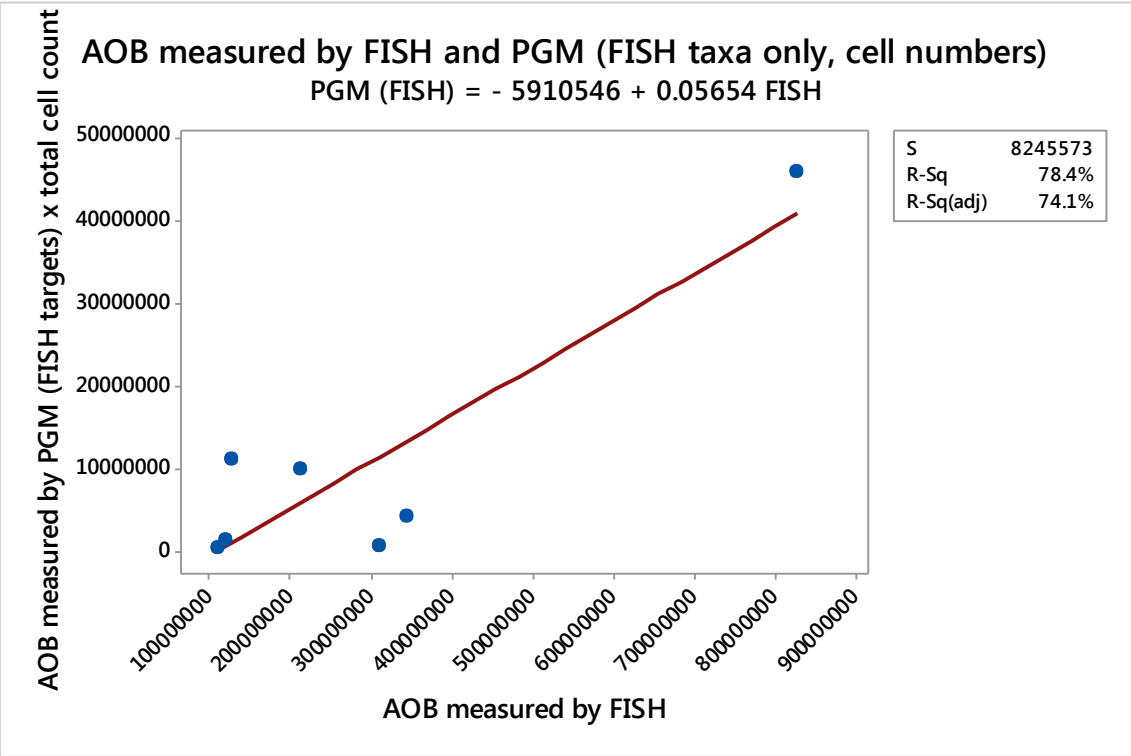


Figure 44: AOB measured by FISH and PGM where PGM data is relative abundance for only those AOB taxa targeted by FISH probes NSO1225, Neu and 6a192 multiplied by total cell count

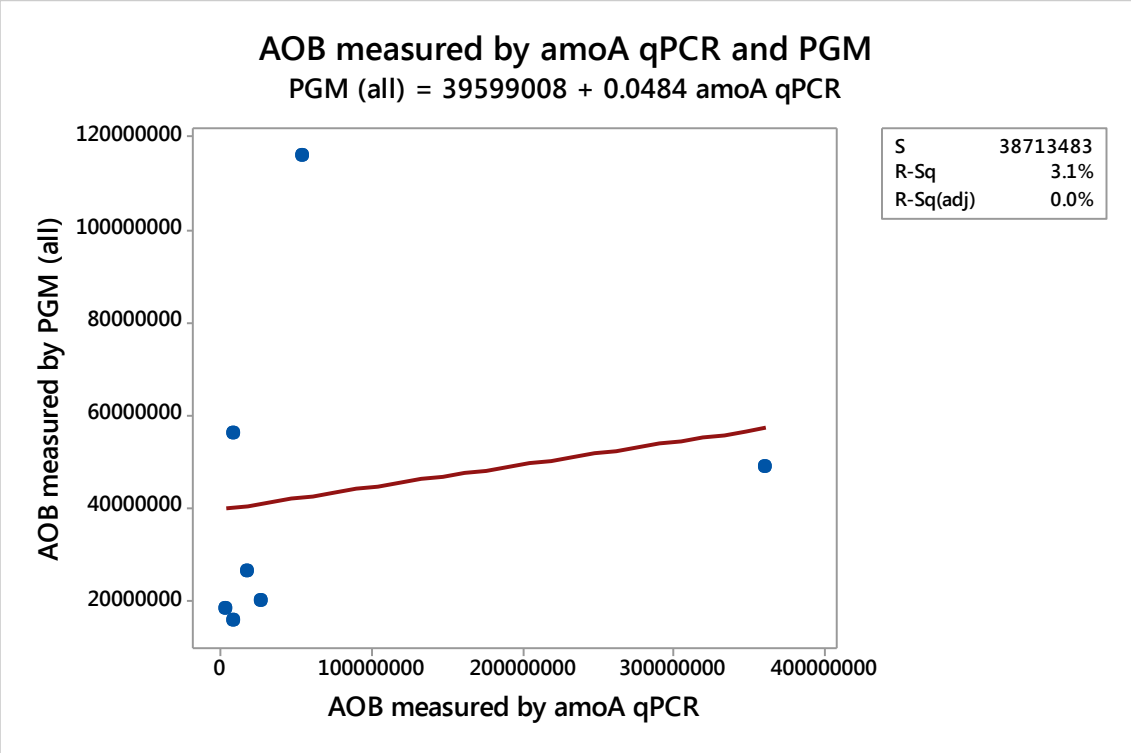


Figure 45: Abundance of ammonia oxidising bacteria (AOB) by quantitative polymerase chain reaction (qPCR) targeting the ammonia mono-oxygenase A gene (amoA) versus the relative abundance of AOB detected by Personal Genome Machine (PGM) multiplied by total cell counts

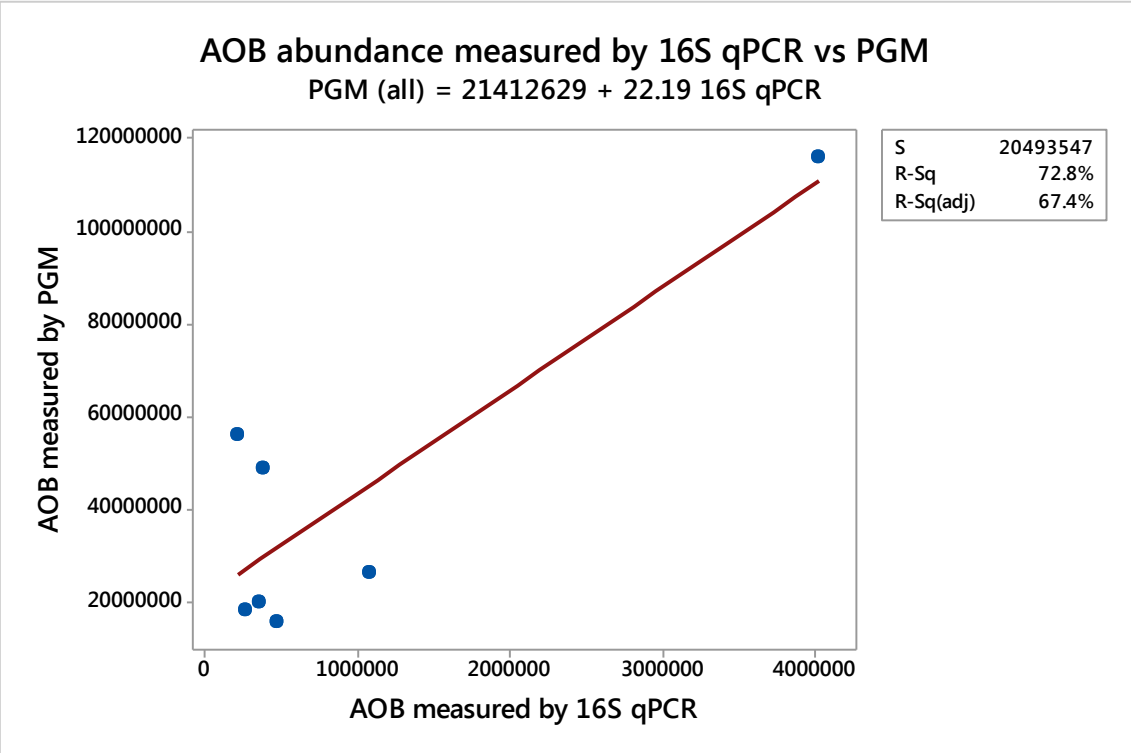


Figure 46: Abundance of ammonia oxidising bacteria measured by quantitative polymerase chain reaction (qPCR) targeting the 16S gene versus AOB relative abundance measured by PGM multiplied by total cell count