Investigations into the existence of unique environmental *Escherichia coli* populations

Report to the
Water Research Commission

by

SC MacRae[1], T Seale[1], ET Steenkamp[1], VS Brözel[1,2] & SN Venter[1]
[1] Department of Microbiology and Plant Pathology, University of Pretoria,
[2] Department of Biology and Microbiology, South Dakota State University

WRC Report No. 1967/1/13
ISBN 978-1-4312-0442-7

July 2013

Obtainable from
Water Research Commission
Private Bag X03
GEZINA, 0031

orders@wrc.org.za or download from www.wrc.org.za

This report is accompanied by a CD containing the text files with the sequence alignments of the four genes used in this study.

EXECUTIVE SUMMARY

BACKGROUND

Based on the assumption that *Escherichia coli* does not multiply or survive for long periods outside the intestines of warm-blooded animals, *E. coli* is used as an indicator of recent faecal contamination. However, several recent studies have reported that some *E. coli* strains are capable of surviving and multiplying in the environment and were present in the absence of any obvious faecal contamination. Based on sequence data of multiple genes it was shown that some of the *E. coli* isolates obtained from environmental sources did not belong to the species *E. coli*. They formed part of unique clades (cryptic species) but could phenotypically not be distinguished from *E. coli*. Should unique environmental *E. coli* populations be present in aquatic environments, the suitability of *E. coli* as an indicator organism is questioned.

The overall goal of the study was to investigate whether or not unique environmental populations of *E. coli* exist in aquatic environments. These *E. coli* populations were expected to be genetically distinct from their commensal and pathogenic counterparts potentially due to their adaptation to long-term survival in the external environment. If these environmental strains could be effectively characterised or identified and their ecology and risk to humans be better understood, the use of *E. coli* as an indicator organism could be improved.

AIMS

The following were the aims of the project:

1. Obtain populations of *E. coli* isolates from the chosen locality.
2. Use genomic fingerprinting procedures and cluster analysis to genetically characterise populations from the different samples.
3. Determine how the various isolates and clusters are related to one another and to isolates that have been studied previously by making use of phylogenetic analyses of multiple house-keeping genes.
4. Development of easily scorable neutral and pathogenesis-related marker sets, by making use of existing genome sequence information for *E. coli*.
5. Analysis of population dynamics within and among the populations of *E. coli* to determine how genetic information is shared among populations and whether distinct environmental populations of this bacterium do exist. Recommendations on how this data will impact on the ability to assess and evaluate the associated health impacts will be made.

METHODOLOGY

The diversity and dynamics of an *E. coli* population in an aquatic ecosystem was studied by initially focusing on the *E. coli* population associated with the Rietvlei Dam. Isolates obtained from the Roodeplaat Dam, and from aquatic plants sampled at 6 other dams were also included to investigate the population dynamics of *E. coli*. Only strains conforming to the classical description of *E. coli* and that could be isolated on media typically used in water quality studies were included. Although atypical strains may exist they would not jeopardise the use of *E. coli* as an indicator organism if they were not detected by the normal enumeration procedures.

In order to determine whether the *E. coli* diversity in the aquatic environment only reflected the diversity of *E. coli* found in humans and warm-blooded animals, four core genes, *rpoS, uidA, mutS* and *fadD* were

sequenced and analysed. The *rpoS* gene was selected as it was previously shown to group strains monophyletically. The *mutS* (methyl-directed mismatch repair) and *fadD* (fatty-acyl CoA synthetase) genes were selected based on their variability and the *uidA* (β-glucuronidase) gene was included as this is a gene unique to *E. coli*. The sequence data was used for the construction of Maximum-Likelihood trees.

When trying to establish whether unique bacterial populations do exist, it is important to determine how populations are structured and to quantify the levels of differentiation and information sharing among sub-populations. The DNA sequence information for two gene regions (rpoS and uidA) was used to calculate population genetic parameters. The program Structure was used to indicate population structure and gene flow and genetic differentiation were calculated with the use of the program DnaSP.

RESULTS

All the aims of the project were achieved. A large set of *E. coli* strains (> 250) associated with the aquatic environment was collected. Based on the sequences of four core gene regions, their phylogenetic relationship with each other and other reference strains could be determined. This study revealed that there was a high level of diversity within the *E. coli* population isolated from aquatic environments. Although, many of the strains isolated from the aquatic environment could not be distinguished from the sewage isolates, the phylogenetic analysis of the sequences of four selected core genes revealed at least two possible environmental *E. coli* groups amongst those strains isolated from aquatic plants.

The sequence data was also used to determine the population structure of these aquatic isolates. Gene flow and genetic differentiation analyses confirmed that these plant-associated clusters showed some level of genetic separation from the rest of the *E. coli* population and that the two environmental clusters have undergone significant population sub-division.

CONCLUSIONS

Based on the findings from this study it can be concluded that:

- A highly diverse *E. coli* population was present in the aquatic environment sampled.
- The *rpo*S phylogeny confirmed that all isolates belonged to *E. coli sensu stricto* and that none of the isolates grouped with the five *E. coli sensu lato* clades known to contain unique environmental isolates.
- The presence of at least two possible environmental *E. coli* clusters was observed. These isolates were collected from aquatic plants and decaying plant material. Although the current data indicated that all isolates still belonged to one population, these two plant-associated clusters were found to be genetically distinct and have undergone significant population sub-division.
- Many of the *E. coli* isolates obtained from the aquatic environment grouped with sewage isolates and would likely have the ability to circulate within the human population. This indicated that in most cases the presence of *E. coli* could still be used to evaluate the safety of water for human use.

RECOMMENDATIONS

The existence of unique environmental *E. coli* populations associated with aquatic plants raises the question whether or not such populations also exist in other aquatic environments. Important environments such as the various types of aquifers should be investigated. The potential link between these isolates and other

components of the aquatic ecosystem, such as aquatic invertebrates and water birds, should also receive further attention.

These results also raise specific questions concerning the use of *E. coli* as a faecal indicator. As it was demonstrated that certain unique groups of *E. coli* strains can survive and proliferate in aquatic environments, it is important to establish whether these strains could also survive in the human gut and pose a threat to water users by harbouring genes associated with pathogenic *E. coli*. On the other hand, if these strains of *E. coli* are not circulating within the human population, it would be important to know how often and at what levels they are detected when *E. coli* is used as a water quality indicator.

Sequencing the genomes of some of these environmental *E. coli* isolates detected during this study will provide an ideal opportunity to address some of the questions raised above. Comparative genomics will be ideal to determine whether these “true” *E. coli* are also undergoing reductive evolution and have lost their ability to grow and cause disease in the human gut.

A genomics approach could also assist in developing more specific detection methods for *E. coli* associated with humans and warm-blooded animals. Based on the genome data, the unique metabolic capabilities of environmental strains could be detected and then used to differentiate them from the *E. coli* associated with human gastrointestinal tract.

ACKNOWLEDGEMENTS

The project team wishes to thank the members of the Reference Group who provided valuable suggestions and sound advice. We also wish to thank our collaborators who assisted with the sampling.

Reference Group	Affiliation
Dr K Murray	Water Research Commission (Chairman)
Dr TG Barnard	University of Johannesburg
Prof CC Bezuidenhout	North-West University
Dr M Du Preez	CSIR
Prof MNB Momba	Tshwane University of Technology
Prof A Okoh	University of Fort Hare
Prof N Potgieter	University of Venda

Collaborators	Affiliation
Paul Botes	Department of Water Affairs
Leanne Coetzee	City of Tshwane (currently CSV Water)
Jan Swart	City of Tshwane
Nico van Blerk	ERWAT

The project team would like to thank the Water Research Commission for financing the project.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS & ABBREVIATIONS

AFLP	Amplified Fragment Length Polymorphism
bp	base pairs
DNA	Deoxyribonucleic acid
dNTP	deoxyribonucleotide triphosphate
MLEE	Multi-locus enzyme electrophoresis
MLST	Multi-locus sequence typing
PCR	Polymerase chain reaction
PFGE	Pulse Field Gel Electrophoresis
rpm	revolutions per minute
rRNA	Ribosomal ribonucleic acid
UV	Ultra violet

CHAPTER 1: AIMS AND APPROACH

1.1 INTRODUCTION

Commensal (non-pathogenic) and pathogenic *Escherichia coli* strains are both commonly associated with the gastrointestinal tracts of warm-blooded animals. These strains also spend a considerable part of their life in an environment outside of their primary host (Gordon, 2001). Savageau (1983) therefore suggested that *E. coli* inevitably have two habitats. He supported the established paradigm, that the primary environment exists in the gastrointestinal tract of the mammalian host whereas the external environment (water, soil and sediment), is the secondary environment. These two habitats differ immensely in terms of biotic and abiotic conditions. Bacteria in water, soil and sediments are exposed to lower temperatures, UV radiation, limited available nutrients, environmental pollutants and predation, which collectively result in the decrease in density of specific strains to undetectable levels. Based on this information it is often concluded that secondary habitats do not actively support the growth of *E. coli* (Solo-Gabriele et al., 2000, Walk et al., 2007). Because of its close association with the gastrointestinal tract of humans and animals it was assumed that *E. coli* does not survive for long periods in external environments, and that it is unable to multiply outside of the intestines of humans and animals. The presence of *E. coli* in these other environments is apparently maintained by the constant input of isolates from the primary habitat. For this reason, *E. coli* is used as an indicator of recent faecal contamination (Winfield and Groisman, 2003).

In contrast, several recent studies have reported that some *E. coli* strains are capable of surviving in soil and water for longer periods and where present in the absence of any obvious faecal contamination (Gordon et al., 2002; Power et al., 2005; Solo-Gabriele et al., 2000, Walk et al., 2007). It is therefore likely that unique environmental *E. coli* strains exist in aquatic environments despite the apparent absence of any faecal contamination. Furthermore, these *E. coli* may be genetically different from their commensal and pathogenic counterparts. If this is indeed the case, the suitability of *E. coli* as an indicator organism is questioned. If these environmental strains could be effectively characterised or identified and their ecology and risk to humans be better understood, the use of *E. coli* as an indicator organisms could be improved.

1.2 PROJECT AIMS

The overall goal of the study was to investigate whether or not natural populations of *E. coli* are structured according to habitat, and if so whether or not unique environmental strains of *E. coli* exist in nature. It was hypothesised that any imposed separation in terms of habitat would reflect at the genetic and genomic levels. Therefore environmental *E. coli* populations were expected to be genetically distinct from their commensal and pathogenic counterparts potentially due to their adaptation to long term survival in the external environment. If such unique populations do indeed exist, and could be genetically differentiated from the others it would be possible to investigate incidences as mentioned above to exclude these strains from water quality analyses. This ability will also improve the overall use of *E. coli* as an indicator of faecal contamination.

The following were the aims of the project:

1. Obtain populations of *E. coli* isolates from the chosen locality.
2. Use genomic fingerprinting procedures and cluster analysis to genetically characterise populations from the different samples.

3. Determine how the various isolates and clusters are related to one another and to isolates that have been studied previously by making use of phylogenetic analyses of multiple house-keeping genes.
4. Development of easily scorable neutral and pathogenesis-related marker sets, by making use of existing genome sequence information for *E. coli*.
5. Analysis of population dynamics within and among the populations of *E. coli* to determine how genetic information is shared among populations and whether distinct environmental populations of this bacterium indeed exist. Recommendations on how this data will impact on the ability to assess and evaluate the associated health impacts will be made.

All the project aims listed above were met during this study conducted at the University of Pretoria.

1.3 SCOPE AND LIMITATIONS

1.3.1 Study Approach

For this study a number of decisions related to the experimental approach and sample and site selection were made. In the initial study only the *E. coli* populations present in an aquatic ecosystem and those associated with human waste were considered. The rationale underpinning the study design was that, while *E. coli* can be isolated from numerous environments, aquatic environments are prime candidates to harbour unique environmental populations. The hypothesis was therefore that strains unique to the aquatic environment, but not detected in sewage effluent, will represent the true environmental strains.

River systems used as sources of drinking water typically experience continuous flow of water and have a relative short hydraulic retention time and that the bacterial populations present could be very dynamic. The decision was therefore made that a reservoir would be a better study site as there is a higher likelihood to find stable bacterial populations within this system. For this reason the Rietvlei Dam, south of Pretoria was selected as the primary study site. Isolates obtained from the Roodeplaat Dam, and from aquatic plants sampled at these two as well as 6 other dams were also included to investigate the population dynamics of this species.

1.3.2 Bacterial isolations

It was decided that only strains conforming to the classical description of *E. coli* and that could be isolated on media typically used in water quality studies would be included. Although atypical strains may exist they would not interfere with water quality analyses if they are not detected by the normal enumeration procedures. To allow for meaningful interpretation of the data, studies investigating the structure and dynamics of bacterial populations have indicated that at least 100 isolates should be obtained. The final number of isolates was substantially higher than this initial target.

1.3.3 AFLP analysis

The initial proposal indicated that all the *E. coli* isolates obtained would be typed by Amplified Fragment Length Polymorphism (AFLP) analysis, a genomic finger printing method (Vos et al., 1995). It was hoped that the AFLP data would clearly reveal the level of diversity among the various isolates and if separate environmental clusters were present within the collection of isolates. It was, however, decided to abandon this approach as a parallel study conducted at the Roodeplaat Dam indicated that sequencing data provided a more defined analysis of the grouping of isolates. Instead of the AFLP analysis, sequencing of the *rpoS* gene was initially used to group and identify all isolates.

1.3.4 Development of population markers

The detection of stratification amongst populations within a particular species (development of separate and genetically unique populations) is usually done using unlinked genetic markers. Finding such markers for bacterial populations, especially for *E. coli*, is not easy. An initial detailed evaluation of the literature as well as comparisons of the published and available *E. coli* genome sequences highlighted the difficulty in finding a non-coding region that is present in all *E. coli* isolates. It was subsequently realised that even if such a sequence could be located, it will be unlikely that this sequence will also be present in some of the environmental populations of *E. coli*, as most of the genomes sequenced belonged to pathogenic isolates. For this reason, it was decided to change focus and rather look for highly variable genes that are part of the core genome. During the present study four, gene core genes, *rpoS, uidA, mutS* and *fadD* were sequenced.

1.3.5 Analysis of population dynamics

In order to establish if unique populations do exist, it is important to determine how populations are structured and to quantify the levels of differentiation and information sharing among sub-populations. Overall, the results of these population analyses were used to indicate whether distinct environmental populations of *E. coli* do indeed exist in the localities studied. For this analysis, isolates from another parallel study performed at the Roodeplaat Dam was also included. The DNA sequence information for two gene regions (*rpoS* and *uidA*) available for this collection of isolates were used to calculate population genetic parameters. The program Structure was used to indicate population structure and gene flow and genetic differentiation were calculated with the use of the program DnaSP.

CHAPTER 2: BACKGROUND

2.1 INTRODUCTION

Escherichia coli is one of the most versatile and widely recognised microorganisms. Its flexibility has allowed for its exploitation in recombinant DNA technology making it the workhorse of many laboratories and one of the most widely-used model organisms. Aside from its uses in the laboratory, *E. coli* is an important inhabitant of the gastrointestinal tract of humans and warm-blooded animals. The physiology, biochemistry and genetics of *E. coli* have been studied extensively over many decades. However, these studies have focussed predominately on the pathogenic and commensal isolates because it was believed that the replication and growth of this bacterium were restricted to the gastrointestinal tract of humans and animals. It is generally believed that many of the *E. coli* strains are harmless commensals (i.e., bacteria that benefit from the host, while the host is neither benefited nor harmed) but others are important pathogens for both humans and animals.

The genus *Escherichia* is a member of the class *Gammaproteobacteria* in the phylum *Proteobacteria* and belongs to the family *Enterobacteriaceae*. This family represents a large assemblage of Gram-negative bacteria that include pathogens of plants and animals and harmless symbionts. Some of the genera include *Citrobacter, Enterobacter, Escherichia, Klebsiella, Pantoea* and *Salmonella*. Within the *Enterbacteriaceae, Escherichia* is most closely related to *Salmonella* (Farmer, 1995).

Analyses of DNA sequences for the 5S and 16S ribosomal RNA (rRNA) gene suggests that *Salmonella* and *Escherichia* diverged from a common ancestor between 100 to 150 million years ago (Doolittle et al., 1996, Welch, 2006). The adaptation of commensal *E. coli* to the gastrointestinal tract of animals is considered as a defining factor in its divergence from its common ancestor with *Salmonella*. Nevertheless, genome sequence data suggest that *Salmonella enterica* serovar Typhimurium (*S*. Typhimurium) and non-pathogenic *E. coli* share up to 80% homology, and that their genomes are mostly superimposable (Lavigne and Blanc-Potard, 2008).

In addition to the species E*. albertii, E. adecarboxylata, E. blattae, E. fergusonii, E. hermanii* and *E. vulneris*, the genus *Escherichia* also include *Shigella* species. *Shigella* and *E. coli* have always been considered as close relatives. *Shigella* was originally given the name *Bacillus dysenteriae* as it was identified as the cause of bacillary dysentery, whereas *E. coli* (previously named *Bacillus coli*) was only known as a commensal at the time (Pupo et al., 2006). Factors that distinguish *Shigella* from *E. coli* are that *Shigella* is non-motile and unable to ferment lactose. This sometimes resulted in the incorrect classification of some pathogenic *Shigella* strains that exhibited *E. coli* characteristics and maintain the ability to ferment sugars (Pupo et al., 2006). However, based on phylogenetic data *Shigella* should be considered as a subgroup of *E. coli* despite their taxonomic separation into two genera, with *Shigella* often being associated with the more pathogenic strains causing dysentery, similar to that caused by enteroinvasive *E. coli* (EIEC) (Escobar-Páramo et al., 2003, Hartl and Dykhuizen, 1984).

E. coli is distributed worldwide with an estimated total population size of 10^{20} (Tenaillon et al., 2010) and occurs in high densities in the gastrointestinal tracts of humans and other warm-blooded animals. The majority of *E. coli* strains are thought to be transient in the gastrointestinal tract with little or no effect on the host, but some are able to persist and form an integral part of the gut microflora (Walk et al., 2009). However, *E. coli* and other *Proteobacteria* only constitute a small fraction of the bacterial diversity thought to make up the gut microbiota. This is not surprising as *E. coli* is a facultative anaerobe existing in a strictly

anaerobic environment (Eckburg et al., 2005). Yet the gut remains its primary habitat where it exists as a predominant aerobic organism (Tenaillon et al., 2010).

The relationship between *E. coli* and its host has the ability to fluctuate between commensalism, mutualism and opportunistic pathogenesis (Tenaillon et al., 2010). As long as the bacterium does not acquire genetic elements encoding virulence factors, it will remain a commensal organism. At any one time, an individual may be colonised by a predominant *E. coli* strain and over time that strain will most likely become the dominant type. This suggests that there is a strong relationship between the host and strain. Host characteristics such as body mass, diet and gut morphology may all play a role in the distribution of strains and phylogenetic groups. Although, there is some overlap in host range between humans and animals, this may be a result of a host picking up a transient strain from the environment (Hartl and Dykhuizen, 1984).

As a commensal organism, *E. coli* is well adapted to life in the gastrointestinal tract (Hartl and Dykhuizen, 1984, Tenaillon et al., 2010). These bacteria are capable of growing in the presence of bile salts and are located in the mucosal layer covering the epithelial cells throughout the intestinal tract and attach there via type 1 pili. In humans, they are consequently shed from the mucosal layer and excreted in the faeces resulting in approximately 10^7 to 10^9 colony-forming units per gram of faeces. The mucosal layer provides a nutrient rich environment to which *E. coli* has adapted using micro-aerobic and anaerobic respiration. Not only does this environment provide nutrients but also protection from certain stresses and in response, *E. coli* and the other commensals benefit the host by preventing colonisation of the gut by pathogens.

Many strains of *E. coli* are intrinsic pathogens as they contain certain virulence characteristics. These virulence factors include, amongst others, toxin production, invasive enzymes and phagocytosis resistance, which allow *E. coli* to overcome the hosts' defences and cause disease (Hartl and Dykhuizen, 1984). They majority of virulence factors are associated with plasmids and pathogenicity islands which non-pathogenic strains can quire through horizontal gene transfer (Wirth et al., 2006).

The majority of studies performed on *E. coli* have been focused on those strains causing disease. Pathogenic strains of *E. coli* include extraintestinal pathogenetic *E. coli* (ExPEC) which can cause neonatal meningitis and urinary tract infections. Shiga toxin-producing *E. coli* (STEC) including enterohemorrhagic *E. coli* (EHEC), specifically *E. coli* O157:H7 are responsible for many outbreaks of food and water-borne disease. EHEC causes bloody diarrhoea and life threatening conditions such as hemorrhagic colitis and haemolytic uremic syndrome (Welch, 2006, Ishii and Sadowsky, 2008). In addition, there are at least four other recognised clinical diarrheagenic isolates: enteroaggregative *E. coli* (EAEC) that can cause persistent diarrhoea lasting up to two weeks or longer; enteropathogenic *E. coli* (EPEC) that causes of watery diarrhoea in infants predominantly in developing countries; enterotoxigenic *E. coli* (ETEC) that is responsible for travellers' diarrhoea; and lastly enteroinvasive *E. coli* (EIEC) that is genetically, biochemically and pathogenetically closely related to *Shigella* and that causes invasive inflammatory colitis and dysentery by invading the intestinal epithelial tissue (Ishii and Sadowsky, 2008, Rasko et al., 2008 and Welch, 2006).

In this literature overview, the genetic diversity within an *E. coli* population was investigated, with special interest in its survival in secondary environments. Furthermore, it was interesting to look into the effects of the secondary environment on *E. coli* population structure with multiple reports of its existence in the external environment outside of the host will be discussed. In addition, *E. coli* as an indicator organism was discussed as well as the consequences of its occurrence in the external environment on its use as an indicator for water quality. The role of the external environment on the overall genetics of *E. coli* as a population as well as multiple methods that have been used to characterise *E. coli* populations were also addressed.

2.2 *E. COLI* AS AN INDICATOR ORGANISM

Faecal indicator organisms are used throughout the world to assess the microbial safety of various water systems (Anderson et al., 2005). This is because they reside in the gut of humans or animals in close association with the host. The presence of these organisms in soil and water systems indicates recent faecal contamination, and an increase in the levels of faecal coliforms (faecal bacteria) provides a warning for the possible presence of pathogens, a failure in the treatment of the water or faults in the distribution system (Ishii and Sadowsky, 2008).

The most commonly-used group of indicators are faecal coliforms. Faecal coliforms are typically Gram-negative, facultative anaerobic, nonspore-forming bacteria that have the ability to ferment lactose. Faecal coliforms themselves do not normally cause serious illness, however they are easy to culture and their presence in water indicates the possibility that other faecal pathogens including enteric bacteria (diarrheagenic *E. coli*, *Shigella*, *Salmonella* and *Campylobacter*), viruses (Norovirus and Hepatitis A), and protozoa (Giardia and Cryptosporidium) may be present (Ishii and Sadowsky, 2008). In addition to species of *Escherichia*, this group includes isolates of the genus *Citrobacter*, *Enterobacter* and *Klebsiella* (Elliot and Colwell, 1985). The primary requirements for representing a suitable faecal indicator are as follows: an indicator organism should be present in higher numbers than the pathogen, survive similar conditions as potential faecal-derived pathogens, be present when the pathogen is there and absent when it is not and most importantly be non-pathogenic (Ishii and Sadowsky, 2008).

The use of *E. coli* as an indicator organism is based on a number of assumptions. The first is that this bacterium is primarily associated with the gastrointestinal tract of humans and animals and therefore shows faecal specificity (Brennan et al., 2010). The second assumption is linked to the first assumption and states that *E. coli* is unable to replicate and multiply in the environment outside of the host, due to the extreme changes in the environmental conditions (Brennan et al., 2010). Accordingly, density of *E. coli* in the secondary environment would be directly proportional to the constant faecal input of isolates from the primary host (Winfield and Groisman, 2003, Power et al., 2005). The third assumption is that all cells in the external environment possess a clonal quality in that they have identical characteristics in terms of their reproduction and survival in the external environment (Gordon, 2001, Power et al., 2005). Here it is assumed that the clonal composition of the *E. coli* strain identified in the soil or water represents the same clonal composition as the *E. coli* in the host responsible for the faecal contamination (Gordon, 2001).

Recent studies have shown that the basic assumptions regarding the biology of *E. coli* and its use as faecal indicator organism might not always be true (Gordon et al., 2002; Power et al., 2005; Solo-Gabriele et al., 2000, Walk et al., 2007). For example, we now know that *E. coli* is capable of proliferation in many environments and not only the gastrointestinal tract. In addition, there is little evidence of a strict relationship between *E. coli* and specific hosts nor for temporal stability in the clonal composition of populations. The most important problem with using *E. coli* as a way to track faecal contamination may be that there appear to be significant changes in the composition of the *E. coli* community during the changeover from the host to the external environment. The environment outside of the host differs greatly and therefore strains that may initially be clonal adapt to the external environment by the uptake of additional genetic elements. There is some evidence, although limited, which suggests there is little similarity between *E. coli* populations in the host and *E. coli* populations in the external environment where the contamination occurs (Gordon, 2001).

2.3 *E. COLI* IN THE ENVIRONMENT OUTSIDE OF THE HOST

According to Gordon (2001) most *E. coli* find themselves in an environment outside of their host at some stage and they may spend up to half their life in the environment outside of the host. Savageau (1983) suggested that *E. coli* inevitably have two habitats, with the primary environment represented by the

gastrointestinal tract of the human or animal host. The external environment that can include water, soil and sediment, represents the secondary environment in which *E. coli* can exist (Savageau, 1983).

These two habitats differ immensely in both their biotic and abiotic conditions. The environment within the host is characterised by readily-available nutrients and carbon sources, constant temperature, microbial competition and protection from predation (Brennan et al., 2010). In addition, the gastrointestinal tract of the host contains an overabundance of bacterial species, which have co-evolved with one another forming an array of symbiotic relationships. In contrast, cells in the secondary environment may be exposed to lower temperatures, UV radiation, limited available nutrients, limited moisture, environmental pollutants and predation. All these factors ultimately result in the decrease in density of specific strains in the secondary environment, often to undetectable levels (Ishii and Sadowsky, 2008). As a result, it is often concluded that the external environment does not actively support the growth of *E. coli*, forming the basis of its use as an indicator organism (Solo-Gabriele et al., 2000, Walk et al., 2007).

In recent years there has been evidence suggesting that *E. coli* are capable of surviving and even multiplying in the external environment, in the absence of faecal contamination, in both tropical and temperate climates (Anderson et al., 2005, Gordon et al., 2002, Ishii et al., 2006, Power et al., 2005, Solo-Gabriele et al., 2000, Walk et al., 2007). Although most *E. coli* strains are commensals, many strains have diverged to take on a pathogenic lifestyle. There is a growing body of data suggesting that others may have evolved to take on a free-living lifestyle, which is consistent with one of the hypotheses developed for explaining the origin of free-living *E. coli*. According to this hypothesis, these bacteria originated from faecal contamination in the past and over time some strains have adapted to replicating outside of their mammalian host and eventually form part of the natural microbiota of the external environment. A second school of thought is that free-living *E. coli* strains were always part of the microbiota in the external environment and that some strains acquired the ability to cause disease in human and animal hosts. If either of these two scenarios is correct then the use of *E. coli* as an effective indicator organism is questionable (Power et al., 2005).

2.3.1 *E. coli* associated with water, sand, sediment and algae

The results of a long-term study in an Australian lake important for water supply to Sydney have shown that annual coliform blooms have occurred during the past 30 years (Power et al., 2005). The researchers identified three *E. coli* strains responsible for the bloom events, which all possessed a Group 1 capsule. The encapsulated strains appeared to be free-living, suggesting that the possession of a capsule can greatly improve the survival of the bloom strains. These *E. coli* Group 1 capsules were remarkably similar to the capsules produced by *Klebsiella* spp such as *K. pneumonia*, which is also a coliform and an opportunistic pathogen, although *K. pneumoniae* is ubiquitous in the environment. These findings thus indicate that Group 1 capsules probably play an important role in the survival of these bacteria outside of their mammalian hosts. Furthermore, the high levels of *E. coli* observed in the lake could not be linked to faecal contamination, suggesting that these bloom strains are able to survive and multiply in the external environment (Power et al., 2005).

The results of a number of studies have shown that recreational beaches are subject to faecal contamination from sewage and agricultural runoff, wild and domestic animals and the recreational users themselves. Wheeler Alm et al. (2003) provided evidence that freshwater beach sand and sediment act as a reservoir for faecal indicator organisms. They concluded that the amount of *E. coli* in the water was not linked to seasonal fluctuations and that E. coli persisted at various depths throughout the sediment. These results concurred with results obtained from a study by Whitman et al. (2006) where *E. coli* was found to persist in forest soils and sediment. This suggested that the soil environment provided protection that may be a major factor in *E. coli* survival where the temperature of the air and water are constantly fluctuating (Sampson et al., 2006).

Other studies showed that sand or sediment is often the main source of *E. coli* in freshwater systems (Byappanahalli et al., 2003a, Ishii et al., 2007 and Solo-Gabriele et al., 2000). Here the *E. coli* concentrations correlated to tidal cycles or an increase of water during heavy rainfall periods causing re-suspension of the sediment and consequently increasing the faecal bacteria counts (Whitman et al., 2006). *E. coli*, once established in the soli can persist in high numbers and following contact with water in high tide or rain, can act as a constant source of *E. coli* into neighbouring water sources (Ishii et al., 2007; Solo-Gabriele et al., 2000).

The growth and survival of *E. coli* in the secondary environment has been associated with macro-algae in the genus *Cladophora*. Byappanahalli et al. (2003b) investigated the possibility that *Cladophora* supports the growth of *E. coli*. *Cladophora* represents macrophytic green algae that grows as dense mats and strands in freshwater streams and lakes. High levels of indicator bacteria have been associated with the presence of *Cladophora* algal mats, which led researchers to hypothesise that *Cladophora* serves as an environmental reservoir for *E. coli* and other possible indicator bacteria. They proposed that algae serve as attachment sites where bacteria can avoid harmful environmental conditions such as UV radiation, predation and poor nutrient availability. Byappanahalli et al. (2003b) also demonstrated that not only does *Cladophora* provide a favourable environment for *E. coli* growth but also it may provide a primary source of nutrients via algal exudates. They showed that *E. coli* growth increased when *Cladophora* leachate concentrations increased.

2.3.2 Environmental conditions affecting *E. coli* outside the host

There is evidence that the survival of *E. coli* in the secondary environment has been linked to water temperature and the presence of sand or other particles and green algae (Sampson et al., 2006, Solo-Gabriele et al., 2000). Soil and sediments in sub-tropical and tropical regions may provide favourable conditions by providing a site of high nutrients, protection from UV and protozoan grazing and warm temperatures, allowing the colonisation of *E. coli* populations (Brennan et al., 2010, Wheeler Alm et al., 2003). It has been suggested that *E. coli* can maintain autochthonous populations should the conditions remain favourable. Results from a study by Ishii et al. (2006) indicated that the same strain of *E. coli* survived the winter months with freezing temperatures and then were able to multiply when the temperatures increased in the summer months. In addition, they discovered that *E. coli* does not multiply in cooler waters, although it is able to survive for longer periods at lower temperatures.

The ability of *E. coli* to adapt and survive in the secondary environment may also be a result of its versatility in acquiring energy (Luchi and Lin, 1993). *E. coli* is able to survive on minimal carbon and nitrogen sources, as well as phosphorous and sulphur. It is also able to utilise various aromatic compounds such as benzoic acid and phenylacetic acid as an energy source. It is thus likely that because of this bacterium's versatility in utilisation of energy sources, growth at varying temperatures and its ability to grow in both aerobic and anaerobic conditions, it is able to integrate into the microbial communities in different environments (Bennett et al., 1992, Ishii and Sadowsky, 2008).

2.3.3 Genetic diversity of naturalised *E. coli*

The persistence and proliferation of *E. coli* in the secondary environment raise the questions: Are environmental *E. coli* strains genetically distinct from their host-associated counterparts and do they still have the ability to circulate through human and animal hosts? The adaptation of *E. coli* to the external environment may be a result of certain genotypes being favoured by natural selection in different environments. Whittam (1989) tested this hypothesis by comparing the clonal composition of *E. coli* populations in the primary (avian gastrointestinal tract) and secondary (litter, water, and soil) environments. The results of this study revealed that the two different environments consisted of genetically distinct subpopulations. This study also showed that there is a significant change in the genetic composition of

E. coli populations from the primary to secondary environment, which may be a consequence of selection for specific clonal characteristics in each habitat (Whittam, 1989). An important conclusion from this study was that *E. coli* populations isolated from primary and secondary environments were clonally distinct, further supporting the idea that populations of free-living *E. coli* exist in nature. This would imply that the *E. coli* population found in the environment would be comprised of strains with the ability to grow in the environment in addition to strains derived through faecal contamination.

Whittam's study also set out to determine how *E. coli* adapted to the changes encountered when moving from the primary to the secondary environment (Whittam, 1989). He suggested that *E. coli* deal with the change by having a dual regulation system. Such dual regulations systems have been identified and characterised in the lac operon involved in lactose metabolism (Malan and McClure, 1984) and in the translation of secA, encoding a translocation ATPase, involved in secretion of proteins across the inner membrane of *E. coli* (McNicholas et al., 1997).

A study by Gordon et al. (2002) suggests some *E. coli* strains are better adapted to the external environment. They studied the genetic structure of *E. coli* populations in the primary and secondary environments where the faecal contribution into the secondary environment was known. Here they found that some strains recovered from the septic tank of a household were genetically distinct from the strains found in the human sources and that the source of these strains was unknown. Furthermore, they found that these strains grew better at lower temperatures, therefore validating the suggestion that certain *E. coli* strains are better suited to the secondary environment. This also supports the suggestion made by Whittam (1989), where selection may be the main driving force in the transition from the primary to secondary environment. Walk et al. (2007) set out to characterise the genetic diversity and the population structure of *E. coli* obtained from the sand and water of freshwater beaches, using both phenotypic and genotypic methods. They discovered that overall, the genetic diversity was widespread and several genotypes were consistently recovered, therefore suggesting that natural selection played a role in favouring certain genotypes. This data suggests that some *E. coli* genotypes are well adapted to the secondary environment as previously shown by Power et al. (2005).

Brennan et al. (2010) suggested that naturalised *E. coli* persisting in the soil are genetically distinct groups that have adapted physiologically to the soil environment by having increased environmental fitness. They discovered that *E. coli* isolated from soil demonstrated a level of environmental fitness greater than that of the laboratory strains. Therefore, when soil conditions are favourable, adapted strains can become naturalised and are in a better position to colonise a specific niche and thereby facilitate their integration into the indigenous microbial population (Bergholz et al., 2011). Here soil environments may selectively sort *E. coli* strains. These naturalised *E. coli* populations can then act as a reservoir for repeated contamination of water bodies and may increase the health risks associated with recreational water.

Byappanahalli et al. (2006) observed that soil-borne *E. coli* had similar HFERP (horizontal fluorophore-enhanced repetitive extragenic palindromic PCR) DNA fingerprints that clustered together in distinct groups. They discovered that *E. coli* isolated from the soil formed a unique group, different from representative faecal isolates. Soil was identified as a possible habitat for *E. coli* populations, provided that it is able to persist and become an integral part of the soil microbiota. Ishii et al. (2006) went further to state that some *E. coli* strains have become naturalised and that these naturalised strains could be repetitively isolated in specific soils and at the same locations over multiple seasons. In habitats such as soil and sediments already colonised by indigenous microbial populations, it raises the question of how does *E. coli* survive and compete for a niche.
Byappanahalli et al. (2007) tested the hypothesis that *E. coli* associated with *Cladophora* are genetically diverse. Using HFERP of over 800 isolates, they were able to demonstrate that *E. coli* isolates from *Cladophora* did in fact show a high level of genetic diversity. In addition, they showed that the *Cladophora*-associated *E. coli* formed a unified genetic group when compared to faecal strains obtained from humans and animals, even though their original source remains unknown. These results concur with previous studies

suggesting that *E. coli* populations can grow naturally in environments such as water, soil and algae, and therefore, compromising their use as indicator organisms (Byappanahalli et al., 2003b).

The existence of these naturalised *E. coli* raises the question of how these environmentally-fit populations arise. It may be that these populations already exist in the host as a minority and upon arrival in a favourable external environment, they are able to dominate due to natural selection and outcompete less competitive strains. Alternatively, strains may adapt upon arrival in the external environment by acquisition of advantageous genetic elements or activation of different metabolic pathways. In the latter situation, strains would have to survive the initial adaptation period and undergo certain selection pressures. In addition, through selection the strains may have established themselves and are not circulating through the host anymore.

2.4 *E. COLI* POPULATION DIVERSITY

Although *E. coli* is primarily known as a model organism, it is not a single clonal organism. Phenotypically they vary in antibiotic resistance profiles, carbon utilisation patterns, ability to cause disease, flagellar motility and biofilm formation (Anderson et al., 2006, Durso et al., 2004, Yang et al., 2004). This diversity within the species can be a consequence of acquisition of new genes via horizontal gene transfer, mediated by either bacteriophages or plasmids (Ishii and Sadowsky, 2008). In addition, mutations should not be overlooked as they also play an important role in the diversification of *E. coli*.

A study by Cooper and Lenski (2000) observed that *E. coli* lost the ability to utilise other carbon sources when they were extensively grown on minimal media supplemented with glucose. Here they suggested that specialization of *E. coli* might be a result of accumulating beneficial mutations and elimination of functions that are unnecessary and that decrease fitness. However, this may result in a population retaining mutations that increase fitness in one environment but are detrimental in another. This diversity amongst *E. coli* isolates is thought to be mainly driven by selection pressures where strains exposed to similar environments may share the same characteristics (Ishii and Sadowsky, 2008).

The population structure of *E. coli* is often defined as a balance between mutation and recombination, changing from a clonal population (i.e., a group of identical cells that share a common ancestor indicating that they are derived from the same mother cell) when recombination is low to a panmictic population (i.e., a population where all members are potential recombination partners) when recombination is high. *E. coli* was initially thought to have a clonal population structure before sequencing methods were available and Multi-locus enzyme electrophoresis (MLEE) revealed that there were only a few unique phenotypes (Tenaillon et al., 2010). It was suggested that *E. coli* reproduces clonally but undergoes increased recombination when conditions are harsh or environments change (Hartl and Dykhuizen, 1984, Whittam, 1996). Davis and Gordon (2002) suggested that the host dynamics greatly influence the clonal composition of the *E. coli* population. Similarly, prevailing conditions in the secondary environment determine the clonal composition of free-living *E. coli*.

Diversity within the population was thought to come about by an increase in clones carrying beneficial mutations, and through natural selection, potentially replace pre-existing ones (Whittam, 1996). However, after sequence analyses, numerous studies showed that when drawing phylogenetic trees using different individual genes, the trees were dissimilar. This led to the suggestion that recombination may be more frequent than originally thought. However, it was discovered that recombination events occur resulting in the horizontal transfer of short fragments of genetic material mostly outside of the core genome, which corresponds to a clonal population structure. The short size of the recombination fragments are not significant enough to blur the phylogenetic signal produced by the rest of the genome that is not involved in recombination (Tenaillon et al., 2010).

2.5 *E. COLI* GENOMIC DIVERSITY

E. coli strains show a significant difference at a genomic level in respect of their gene content (Bergthorsson and Ochman, 1995). Fourteen natural strains selected from the *E. coli* reference collection (ECOR; Ochman and Selander, 1984) were analysed by Bergthorsson and Ochman (1995) to investigate differences in genome size. In comparison to laboratory isolates of *E. coli* K-12 and *Salmonella* typhimurium LT2, the natural isolates showed differences in genome sizes of up to 650 kb. The results of this study suggested that the acquisition and removal of genetic information is not evolutionarily constant among laboratory strains but may be beneficial to natural strains adapted for growth in variable environments. Strains may lose or acquire genetic information as an adaptive response to a new environment resulting in possible genetic differentiation between strains in the host and those existing in the external environment.

A comparative study by Rasko et al. (2008) showed that of 17 *E. coli* genomes, including commensal and pathogenic isolates, the average genome size was 5 020 genes. The conserved core genome size was calculated to consist of approximately 2 200 genes and functional annotation of these genes suggested that they are involved in core metabolic processes. They calculated the size of the pan-genome to be more than 13 000 genes and suggest that the pan-genome of *E. coli* be considered as open. An open species pan-genome indicates that the species is still undergoing evolution and diversification by the acquisition and removal of specific genetic elements thereby creates a high level of flexibility in the genome, which allows *E. coli* to take on various adaptive paths (Tenaillon et al., 2010).

With such a large pan-genome, *E. coli* has the opportunity to diversify and acquire new genes depending on the environment it encounters. *E. coli* faces variable environments and strong selective pressures in each host and with a large pan-genome, subsets of *E. coli* strains are able to acquire certain genes or genomic islands that are favoured in a specific environment. Baur et al. (1996) discovered that *E. coli* could develop natural genetic competence in conditions similar to those found in river and spring water with calcium concentrations higher than 1 mM. The development of competence involves DNA binding and uptake followed by processing and finally expression. Their results suggest that the development of natural genetic competence is biologically possible but successful transformation is dependent on the *E. coli* stain and condition of the transforming DNA.

Rasko et al. (2008) suggested that commensal *E. coli* have the potential to become pathogenic by acquiring the appropriate pathogenic genes via horizontal gene transfer. In contrast, pathogenic strains may also lose their pathogenic genes and revert to a commensal state, through the loss of plasmids encoding pathogenicity factors (Rasko et al., 2008). With the ability to acquire and lose genetic information within a large pan-genome, the possibility for *E. coli* to inhabit various environments, including the environment outside of the host, is inevitable. The secondary environment may play a vital role in the generating and maintaining the genetic diversity of the *E. coli* population by selection of tolerant and persistent strains (Bergholz et al., 2011). Whittam (1996) refers to this as niche-specific selection.

2.6 CHARACTERISATION OF *E. COLI* POPULATIONS

2.6.1 Traditional approaches

Two techniques have traditionally been used to study the population structure of *E. coli*. The first method is serotyping, which was developed in the 1940s where *E. coli* was separated into serotypes based on the presence or absence of combinations of 173 O antigens, 80 K antigens and 56 H antigens (Tenaillon et al., 2010). The O antigen corresponds to the lipopolysaccharide of the cell wall, K antigens correspond to the polysaccharide capsule or envelope and lastly, the H antigen corresponds to the proteins that are involved in the formation of the flagellum, all of which are established on chromosomal genes (Hartl and Dykhuizen,

1984). PCR techniques have now been developed for typing of these antigens (Clermont et al., 2000, Yang et al., 2007). Nevertheless, most of the serotyping studies on *E. coli* were based on only the diverse and pathogenic strains associated with the gut of humans and animals using the ECOR collection. The ECOR collection was derived from mammals at various geographical locations (Ochman and Selander 1984) which formed the basis of so many *E. coli* based studies.

The second method that has been traditionally used to study populations of *E. coli* is multi-locus enzyme electrophoresis (MLEE). This method became available in the 1980's and allowed differentiation of *E. coli* strains based on the electrophoretic motility of certain housekeeping enzymes. The method makes use of the electrophoretic mobility of specific chromosomally encoded cytoplasmic enzymes to differentiate between strains and analyse the population genetics (Dijkshoorn et al., 2001, Gordon et al., 2002, Tenaillon et al., 2010, Walk et al., 2007). MLEE has some drawbacks including problems with band resolution and mainly that it determines phenotypes rather than genotypes.

2.6.2 Grouping *E. coli* based on phylogeny

Previous MLEE studies revealed that *E. coli* might have a "subspecific" structure (Gordon, 2004). Further phylogenetic studies have indicated that *E. coli* strains belong to one of five distinct phylogenetic groups (or phylogroups), namely A, B1, B2, D and E, although the members of group E are less common (Gordon, 2004, Gordon et al., 2008). Groups A and B1 are considered to be sister groups with group B1 believed to represent the "ancestral lineage" of *E. coli* (Gordon, 2004, Gordon et al., 2008). Group B2 strains are monophyletic whereas strains belonging to group D are not and possibly represent two or more clades. In addition, strains belonging to groups B2 and D have larger genomes than A and B1 strains and the presence or absence of virulence factors involved in causing extra-intestinal disease may also vary within groups (Gordon, 2004, Gordon et al., 2008).

Strains belonging to the four main groups may also differ in their ecological niche. The majority of commensal *E. coli* strains have been found to belong to group A whereas the more virulent extra-intestinal strains belong mostly to group B2 and some to group D. Strains associated with environmental sources belong mostly to the B1 phylogroup (Walk et al., 2007). With regards to the ecological distribution of the four phylogroups, Gordon (2004) states that strains belonging to groups A and B1 appear to be generalists as they appear to cover a larger range of environments. In contrast, strains belonging to groups B2 and D appear to be more specialised.

Phylogenetic studies have previously been very time consuming and complex because of the need for markers derived from MLEE and ribotyping (Grimont and Grimont, 1986) data. However, a rapid and simple PCR-based method to accurately and effectively group *E. coli* based on phylogeny was developed by Clermont et al. (2000). They used the 72 ECOR strains (Ochman and Selander, 1984), together with diverse *E. coli* strains (i.e. causing neonatal meningitis and neonatal septicaemia, as well as verotoxin producing *E. coli* and *E. coli* from faeces of healthy neonates) to develop three sets of PCR primers for separating strains into their respective phylogroups. The three primer sets target two genes (i.e., *ChuA* and *YjaA*) and an anonymous DNA fragment named TspE4C2. *ChuA* was discovered in enterohemorhagic O157:H7 *E. coli* and the coded protein is involved in heme transport. *YjaA* was identified in the genome sequence of *E. coli* K-12 and its function is still unknown.

Based on the presence or absence of these three diagnostic markers in triplex assays, *E. coli* can be effectively grouped into groups A, B1, B2 and D. The presence of *ChuA* gene designates strains to either phylogroup B2 or D. The presence of the *YjaA* gene then differentiates phylogroup B2 from D and is present in most strains belonging to phylogroup A. Lastly, the presence of the TSPE4.C2 fragment differentiates phylogroup B1 from A, being present in all B1 strains (Clermont et al., 2000 and Gordon et al., 2008). The discovery of these markers allowed for the phylogenetic grouping *E. coli* strains based on the presence or

absence of these three markers. The accuracy obtained for grouping *E. coli* strains was more than 99%, proving that this method can rapidly and effectively group *E. coli* strains comparative to previous methods (Clermont et al., 2000).

Walk et al. (2007) used the triplex assays in conjunction with multi-locus enzyme electrophoresis and multi-locus sequence analysis to determine the population structure and genetic diversity of *E. coli* from six freshwater beaches in the state of Michigan in the USA. They discovered that *E. coli* isolated from the secondary environment belonged predominately to phylogroup B1, suggesting that specific genotypes were favoured by natural selection. Therefore, this B1 phylogroup may have acquired special attributes that have allowed it to survive in the external environment.

In comparison to multi-locus sequence typing, the PCR triplex method was shown to be an effective method for rapid classification of *E. coli* isolates based on phylogeny (Gordon et al., 2008). However, Gordon et al. (2008) discovered an inconsistency with strains that failed to produce any PCR products for the two genes (*ChuA* and *YjaA*) and the anonymous DNA fragment (TSPE4.C2). These strains are assigned to phylogroup A to which they seldom belong and should not be assigned to a phylogroup. They concluded that the Clermont method is a great way to rapidly group E. coli strains based on phylogeny. Overall 85% of strains were correctly assigned to phylogroups.

2.6.3 Pulsed Field Gel Electrophoresis (PFGE)

Pulsed Field Gel Electrophoresis is commonly used to differentiate between *E. coli* strains. PFGE allows for high discrimination between closely related strains when compared to some PCR techniques, such as rep-PCR and enterobacterial repetitive intergenic consensus polymerase chain reaction (ERIC-PCR) (McLellan et al., 2003). PFGE was originally developed by Schwartz and Cantor, (1984) where DNA molecules up to 20 000 kb can be separated by making use of agarose gel electrophoresis in which the electric field is applied in different directions. This allows for the separation of very large pieces of DNA as opposed to the standard gel electrophoresis (Olive and Bean, 1999, Ribot et al., 2006). Following restriction digestion of genomic DNA with soft agarose plugs, the unique restriction patterns of each isolate are then compared to one another to determine relatedness (Tenover et al., 1995). PFGE is often used to measure the degree of relatedness among strains of the same species and Böhm and Karch (1992) successfully used PFGE to subtype *E. coli* O157:H7 strains isolates from different geographical regions. They were able to identify clinical strains without any previous knowledge of serotypes. McLellan et al. (2003) showed that fingerprints generated from PFGE gave higher resolution than fingerprints produced by rep-PCR when characterising *E. coli* populations from host sources of faecal pollution. PFGE was able to detect single base pair changes resulting in highly diverse fingerprint patterns with only a few common fragments, making it useful when differentiating between strains of the same species. However, rep-PCR and other PCR-based methods may be more practical when approaching larger datasets (McLellan et al., 2003).

2.6.4 Amplified Fragment Length Polymorphism (AFLP)

An alternative method for characterising the *E. coli* populations is Amplified Fragment Length Polymorphism (AFLP) (Vos et al., 1995). AFLP is a PCR-based DNA fingerprinting method proven important in genotypic analysis. This method requires no prior knowledge of the DNA sequence and can simultaneously detect polymorphisms in different genomic regions at the whole genome level. AFLP is also robust technique with high discriminatory power for bacterial strains below the species level (Brady et al., 2007, Dijkshoorn et al., 2001, Hahm et al., 2003, Leung et al., 2004).

In a study by Gaun et al. (2002) AFLP proved to be the most effective method in differentiating *E. coli* isolates from human and animal sources. In contrast to multiple-antibiotic resistance (MAR) profiles and 16S

rRNA sequence analysis, AFLP correctly classified over 96% of the *E. coli* isolates showing the highest level of discriminatory power among the methods investigated. In addition, Leung et al. (2004) set out to determine the capability of AFLP to differentiate *E. coli* strains, isolated from various geographical regions, based on pathogenicity and host source. In comparison to ERIC-PCR (Hulton et al., 1991), they discovered that AFLP was extremely effective in discriminating *E. coli* strains in terms of host source and pathogenicity.

Hahm et al. (2003) compared methods for subtyping *E. coli* isolates where comparisons were made using multiplex-PCR (Paton and Paton, 1998), rep-PCR (Rademaker et al., 1998), pulse-field gel electrophoresis (PFGE; see above), ribotyping and AFLP. These methods have the maximum potential for strain discrimination, only differing based on the genetic polymorphism being considered. These methods are preferred because of their high discriminatory power, their speed, ease and potential for large-scale screening. Hahm et al. (2003) discovered that the methods were unable to group the isolates identically because they all differed in the genetic polymorphisms they detect, inferring different phylogenetic relationships. PFGE showed the best results in discriminating between subtypes although it is very time consuming, whereas rep-PCR was the quickest and the easiest (McLellan et al., 2003, Ishii and Sadowsky, 2009). AFLP is believed to give similar results to PFGE and is the most flexible due to the range of primers available. This is similar to what was found by Jonas et al. (2003) where AFLP was also found to have the greatest discriminatory power for typing *E. coli* isolates.

2.6.5 Multilocus Sequence Typing (MLST)

In the late 1990's MLEE was replaced by Multilocus Sequence Typing (MLST) (Maiden et al., 1998). MLST has become widely used in characterising various bacterial species. It is based on the same principles as MLEE but rather than differentiating strains based on the electrophoretic mobility of their gene products, it identifies differences in the nucleotide sequences of chromosomal housekeeping genes. MLST has a number of advantages over MLEE, it has better resolution, it is based on DNA sequence information, and results can therefore be standardized. It can also be automated and results are clear-cut (Dijkshoorn et al., 2001, Tenaillon et al., 2010, Walk et al., 2007). Although MLST is best for population genetic studies, it often lacks discriminatory power to differentiate some bacterial strains, due to sequence conservation of housekeeping genes.

MLST data can be analysed in two ways: allele numbers are assigned to unique sequences and combined to form an allelic profile, which determines the sequence type (ST). Therefore, strains sharing the same alleles at all loci are considered to belong to the same sequence type. The number of nucleotide differences at each allele is not taken into account. Downstream analysis of MLST is then based on allele numbers and sequence types. Alternatively, in MLSA the actual nucleotide sequences of each gene are used in downstream phylogenetic analysis (Tenaillon et al., 2010). Application of MLST usually pertains to strains that belong to defined species whereas MLSA is used to improve species descriptions when species boundaries are unclear.

Using an extended MLST approach, Walk et al., (2009) identified and characterised five novel *Escherichia* clades (CI to CV). In their study, they included a range of known species of *Escherichia* and *Shigella*, as well as isolates collected from human, animal and, different from most previous studies, various environmental sources. Based on the DNA sequence information for 22 conserved genes, they were able to show that each of *E. coli*, *E. albertii*, *E. fergusonii* and *Salmonella enteric*a formed monophyletic clades. The remaining monophyletic clusters were named Clade I to Clade V and all of the five clades grouped more closely to *E. coli* than *S. enterica*.

Of the five clades identified by Walk et al. (2009), CI and *E. coli* were identical at most of the 22 loci investigated whereas the remaining *Escherichia* species and clades were monophyletic. This suggests that although *E. coli* and CI have had sufficient time to diverge, various evolutionary processes, such as

recombination, mutation and natural selection, have been acting in maintaining their similarities. In spite of this, Walk et al. (2009) maintain that these emerging clades are a result of those same evolutionary processes. In contrast to *E. coli sensu stricto* and CI, CII and specifically CV are more phylogenetically distinct. CV differs more than *E. fergusonii* and is almost as distinct as *E. albertii*. Here Walk et al. (2009) concluded that CV represents a rare "living fossil" of *E. coli*. In contrast to CI, CIII, CIV and *E. coli sensu stricto*, which are considered young lineages, CV is one of the oldest *Escherichia* lineages.

In addition, clades CIII, CIV and CV were identified as environmental representatives as isolates belonging to these clades were isolated from a variety of sources including surface water, freshwater beaches and various environmental samples. This suggests that these novel clades may have an extensive habitat range. The fact that the novel clades are hard to differentiate from *E. coli* based on traditional phenotypic analysis they demonstrate highly variable genotypes and evolutionary histories. These observations support the growing body of evidence of *E. coli* in the environment outside the host and its varied population structure (Gordon et al., 2002, Ishii et al., 2006, Power et al., 2005, Solo-Gabriele et al., 2000, Walk et al., 2007).

2.7 FUTURE PROSPECTS

The presence of *E. coli* in water systems is a human health risk and *E. coli* is currently used as an important indicator organism to assess water quality. The use of *E. coli* as an indicator organism is based on the assumption that it does not survive for long periods outside of the mammalian gastrointestinal tract and therefore its presence in the water is indicative of recent faecal contamination. Although most human associated *E. coli* may not survive for long periods, there is recent evidence that some *E. coli* strains are capable of surviving in water systems for longer periods and in the absence of any obvious faecal contamination (Gordon et al., 2002; Power et al., 2005; Solo-Gabriele et al., 2000, Walk et al., 2007).

Further indications of the existence of environmental *E. coli* stains come from the drinking water supply industry. A number of water suppliers (e.g. Rand Water; Johannesburg Water) have reported the occasional occurrence of *E. coli* in water distribution networks without any indication of potential faecal contamination. The presence of *E. coli* in these water systems results in the suppliers applying corrective actions at great cost. This may be unnecessary if these *E. coli* isolates represent unique environmental clones not associated with faecal contamination. If unique environmental *E. coli* populations do exist, they would render *E. coli* unreliable as a faecal indicator.

Based on the information currently available, it is likely that unique environmental *E. coli* strains exist in the apparent absence of any faecal contamination and without being associated with a primary host. Furthermore, these *E. coli* may be genetically different from their commensal and pathogenic counterparts as a consequence of their adaptation to the external environment. If this is indeed the case, the suitability of *E. coli* as an indicator organism is highly questionable. If environmental strains can be effectively characterised or identified, then it may save the water industry a considerable amount of time and costs involved in increasing the treatment of supposedly contaminated water. Apart from these economic issues, there are also social and health implications because the main assumption is that the presence of faecal indicators is indicative of an increased human health risk.

The presence of *E. coli* in the secondary environment raises several questions: if *E. coli* persists in soil and sediments, where else are they able to survive? What mechanisms enable these *E. coli* to survive? In addition, what makes them different from other *E. coli*? Future studies should employ comparative genomics to shed light on the mechanisms involved in *E. coli* evolution and adaptation to other environments.

CHAPTER 3: DIVERSITY OF *ESCHERICHIA COLI* ASSOCIATED WITH AQUATIC ENVIRONMENTS

3.1 INTRODUCTION

Although most human associated *E. coli* may not survive for long periods, the review of our current knowledge presented in the previous chapter has highlighted recent evidence that some *E. coli* strains are capable of surviving in water systems for longer periods in the absence of any obvious faecal contamination (Gordon et al., 2002; Power et al., 2005; Solo-Gabriele et al., 2000, Walk et al., 2007). Understanding the diversity, ecology, population structure and relationship between *E. coli* strains is therefore important when addressing the suitability of *E. coli* as indicator to determine water quality. To study the diversity and dynamics of an *E. coli* population in an aquatic ecosystem the Rietvlei dam was selected as study site. It had all the characteristics typically associated with such impoundments where habitats such as sediments, algae and water plants form unique niches but are also in continuous contact with the overall water body, with no physical barriers between them.

The aim of the work reported in this chapter was to determine whether the *E. coli* diversity in the aquatic environment only reflects the diversity of *E. coli* found in humans and warm-blooded animals or if separate and genetically distinct environmental *E. coli* populations also exist in aquatic habitats. Based on the initial survey, the diversity of *E. coli* associated with aquatic plants obtained from various impoundments where also further investigated and compared to the human isolates.

3.2 MATERIALS AND METHODS

3.2.1 Site description and sampling

The Rietvlei Dam is situated in a nature reserve on the Hennops River. The dam supplies about 10% of the drinking water that Pretoria requires through the Rietvlei Dam Water Treatment Works operated by the Tshwane Metropolitan Municipality (Bodenstein et al., 2006). The dam is also important for recreation as it has a yacht and canoe club. The water of the Rietvlei Dam has low nitrate-nitrogen levels, implying that the water is relatively unpolluted. The dam was ideally suited for this study because it has only limited sewage treatment works and urban areas draining into the dam.

3.2.2 Sample collection

The sampling points are indicated on the map (Figure 3.1). Samples were also taken from the inflow and final effluent of the Hartebeestfontein sewage treatment plant in this catchment to get strains that represent the *E. coli* dominant in the surrounding human population.

Based on the fact that the initial results indicated that unique *E. coli* isolates were associated with aquatic plants in the Rietvlei Dam, additional plant samples were obtained. Monitoring staff of the Department of Water Affairs collected 30 aquatic plant samples at 7 other impoundments on two different occasions. These dams included the Rietvlei, Hartbeespoort, Leeukraal, Bon Accord, Roodekopjes, Roodeplaat, Buffelspoort and Klipvoor Dam (Figure 3.2). A number of other *E. coli* strains obtained from laboratories testing the drinking water distribution networks in the vicinity of the dam were also included.

Figure 3-1: Map of Rietvlei Dam indicating sampling sites. (© Google Maps)

Figure 3-2: Map indicating the location of the dams from which aquatic plants were collected. (© Google Maps)

3.2.3 Bacterial isolations

For the raw and final sewage effluent samples, serial dilutions up to 10^{-6} were prepared using sterile quarter strength Ringer's solution (Merck). One hundred microliters of each dilution was plated out on MLGA (Membrane Lactose Glucuronide Agar) (Oxoid) containing 40 g peptone, 6 g yeast extract, 30 g lactose, 0.2 g phenol red, 1 g sodium lauryl sulphate, 0.5 g sodium pyruvate, 10 g agar and 0.2 g X-glucuronide per litre. After 100 µℓ of the samples were plated out, these plates were incubated at 37°C overnight. Other water samples were diluted to 10^{-4} and 100 uℓ was also plated. Samples not yielding any colonies were concentrated by filtering at least 1 mℓ of each sample through a 0.45 µm membrane filter (Whatman). The filters were placed on MLGA and incubated at 37°C overnight.

For the isolation from aquatic plants, a piece of the plant was removed aseptically and placed in 50 mℓ sterile quarter strength Ringer's solution (Merck). After the sample was shaken for a period, 10 mℓ of the Ringers solution was filtered through a 0.45 µm membrane filter (Whatman). The filters were placed on MLGA and incubated at 37°C overnight.

Green colonies were assumed to be *E. coli*, and were randomly selected from the plates. These cultures were then streaked out for single colonies on MLGA, and incubated at 37°C overnight. A number of yellow colonies with a possible green shading were also selected. Each of the colonies selected were verified using Colilert™ (Dehteq). A single colony was picked and added to 5 mℓ of Colilert™ solution in a sterile test tube and incubated at 37°C for 18 h. Colonies that turned the broth yellow and resulted in fluorescence were included in the study. The isolates were stored at -70°C according to the manufacturer's instruction by making use of Microbank™ (Pro-lab Diagnostics) beads.

3.2.4 DNA extractions

DNA was extracted from all the strains included in the study according to manufacturer's instruction using the ZR Genomic DNA II Kit (Zymo Research). DNA was released from the cells with the genomic lysis buffer. This was followed by two wash steps. The DNA was subsequently eluted using 50 µℓ DNA elution buffer and stored at -20^{o}C.

3.2.5 Determination of the phylogenetic groups of the isolates

The protocol described by Clermont et al. (2000) was followed with various modifications. DNA amplification was performed using three separate primer pairs shown in Table 3.1 as opposed to the original triplex method described by Clermont et al. (2000).

Table 3-1: Primer pairs used in the determination of phylogenetic groups described by Clermont et al. (2000)

Primer name	Primer sequence	Expected amplicon size
chuA.1	5'-GAC GAA CCA ACG GTC AGG AT-3'	279 bp
chuA.2	5'-TGC CGC CAG TAC CAA AGA CA-3'	
yjaA.1	5'-TGA AGT GTC AGG AGA CGC TG-3'	211 bp
yjaA.2	5'-ATG GAG AAT GCG AAC CTC AAC-3'	
TspE4C2.1	5'-GAG TAA TGT CGG GGC ATT CA-3'	152 bp
TspE4C2.2	5'-CGC GCC AAC AAA GTA TTA CG-3'	

Each 20 µℓ PCR reaction contained the following: 10 X Reaction buffer, 2.5 mM $MgCl_2$, 250 µM of each nucleotide (dATP, dCTP, dGTP and dTTP), 10 µM of each primer pair, 2.5 U of Taq DNA polymerase (Southern Cross Technologies) and 50 – 100 ng of genomic DNA. The PCR conditions involved an initial denaturation at 94°C for 4 minutes followed by 35 cycles of denaturation of 94°C for 5 seconds, annealing at 55°C for 10 seconds and lastly a final extension of 72°C for 5 minutes. A negative control for each PCR reaction was included where the genomic DNA was substituted with nuclease free water (Promega). Amplification was performed using a Veriti™ Thermal Cycler (Applied Biosystems). The PCR products were mixed with 1 µℓ gel red (Biotium) per 5 µl PCR product and visualised on a 1% agarose (WhiteSci) gel in 1X TAE buffer along with a 1000 bp marker (Fermentas). Electrophoresis occurred at 80 V for 30 minutes and the results were visualised and recorded under UV light. Isolates were then assigned to phylogroups based on the scheme that used the presence or absence of the amplicons to determine the specific phylogroup (Table 3.2).

Table 3-2: Phylogroup assignment of *E. coli* according to the method proposed by Clermont (Clermont et al., 2000) and adapted by Gordon et al. 2008

chuA	*yjaA*	TSPE4.C2	Phylogroup
-	-	-	Unknown
-	+	-	A
-	-	+	B1
+	+	-	B2
+	+	+	B2
+	-	-	D
+	-	+	D

3.2.6 Sequencing of core genes

The initial screening of the diversity represented by the selected *E. coli* isolates was based on the *rpoS* gene sequences of all the isolates. The *rpoS* gene is commonly used for this type of analysis and encodes for the RNA polymerase sub-unit sigma factor 38. The gene was amplified by the polymerase chain reaction (PCR) using the primers indicated in Table 3.2 (Walk et al., 2009). The PCR was set up as follows: 2.5 µℓ Reaction buffer (10X) (Southern Cross Biotechnology), 2 µl of 25 mM $MgCl_2$ (Southern Cross Biotechnology), 2.5 µℓ dNTPs (100mM), 0.5 µℓ of the 10 pmol forward primer (Inqaba Biotec), 0.5 uℓ of the 10 pmol reverse primer (Inqaba Biotec), 0.3 µℓ of the 5 U/µℓ Super-therm polymerase (Southern Cross Biotechnology), 1 µℓ DNA template and 15.7 µℓ nuclease free water (Promega) making up a final volume of 25 µℓ. Amplification was done using the Eppendorf thermocycler (Eppendorf). The amplification cycle was initiated with 94°C for 10 min, followed by 30 cycles of denaturing at 92°C for 1 min, annealing at 60°C for 1 min and extension at 72°C for 1 min. The final extension was done at 72 °C for 5 min.

The amplified product was then visualized by loading 5 µℓ of each sample combined with 1 µℓ of gel red loading dye (Biotium) on a 1% agarose gel made up with TAE buffer (pH 8). The gel was run for 30 minutes at 80 V. The bands were visualized under UV light, band size ranges from 576 bp to 618 bp. For PCR clean-up of the 20 µℓ of the PCR mixture 0.5 µℓ Exonuclease 1 (20 U/µℓ) (Fermentas) and 2 µℓ Fast AP Alkaline Phosphatase (1 U/µℓ) (Fermentas) was added to the PCR mixture. The mixture was incubated at 37°C for 15 min and then at 85°C for 15 min.

The sequencing reaction was set up using 2.5 µℓ sequencing buffer (5X) 0.5 µℓ ABI PRISM BigDye v3.1 (Life Technologies), 0.3 µℓ undiluted forward primer (100 µmol) of each primer pair (Table 3.2), 4.7 µℓ nuclease free water (Promega) and 4 µℓ of the DNA template that was cleaned in the previous step. The reaction was

run starting with 96°C for 5 s, followed by 25 cycles of 96°C for 10 s, 55°C for 5 s and 60°C for 4 min and 15 s. The resulting PCR products were sequenced with an ABI Prism DNA Automated Sequencer (Life Technologies).

Another three core genes showing greater variability than *rpoS* were also selected based the gene selected for a comprehensive MLST scheme designed by Walk et al. (2009). Each of the 22 genes was individually evaluated to determine the highest variability within the genes. This meant that the published genomes and data banks were screened to locate representative gene sequences for each of the 22 housekeeping genes. At least ten sequences were downloaded from Genbank (www.ncbi.nlm.nih.gov/genbank/) for each of the selected genes. The sequences were then aligned and the overhangs were trimmed. The aligned sequences were opened in MEGA5. The genes selected were the *uidA* (β-glucuronidase), *mutS* (methyl-directed mismatch repair) and *fadD* (fatty-acyl CoA synthetase). All three genes were amplified and sequenced using the same methodology described above for the *rpoS* gene. The primers used (Table 3.2) were previously used to study *E. coli* populations (Walk et al., 2009) and formed part of the procedures described by the online MLST database EcMLST (www.shigatox.net).

Table 3-3: Primers used for the amplification of the *E. coli* core genes (Walk et al., 2009)

Primer name	Primer sequence	Expected amplicon size
**rpoS*	5'-CGC CGG ATG ATC GAG AGT AA-3'	618 bp
rpoS	5'-GAG GCC AAT TTC ACG ACC TA-3'	
**uidA*	5'-CAT TAC GGC AAA AGT GTG GGT CAA T-3'	658 bp
uidA	5'-TCA GCG TAA GGG TAA TGC GAG GTA-3'	
**mutS*	5'-GGC CTA TAC CCT GAA CTA CA-3'	596 bp
mutS	5'-GCA TAA AGG CAA TGG TGT C-3'	
**fadD*	5'-GCT GCC GCT GTA TCA CAT TT-3'	580 bp
fadD	5'-GCG CAG GAA TCC TTC TTC AT-3'	

* Indicates forward primer

3.2.7 Determining the phylogenetic relationships amongst isolates

Once the *E. coli* core genes were sequenced for all isolates, the sequences were viewed in BioEdit (Hall, 1999). The sequences were aligned using both ClustalW (Thompson et al., 1994) multiple alignment and MAFFT (Katoh et al., 2009) and the overhangs were then trimmed in BioEdit. The model that fitted the data best was determined using jModelTest software v. 0. 1. 1. (Posada, 2008). Maximum-Likelihood trees (Felsenstein, 1981) with 1000 bootstrap replicates were drawn for each of the genes respectively using PhyML 3.0 (Guindon et al., 2010). *E. fergusonii* was used as the outgroup, except for the *uidA* tree where *Shigella dysenteriae* was used.

3.3 RESULTS

3.3.1 Isolates

During the initial sampling a total of 97 *E. coli* strains were collected from the Rietvlei dam and linked environments. Of these 58 were from the dam water and sediments, 14 were isolated from decaying plant material floating on the dam. A further 16 isolates were obtained from the raw sewage and final effluent of

the Hartebeestfontein treatment works and 9 were isolated from the drinking water distribution network. A list of isolates and the sampling point from which they were sampled are provided in Table 3.4.

After the initial sampling of Rietvlei Dam the Department of Water Affairs assisted with the sampling of aquatic plants from another 7 impoundments in the Highveld region on two further occasions. Four aquatic plant species were sampled and a total of 81 isolates were collected. A list of isolates and the dam and plant species from which they were isolated are provided in Table 3.5.

Table 3-4: List of isolate names, sample types and sampling location for the *E. coli* isolated from the Rietvlei Dam

Isolate Name	Sample type	Sample point
S211G, S212G, S213G, S214G, S215G, S216G, S217G, S218G, S219G, S2110G, S2111G, S2112G, S2113G, S2114G, S226Y, S2F11G, S2F12G, S2F13G, S2F14G, S2F15G, S2F16G, S2F17G, S2F18G, S2F110G, S2F111G, S2F112G, S2F113G, S2F114G, S2F114G, S2F21G, S2F22G, S2F23G, S3F21G, S3F22G, S3F23G	Sediment	Rietvlei Dam
DWWF12G, DWWF13G, DWWF14G, DWWF21G, DWWF22G, DWWF23G, DWWF24G, DWWF25G, DWWF26G, DWWF27G, DWWF28G, DWWF29G, DWWF210G, DWWF211G	Decaying plant	Rietvlei Dam
RAW42G, RAW45G, RAW46G, RAW48G, RAW49G, FINAL21G, FINAL22G, FINAL23G, FINAL24G, FINAL25G, FINAL26G, FINAL27G, FINAL28G, FINAL29G, FINAL210G, FINAL15Y	Sewage	Hartebeestfontein sewerage treatment works
TS1A, TS1B, TS2A, TS2B, TS3A, TS3B, TS4B, TS5A, TS5B, TS6A, TS6B, TS7A, TS7B, TS8B, TS9A, TA10A, TS11A, TS12B, TS13A, TS14B, TS15A, TS15B, TS16A, TS17A, TS17B, TS18A, EP141,	Dam water	Rietvlei Dam
AD11G, ADBF11G	Algae	Rietvlei Dam
EP153, EP154, EP168, SP004, SP1000, NP720, NP707, NP705	Drinking water	Drinking water distribution networks

Table 3-5: List of plant codes, number of isolates and plant hosts for *E. coli* strains isolated from aquatic plants sampled from different dams in the Highveld region

Plant sample	Number of isolates	Scientific name	Common name	Dam
RKD 3	1	*Ceratophyllum demersum*	Water hornwort	Roodekopjes
RKD 2	1	*Myriophyllum aqauticum*	Parrot's feather	Roodekopjes
RKD 1	2	*Myriophyllum aqauticum*	Parrot's feather	Roodekopjes
BA1	4	*Ceratophyllum demersum*	Water hornwort	Bon Accord
BAD 5	2	*Myriophyllum aqauticum*	Parrot's feather	Bon Accord
BAD 4	2	*Myriophyllum aqauticum*	Parrot's feather	Bon Accord
BAD 3	2	*Myriophyllum aqauticum*	Parrot's feather	Bon Accord
BAD 2	2	*Myriophyllum aqauticum*	Parrot's feather	Bon Accord
BAD 1	2	*Myriophyllum aqauticum*	Parrot's feather	Bon Accord
KVD 5	2	*Myriophyllum aqauticum*	Parrot's feather	Klipvoor
KVD 3	1	*Myriophyllum aqauticum*	Parrot's feather	Klipvoor
KVD 1	2	*Myriophyllum aqauticum*	Parrot's feather	Klipvoor
K1	2	*Myriophyllum aqauticum*	Parrot's feather	Klipvoor
HBPD 2	1	*Ceratophyllum demersum*	Water hornwort	Hartbeespoort
HBPD 1	2	*Ceratophyllum demersum*	Water hornwort	Hartbeespoort
H2	3	*Ceratophyllum demersum*	Water hornwort	Hartbeespoort
HBPD 5	1	*Isolepis fluitans*	Water weed	Hartbeespoort
H3	2	*Isolepis fluitans*	Water weed	Hartbeespoort
RVD 4	2	*Isolepis fluitans*	Water weed	Rietvlei
RVD 1	1	*Isolepis fluitans*	Water weed	Rietvlei
RV1	6	*Isolepis fluitans*	Water weed	Rietvlei
RVD 5	1	*Myriophyllum aqauticum*	Parrot's feather	Rietvlei
RVD 3	1	*Myriophyllum aqauticum*	Parrot's feather	Rietvlei
RVD 2	1	*Myriophyllum aqauticum*	Parrot's feather	Rietvlei
LKD 3	1	*Isolepis fluitans*	Water weed	Leeukraal
L1	4	*Isolepis fluitans*	Water weed	Leeukraal
Q02H	13	*Eichhornia crassipes*	Water hyacinth	Roodeplaat
R 1	9	*Isolepis fluitans*	Water weed	Roodeplaat
R 2	3	*Myriophyllum aqauticum*	Parrot's feather	Roodeplaat
B1	5	*Myriophyllum aqauticum*	Parrot's feather	Buffelspoort

3.3.2 Phylogrouping

The isolates collected during the initial phase of the study were all analyzed using the multiplex PCR to determine to which phylogenetic groups these strains belong. The results were analyzed according to the scheme presented in Figure 3.2. Of the total number of 88 isolates, 44 isolates (40.7%) were part of the phylogenetic group A, 38 isolates (35.2%) part of phylogenetic group B1, 15 isolates (13.9%) part of phylogenetic group B2 and 11 isolates (10.2%) formed part of phylogenetic group D.

3.3.3 Initial phylogenetic analysis of Rietvlei isolates

An initial phylogenetic analysis of the *rpoS* sequences of the 97 Rietvlei isolates indicated that all isolates belonged to *E. coli sensu stricto*. None of the isolates grouped with any of the 5 clades previously described by Walk and co-workers (Walk et al., 2009) to form part of the cryptic species of *E. coli*. Most of the isolates could also not be separated from the *E. coli* associated with humans (sewage isolates). It was, however, observed that most of the isolates obtained from aquatic plants and decaying plant material floating on the dam, clustered separately from the strains representing the *E. coli* associated with humans. This indicated that these isolates could belong to environmental populations. This observation directed the sampling strategy and extra *E. coli* isolates (Table 3.4) associated with aquatic plants were collected, sequenced and the data analysed to determine the diversity of these *E. coli* strains.

3.3.4 Selection of additional core genes

It is difficult to infer reliable phylogenetic relationships for a group of bacteria based only on one gene and it was decided to include core genes with greater variability in order to improve the level of resolution at which different *E. coli* populations could be differentiated. The percentage variability within 22 of the core genes of *E. coli* was determined based on at least 10 sequences for each gene and the results are shown in Table 3.6. The two most variable genes, *mutS* (methyl- directed mismatch repair) and *fadD* (fatty-acyl CoA synthetase) were selected. It was also decided to include the *uidA* (β-glucuronidase) gene as this is a gene unique to *E. coli*.

3.3.5 Phylogenetic analyses of selected core genes

Sequence data were obtained for all 178 strains isolated from the Rietvlei Dam and from aquatic plants sampled at 8 different localities. Based on this data, Maximum-Likelihood trees were constructed for each of the 4 selected gene regions. These trees are presented in Figure 3.3 to 3.6. The majority of *E. coli* isolates could not be separated from the sewage isolates but some potential environmental clusters could be observed.

Table 3-6: Nucleotide variability in 22 core genes of *Escherichia coli*

Gene	Number of Variable sites	% variability
mutS	54/380	14.21
fadD	53/483	11
cyaA	44/498	8.84
fumC	35/469	7.46
clpX	38/527	7.21
mltD	36/526	6.84
uidA	35/520	6.73
aroE	17/291	5.84
metG	21/442	4.75
mdh	24/526	4.56
dnaG	18/444	4.05
adk	19/530	3.58
icdA	10/300	3.33
gyrB	15/460	3.26
grpE	12/417	2.88
torC	15/520	2.88
recA	12/510	2.35
aspC	11/513	2.14
lysP	9/477	1.89
purA	9/478	1.88
kdsA	6/431	1.39
arcA	5/435	1.15
Gene already sequenced:		
rpoS	5/472	1.06

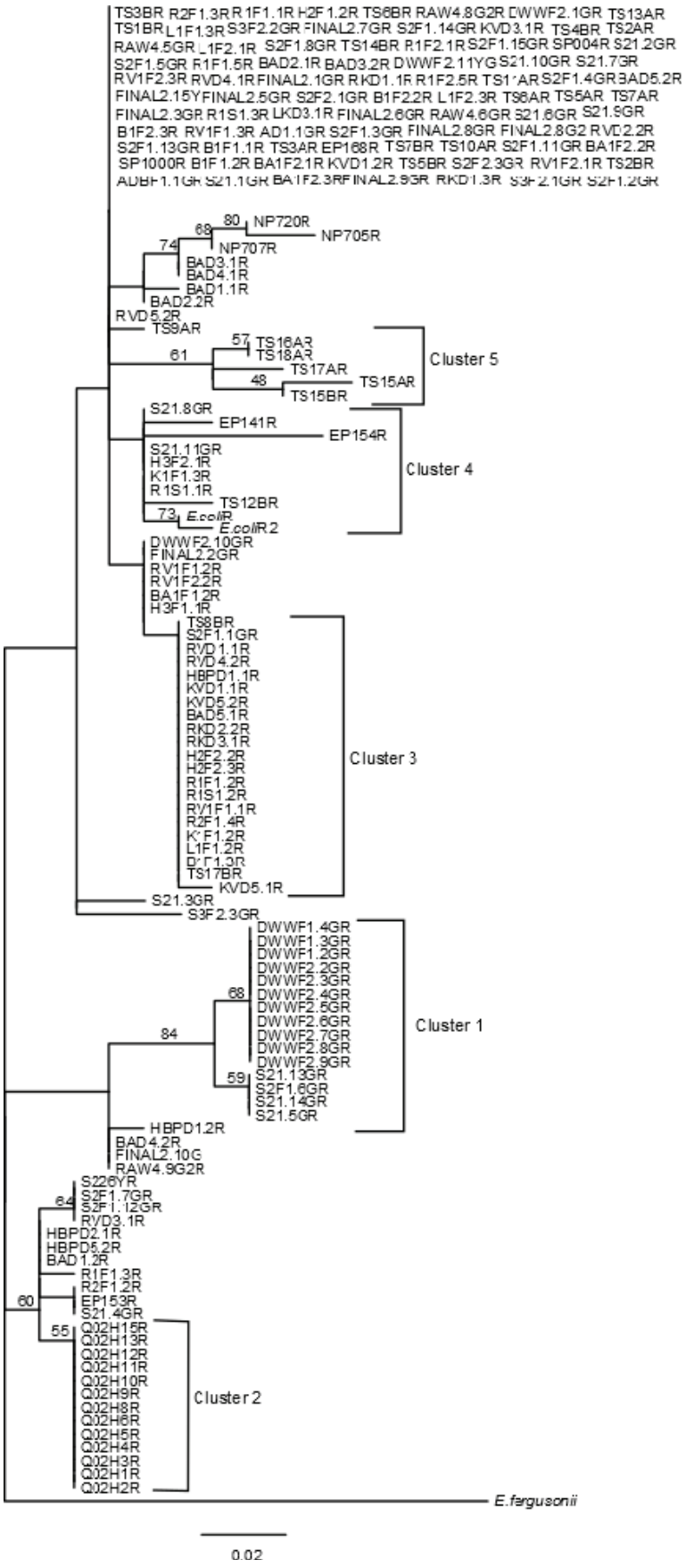

Figure 3-2: A Maximum-Likelihood tree based on the *rpoS* gene indicating the relatedness of *E. coli* isolates associated with aquatic environments.

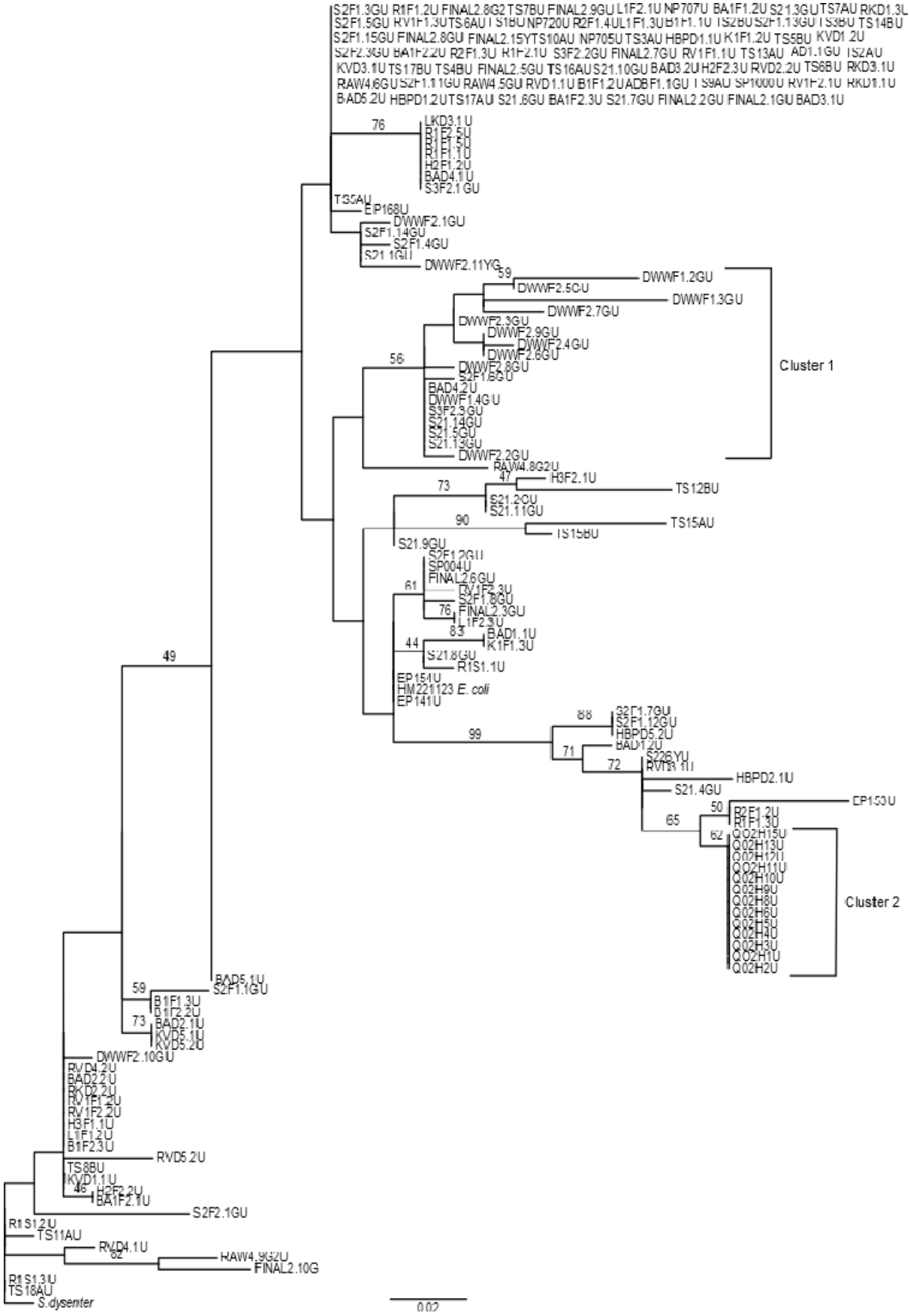

Figure 3-3: A Maximum-Likelihood tree based on the *uidA* gene indicating the relatedness of *E. coli* isolates associated with aquatic environments.

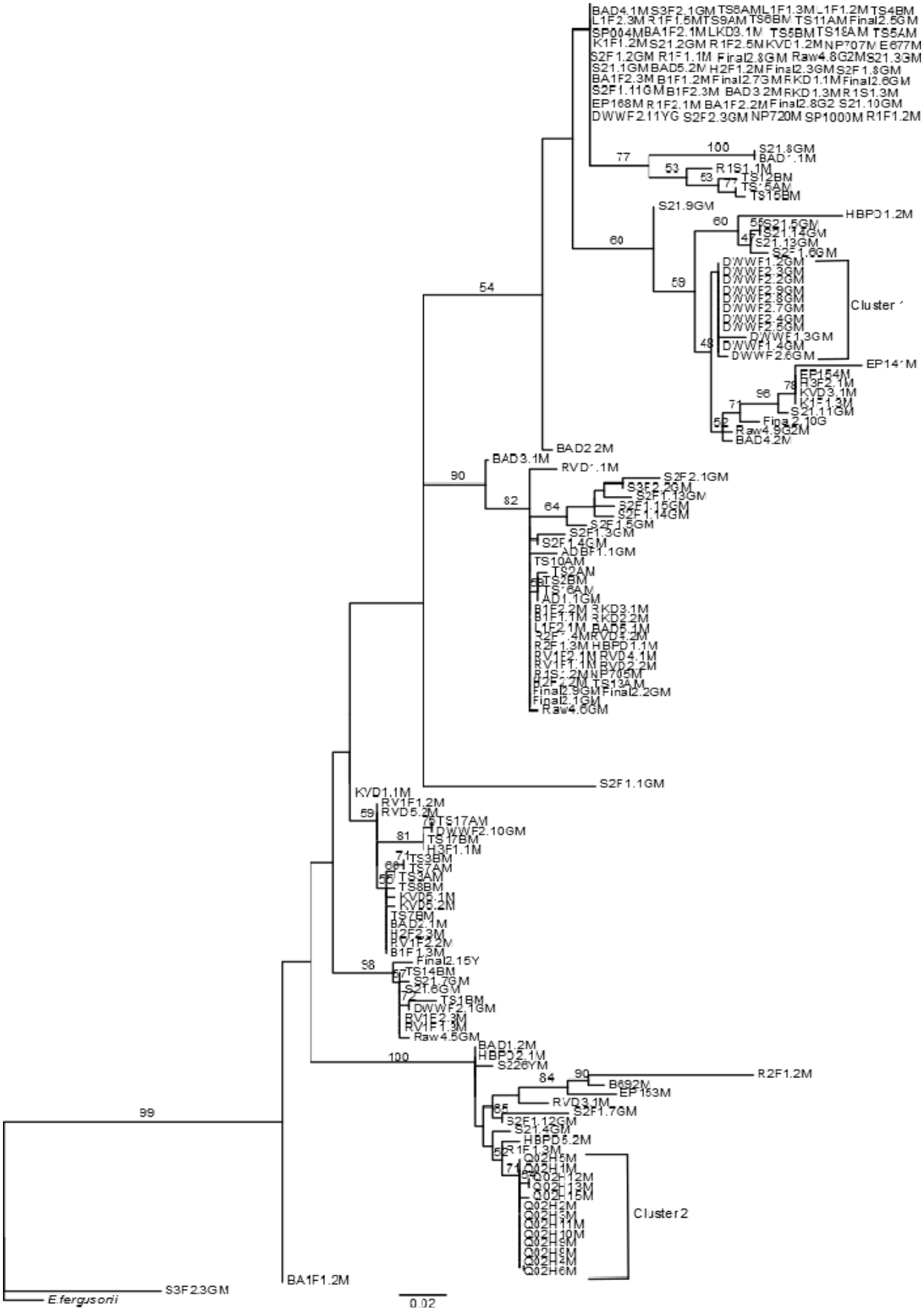

Figure 3-4: A Maximum-Likelihood tree based on the *mutS* gene indicating the relatedness of *E. coli* isolates associated with aquatic environments.

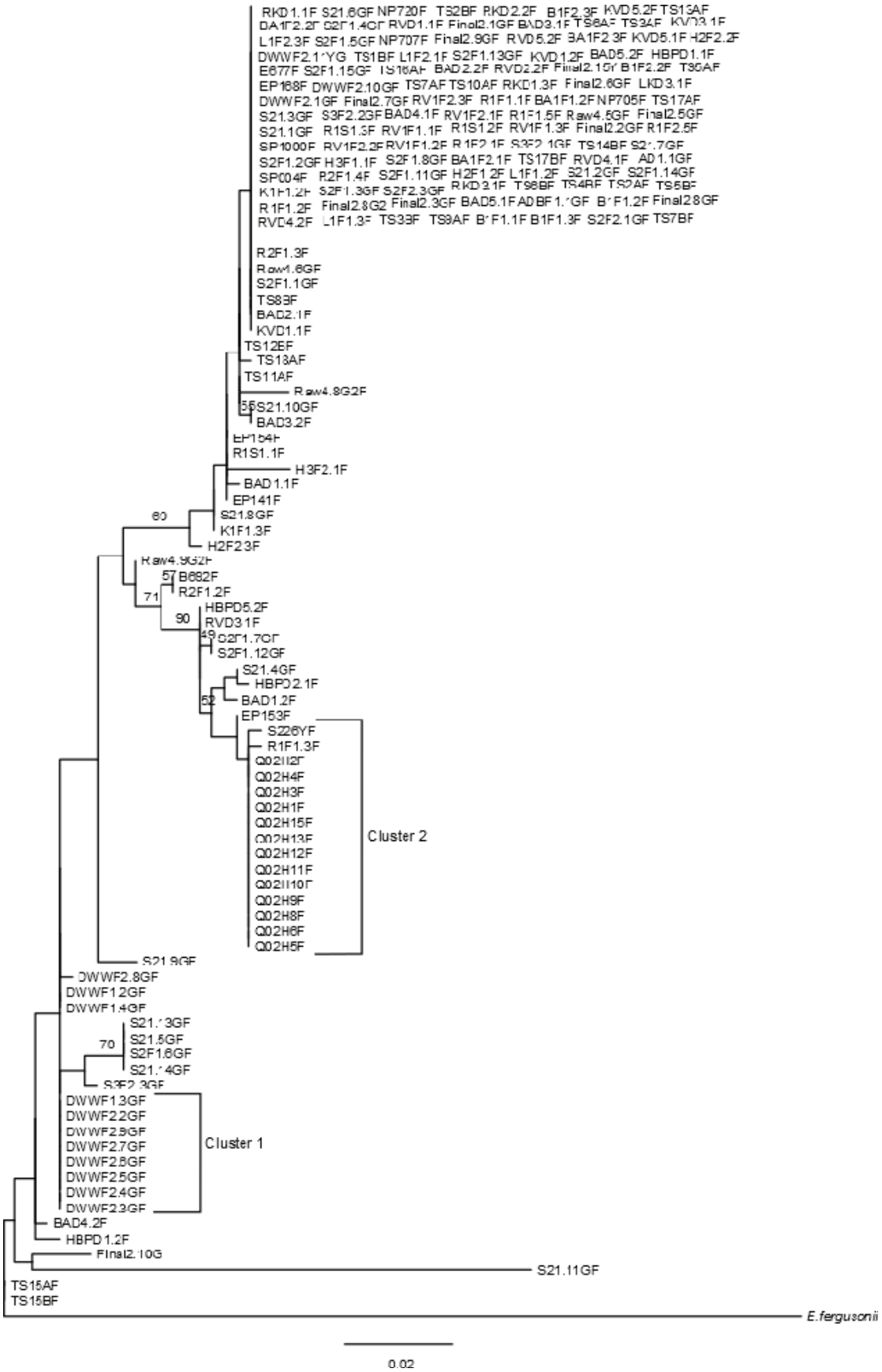

Figure 3-5: A Maximum-Likelihood tree based on the *fadD* gene indicating the relatedness of *E. coli* isolates associated with aquatic environments.

The *rpoS* tree revealed five potential environmental clusters, from which sewage isolates were absent (Figure 3.3) but only two of these clusters had reasonable bootstrap values to support this topology. Cluster 1 included eleven *E. coli* strains isolated from decaying plant material in the Rietvlei Dam. Although 4 strains isolated from the sediment also clustered close to these strains, inclusion in this cluster was not supported by two of the other gene phylogenies. Cluster 2 consisted of thirteen *E. coli* strains isolated from a water hyacinth in the Roodeplaat Dam and this cluster was well supported. Several strains from aquatic plants grouped in the vicinity of this cluster but their inclusion was again not supported by all the other gene phylogenies.

Resolution between *E. coli* isolates improved with the *uidA* and *mutS* sequence analysis (Figure 3.4 and Figure 3.5). As was expected the *mutS* gene showed the greatest improvement in resolution. In both of these trees, Cluster 1 and 2 were again well defined from the rest of the isolates and had the necessary support for the clusters. The *fadD* gene (Figure 3.6) did not show the expected variability and although Cluster 1 and 2 could be clearly observed, they were not well supported.

3.4 DISCUSSION

Based on their *rpoS* sequence, all of the 178 *E. coli* strains isolated during this study grouped with the "true" *E. coli*. None of the strains could be linked to the cryptic species (Clades I –V) described by Walk et al., (2009). The *rpoS* gene was selected as it was shown to group strains monophyletically in the various clusters observed during a study performed by Walk et al. (2009). It is unclear at present why strains belonging to the cryptic clades were not detected in this study. It is possible that these bacteria are present at lower levels in the environment and are easily masked by "true" *E. coli* strains during isolation and cultivation.

Strains belonging to the four phylogroups differ in their ecological niche and ability to grow at varying temperatures (Gordon et al., 2008). It has been stated that environmental *E. coli* are more likely to belong to phylogroups A and B1 (Gordon et al., 2008, Tenaillon et al., 2010). The results of this study support that statement as 70.5% of the initial Rietvlei *E. coli* isolates belong to phylogroups A and B1. The highest percentage of isolates (36.4%) belonged to phylogroup B1. Walk et al. (2007) stated that despite recombination events in nature, the phylogroup B1 is favoured by natural selection in the secondary environment.

Byappanahalli et al. (2006) observed that *E. coli* populations persisting in the secondary environment form cohesive phylogenetic groups when compared to faecal strains. Although, no comprehensive phylogenetic groups were observed similar to those observed by Byappanahalli et al. (2006), a number of strains isolated from the aquatic plants consistently grouped separately in the Maximum-Likelihood trees and formed two well defined clusters. The best resolution was observed for *mutS*, although *rpoS* and *uidA* provided similar results. It is unclear why the *fadD* gene did not show the expected variability for these aquatic isolates.

The possibility exists that the strains associated with the observed environmental clusters may have adapted to the environment outside of the primary host and found a unique niche within the aquatic system. *E. coli* is a highly diverse species with a flexible accessory genome allowing it to obtain specific genes in order to adapt to a various environments (Ihssen et al., 2007; Lukjancenko and Wassenaar, 2010). Horizontal gene transfer of beneficial genes within and between species in such an environment may well contribute to possible genetic differentiation when adapting to new niches.

3.5 CONCLUSIONS

It is well known that *E. c*oli is a highly diverse species. However, the majority of diversity studies are based on human, clinical and pathogenic strains. This work revealed that the *E. coli* population isolated from aquatic environments was very diverse. Sequence analysis of selected core genes proved an effective method in differentiating between *E. coli* strains. Although, many of the strains could not be distinguished from the sewage isolates, possible environmental clusters became apparent. The *E. coli* genome is shaped by its environment indicating that the majority of *E. coli* isolates obtained for this study are from the same environment or were at some point exposed to similar environmental conditions. However, the high level of diversity observed indicates that some *E. coli* strains may have become adapted to a niche within the aquatic environment.

Based on the study, possible *E. coli* environmental groups were observed within aquatic environments particularly those strains isolated from aquatic plants. The high level of diversity among the isolates and the indication of possible environmental groups raise questions concerning population structure and gene flow between strains. In the following chapter, these questions are addressed using an extended set of isolates.

CHAPTER 4: POPULATION STRUCTURE OF AQUATIC *E. COLI* STRAINS

4.1 INTRODUCTION

Many bacterial species can occupy a range of different environments resulting in populations that are complex and difficult to define (Nakamura et al., 2004, Lukjancenko and Wassenaar, 2010) due to their ease of dispersal and the genetic diversity of accessory genes (Lawrence and Hendrickson, 2005). For these reasons, bacterial populations rarely conform to the simplicity of some of the models commonly used to describe eukaryotic populations (Spratt and Maiden, 1999) Knowledge of the distribution and interaction of genes within a population is however, very important in understanding how processes such as natural selection, genetic drift, gene flow and recombination affect the overall evolution and ecology of a species. The short generation time of bacterial populations allows these evolutionary changes to be monitored on a feasible time scale.

Some bacterial populations conform to a clonal model where there is little or no genetic exchange or recombination and genetic change is introduced by the accumulation of mutations. Lineages within the population will then arise or fade out as a consequence of selection or other stochastic events (i.e., a random or unpredictable event such as bottle-necking). Conversely, populations may undergo frequent genetic exchange via horizontal gene transfer (HGT) and if individuals are in constant contact with each other there may be no restriction on gene exchange between individuals, resulting in a population with little or no structure. Most bacterial populations, including those of *E. coli*, are thought to fall somewhere between these two extremes, exhibiting some clonal structure due to recent clonal descent disrupted by varying degrees of horizontal gene transfer (Spratt and Maiden, 1999).

The population structure of a bacterial species is largely determined by its dispersal throughout an environment (Trevors, 1998). In aquatic environments, where bacteria constitute approximately 90% of the microorganisms (Hahn, 2006), the population structure of a species is shaped by factors such as water chemistry, water temperature, predation, nutrient availability, exposure to UV, protection and habitat size (Schauer et al., 2005, Hahn, 2006). According to Cohen (2002), bacteria occupying an ecological niche become genetically distinct or isolated from their neighbours in an adjacent niche, provided there is little or no gene exchange or recombination. Not only is gene flow limited physically when organisms occupy different niches, but different environments also dictate different modes of survival. If genetic exchange was absent between two individuals, then the possibility that those individuals would diverge from one another to form isolated populations and species is highly likely (Hahn, 2006). In general, freshwater bacteria share little or no significant overlap taxonomically with terrestrial or marine bacteria (Hahn, 2006). However, this may not be the case concerning *E. coli* as the members of this species are found in two drastically different environments (Savageau, 1983, Prosser et al., 2007).

The biphasic life style of *E. coli* suggests a complex interplay between the various processes determining its population structure. As *E. coli* cycles between the primary and secondary environments, its population structure is shaped by the phenotypic and genetic selective pressures inherently associated with both environments (Savageau, 1983). In the secondary environment, ecological differentiation at the intraspecific level may be driven by the differential adaptation to water chemistry and the geological backgrounds of the catchment area habitat (Schauer et al 2005). In the primary environment, ecological differentiation is probably driven by host-associated properties. In other words, *E. coli* populations in the primary environment undergo host-associated evolution followed by host-independent evolution once they are in the secondary

environment (Oh et al., 2012). It is unclear which one of these environments has the greater influence on the population structure of the species.

Ihssen et al (2007) showed that some *E. coli* genes are highly conserved in both environmental and human isolates. Using comparative genomic hybridisation and physiological characterisation, they revealed that genes specifically involved in carbon utilisation show little variation between environmental and human strains, suggesting that these catabolic pathways are maintained through vertical inheritance. However, the open pan-genome of *E. coli* may account for the metabolic and ecological diversity within the species. Based on 61 genomes analysed, the *E. coli* pan-genome consists of 15 000 unique genes (Lukjancenko and Wassenaar 2010) and will probably increase as more genomes are sequenced. This emphasises the potential for *E. coli* to adapt to various non-host environments.

Recent genomic studies have revealed that environmental and human isolates of *E. coli* have numerous genes specific to each set of strains. For example, the genomes of commensal *E. coli* encode for more genes associated with survival in the human gut (Luo et al., 2011). Furthermore, genomic studies of numerous *E. coli*-like strains isolated from environmental sources are distinct from human-associated *E. coli* (Oh et al 2012). In fact, these *E. coli*-like strains form a number of discrete lineages or clades that have apparently lost the ability to colonise the human host (Walk et al., 2009; Oh et al., 2012). Whole genome DNA-DNA hybridisation studies using a multi-genome *E. coli* microarray revealed that these *E. coli*-like environmental isolates lack sets of genes coding for stress response and defence mechanisms, and attachment to human epithelial cells (Luo et al., 2011). These data suggest that these *E. coli*-like strains originate from true *E. coli* through a type of reductive evolution by losing genes that were no longer needed in a specific niche.

Despite the discovery of unique *E. coli*-like lineages that potentially represent cryptic species of Escherichia (Walk et al., 2009), some strains obtained from environmental sources represent true *E. coli*. The work reported in Chapter 3 showed that all of the examined strains represented true members of this species, regardless of their source (e.g., sediments, dam walls, aquatic plants, open water). However, considerable genetic variation was observed among the various *E. coli sensu stricto* strains associated with the various sample types. In theory, the population of *E. coli sensu stricto* in this water body should be homogenous, as the main source of *E. coli* in aquatic environments is believed to be faecal material from humans and animals (Winfield and Groisman, 2003; Power et al., 2005). Because of the intrinsic differences between the primary and secondary environments of this bacterium, the possibility of population differentiation associated with specific niches within such a water system cannot be excluded.

The aim of this part of the project was to study the phylogenetic and population genetics of a collection of *E. coli sensu stricto* strains originating from two freshwater environments. For this purpose the DNA sequence information for two gene regions (*rpoS* and *uidA*) were used to calculate population genetic parameters reflecting gene flow and genetic differentiation. The specific questions addressed were as follows: (i) Are there unique populations of *E. coli* present in the aquatic environments that were sampled? (ii) Can the observed diversity of *E. coli* in aquatic environments primarily be attributed to human-linked contamination?

4.2 MATERIALS AND METHODS

4.2.1 Strains included in population studies

A dataset consisting of *rpoS* and *uidA* sequences for 293 *E. coli sensu stricto* strains were compiled. It consisted of the original isolates obtained during the sampling of the Rietvlei Dam (Chapter 3) as well as a similar collection obtained from the Roodeplaat Dam during a parallel study. The dam was originally

constructed as an irrigation dam and but has also become popular for recreational use. It is also an important source of water for the City of Tshwane. Roodeplaat Dam forms part of catchment area draining a large part of the City of Tshwane. There are two sewage treatment works in the vicinity, namely Zeekoegat and Baviaanspoort sewage treatment works, which both release their treated effluent into the dam. This causes highly eutrophic conditions, which result in blooms of algae and cyanobacteria and dense covering by water hyacinth (*Eichhornia crassipes*).

Water, algae, water hyacinth and sediment samples were collected at various depths around the dam. Algal, sediment and water samples were also collected from the Hartbeesspruit River leading into the Dam. For both Zeekoegat and Baviaanspoot sewage treatment works, samples were collected from both the raw sewage coming into the works and the treated effluent being released into the dam. Sewage samples were collected as to represent the *E. coli* strains circulating in the human and animal populations. A list of isolates and the sampling point from which they were sampled are provided in Table 4.1.

Table 4-1: List of isolate names, sample types and sampling location for the *E. coli* isolated from the Rietvlei Dam

Isolate Name	Sample type	Sample point
ZA1.2, ZA1.3, ZA1.4, ZA1.5, ZA1.6, ZA1.7, ZA1.8, ZA1.9, ZA2.1, ZA2.2A, ZA2.2B, ZA2.3, ZA2.4, ZA2.5, ZA2.6, ZA2.7, ZA2.9, ZB1.2, ZB1.3, ZB1.4, ZB1.5, ZB1.6, ZB1.7, ZB1.8, ZB1.9, ZB1.10, ZB2.1, ZB2.2, ZB2.3, ZB2.4, ZB2.5, ZB2.6, ZB2.7, ZB2.9, ZB2.10	Sewage	Zeekoegat sewage treatment works
B1.2, B1.3, B1.5, B1.6B, B1.7, B1.9, B1.10, B2.1, B2.2A, B2.2B, B2.3, B2.4, B2.5, B2.6, B2.7, B2.8, B2.10	Sewage	Baviaanspoort sewage treatment works
Q011, Q013, Q014, Q021, Q023, Q024, Q025, Q026, Q027, Q028, Q073, Q072, Q076, Q077, Q078, Q0710, Q082, Q083, Q084, Q085, Q086, Q087, Q088, Q0810, Q0811, Q0812, Q0813, Q091, Q093, Q094, Q095, Q096, Q097, Q098, Q099, Q0910, Q0911, Q0912, Q0913, Q0914, Q0915, Q105, Q106, Q107, Q109, Q1010A, Q1010B, Q1012, Q1013, Q1014, Q1015	Dam water	Roodeplaat Dam
Q02H1, Q02H2, Q02H3, Q02H4,Q02H5, Q02H6, Q02H8, Q02H9, Q02H10, Q02H11, Q02H12, Q02H13, Q02H15	Water Hyacinth (included in Chapter 3)	Roodeplaat Dam
Q09A2, Q09A3, Q09A4, Q09A5, Q09A6, Q09A8, Q09A9, Q09A10, Q09A11, Q09A12, Q09A13, Q09A14, Q09A15	Algae	Roodeplaat Dam
KW3, KW4, KW5, KW6A, KW6B, KW7, KW8, KW9, KW10, KW11, KW12A, KW12B, KW13, KW14, KW15	Water	Hartbeesspruit
KS1A, KS1B, KS2, KS3, KS4, KS5, KS6, KS7, KS8, KS9, KS10, KS11A, KS11B, KS12, KS13, KS14A, KS14B, KS15	Sediment	Hartbeesspruit
KA1A, KA1B, KA4, KA5, KA6, KA7, KA8A, KA8B, KA9, KA10 , KA11, KA12, KA13A, KA13B, KA14, KA15	Algae	Hartbeesspruit
JA1, JA2, JA3, JA4, JA5, JA6A, JA6B, JA7, JA8, JA9, JA10, JA11, JA12, JA13, JA14, JA15	Algae	Roodeplaat Dam Jeti

4.2.2 Population genetic analysis

To determine whether the population of *E. coli* is structured into sub populations, the software Structure (Version 2.3) (Prichard et al., 2000) was employed. Structure uses a model-based clustering method, which uses a Marcov Chain Monte Carlo (MCMC) methodology for inferring population structure (Falush et al., 2007). Structure software identifies genetically distinct populations and assigns individuals to populations based on genotype and allele frequencies. The admixture model was applied, assuming mixed ancestry where individuals within a population are thought to have inherited a fraction of its genome from an ancestor in the population. The aligned sequence information for the *uidA* and *rpoS* genes for all *E. coli* isolates collected from both the Roodeplaat and Rietvlei dams were used for the analysis. These datasets were formatted using MEGA (Version 5) (Tamura et al., 2011) and each nucleotide of the gene sequence was recognised as an individual locus. The Structure analysis was run assuming the presence of 1 to 20 populations (K=1 to K=20), with a burn-in length of 200 000 and a run length of 2 000 000.

The aligned *rpoS* and *uidA* gene sequences were also used to calculate gene flow and genetic differentiation using the software DnaSP Version 5.10.01 (Librado and Razos, 2009). For these analyses populations were defined based on the geographical location and sample site. Gene flow was analysed by comparing the F_{st} and Nm values between defined populations following Hudson et al. (1992). Genetic differentiation was measured by using the sequence-based statistic K_{ST}*, which gave an indication of population subdivisions. Statistical confidence for the genetic differentiation between populations was evaluated with permutation testing using 10 000 replicates.

4.3 RESULTS

4.3.1 Population structure analysis

To investigate the population structure and determine the number of populations present in the data set, the program Structure was used. Results indicated that all isolates belonged to one population (K=1). This result was observed for both the *rpo*S and *uidA* genes. The estimated ln probability of data and the variance of ln likelihood for K=1 were the lowest in both cases, therefore suggesting that all these isolates belonged to one population (Prichard et al., 2000). Values for both genes are presented in Table 4.2 and 4.3.

4.3.2 Gene flow and genetic differentiation

As Structure suggested the presence of only one population, it was necessary to investigate gene flow within this population. In F-statistics, the F_{st} value refers to the fixation index where it is a measure of the genetic differentiation between sub-populations. It provides a good idea as to the level of sub-population structure in any population. It is explained as the proportion of total genetic variance contained in sub-populations relative to the total genetic variance (t). An F_{st} value of zero indicates no genetic variation and no divergence between populations whereas an F_{st} value of one suggests total genetic differentiation where sub-populations become fixed and completely isolated. In terms of gene flow, if a gene is fixed and populations are isolated, there is total differentiation and the F_{st} value will approach one. Conversely, a low F_{st} value approaching zero indicates a high level of gene flow as there is little or no genetic differentiation. In addition, when considering gene flow between groups of populations, a second value to consider is the Nm value. Nm refers to the number of migrants successfully entering the population per generation (Whitlock and McCauley, 1999). Therefore, if the Nm value is high then there is no restriction on gene flow and migrants can freely enter the population

Table 4-2: Structure results showing estimated ln probability of data and the variance of ln likelihood for K=1 to K=20 for the *rpoS* gene

Estimated number of populations (K)	ln P(D)	Var[lnP(D)]
1	-2423.2	52.2
2	-1534.6	110
3	-1481.9	223.6
4	-1357.7	163
5	-1373.6	260.6
6	-1150.3	330.3
7	-1069.1	244.8
8	-1068.7	253.6
9	-1078.1	268.9
10	-1079.6	283.6
11	-1087.5	296.9
12	-1122.9	359.4
13	-1185.9	455.3
14	-2013.2	2209.9
15	-1114.9	478.1
16	-1111.3	452.1
17	-1069.3	422
18	-1144.1	496.5
19	-1129.8	512.1
20	-1241.1	709.1

Table 4-3: Structure results showing estimated ln probability of data and the variance of ln likelihood for K=1 to K=20 for the *rpoS* gene

Estimated number of populations (K)	ln P(D)	Var[lnP(D)]
1	-4849.8	65.4
2	-3852.6	145.8
3	-2984	340.6
4	-2306.7	242.2
5	-2208.1	576.6
6	-2041.4	291.8
7	-1912.5	250.1
8	-1944	374.8
9	-1712.7	382
10	-1696.2	446.2
11	-1654.9	370.1
12	-1926.7	921.5
13	-2245.7	1649.4
14	-1739.2	572.3
15	-2444.3	2040.8
16	-2075.8	1258
17	-2216.5	1558.3
18	-1727.9	709.8
19	-2249	1719.7
20	-2026.2	1471.9

Furthermore, population subdivision can be determined using the sequenced-based statistic K_{ST}* which gives a good indication of genetic differentiation. Taking into account the number of nucleotide differences within each sequence type, the statistic K_{ST}* determines the likelihood of different sub-populations. Under the null hypothesis, namely that two sub-populations are not genetically different, K_{ST}* should be close to or equal to zero. Therefore, a high K_{ST}* value with a low probability (P)-value less than 0.05 will lead to the null hypothesis being rejected (Hudson et al., 1992b).

The gene flow and genetic differentiation results showed similar patterns for both the *rpoS* and *uidA* genes, results are shown below in Table 4.4 and Table 4.5 respectively. Tests for genetic differentiation, for both genes, suggest that the null hypothesis be rejected, that is, these two populations show some population subdivision between the Roodeplaat and Rietvlei Dam. However, there is no restriction of gene flow between the isolates from the Roodeplaat and Rietvlei Dam for both genes (*rpoS* F_{st} = 0.0229; Nm = 21.37 and *uidA* F_{st} = 0.0548; Nm = 8.63). Similar results were observed for water, algae and sewage populations when compared to the rest of the population. The null hypothesis was rejected and sub-populations showed some genetic differentiation but there was little or no restriction of gene flow.

When comparing sediment isolates with the remaining isolates from both the Roodeplaat and Rietvlei Dams, the null hypothesis was accepted. Here, the sub-populations are not genetically differentiated. High levels of gene flow are also observed, for both genes, between sediment isolates and the remaining isolates from both dams (e.g. *rpoS* F_{st} = 0.00873; Nm = 56.79 and *uidA* F_{st} = 0.00739; Nm = 67.14, respectively). These results were expected as sediment and water environments are in constant contact with each other resulting in no physical restriction on gene flow. Comparisons between all water isolates and all sediment isolate gave the same results.

Conversely, genetic differentiation between the Roodeplaat water hyacinth isolates and the Rietvlei Dam isolates from decaying plant material strongly rejected the null hypothesis for both genes. When compared to the remaining isolates from both dams, these two groups showed significant population subdivision. In addition, gene flow was highly restricted between these two groups and between the remaining populations. There is little gene flow observed between water hyacinth isolates (Cluster 3) and the remaining isolates from both dams, (*rpoS* F_{st} = 0.660; Nm = 0.26 and *uidA* F_{st} = 0.711; Nm = 0.20). Similar results are observed between Rietvlei Dam decaying plant isolates (Cluster 3) and the remaining isolates from both dams, (*rpoS* F_{st} = 0.479; Nm = 0.54 and *uidA* F_{st} = 0.175; Nm = 2.35). In addition, there is also little or no gene flow between the Roodeplaat Dam water hyacinth isolates and the Rietvlei Dam decaying plant isolates (*rpoS* F_{st} = 0.714; Nm = 0.20 and *uidA* F_{st} = 0.826; Nm = 0.17) indicating that these clusters may be in the process of becoming isolated groups.

Table 4-4: Gene flow and genetic differentiation estimates based on *rpoS* sequence data of isolates from the Roodeplaat and Rietvlei Dam

Populations compared [a]	Gene flow [b]		Genetic differentiation [c]
	F_{ST}	*Nm*	K_{ST}^{*}
All Roodeplaat Dam isolates vs. All Rietvlei Dam isolates	0.0229	21.37	0.00787 (0.007**)
Roodeplaat Dam water hyacinth vs. Remaining Roodeplaat and Rietvlei Dam isolates	0.660	0.26	0.0752 (0.000***)
Rietvlei Dam decaying plant material vs. Remaining Roodeplaat and Rietvlei Dam isolates	0.479	0.54	0.0592 (0.000***)
All sediment isolates vs. Remaining Roodeplaat and Rietvlei Dam isolates	0.00873	56.79	0.00139 (0.186^{ns})
All algae isolates vs. Remaining Roodeplaat and Rietvlei Dam isolates	0.0130	38.11	0.00259 (0.091^{ns})
All sewage isolates vs. Remaining Roodeplaat and Rietvlei Dam isolates	0.0249	19.59	0.00702 (0.012*)
All water isolates vs. Remaining Roodeplaat and Rietvlei Dam isolates	0.0583	8.07	0.0133 (0.001**)
Rietvlei Dam decaying plant isolates vs. Roodeplaat Dam Water Hyacinth isolates	0.714	0.20	0.562 (0.000**)

[a] Populations were defined based on their geographical location and sample site.

[b] Gene flow estimated as described by Hudson et al. (1992a)

[c] Genetic differentiation estimated as described by Hudson et al. (1992b and 2000)

ns = not significant

*0.01 < P > 0.05

**0.001 < P > 0.01

***P < 0.001

Table 4-5: Gene flow and genetic differentiation estimates based on *uidA* sequence data of isolates from the Roodeplaat and Rietvlei Dam

Populations compared [a]	Gene flow [b]		Genetic differentiation [c]
	F_{ST}	*Nm*	K_{ST}^{*}
All Roodeplaat Dam isolates vs. All Rietvlei Dam isolates	0.0548	8.63	0.0172 (0.000***)
Roodeplaat Dam water hyacinth vs. Remaining Roodeplaat and Rietvlei Dam isolates	0.711	0.20	0.0642 (0.000***)
Rietvlei Dam decaying plant vs. Remaining Roodeplaat and Rietvlei Dam isolates	0.175	2.35	0.0131 (0.000***)
All sediment isolates vs. Remaining Roodeplaat and Rietvlei Dam isolates	0.00739	67.14	0.00126 (0.146^{ns})
All algae isolates vs. Remaining Roodeplaat and Rietvlei Dam isolates	0.0764	6.05	0.0137 (0.000***)
All sewage isolates vs. Remaining Roodeplaat and Rietvlei Dam isolates	0.0203	24.08	0.00532 (0.004**)
All water isolates vs. Remaining Roodeplaat and Rietvlei Dam isolates	0.0226	21.60	0.00837 (0.001**)
Rietvlei Dam decaying plant isolates vs. Roodeplaat Dam Water Hyacinth isolates	0.826	0.17	0.513 (0.000***)

[a] Populations were defined based on their geographical location and sample site.
[b] Gene flow estimated as described by Hudson et al. (1992a)
[c] Genetic differentiation estimated as described by Hudson et al. (1992b and 2000)
ns = not significant
*0.01 < P > 0.05
**0.001 < P > 0.01
***P < 0.001

4.4 DISCUSSION

The initial indication (Chapter 2), namely that unique environmental *E. coli* clusters exist, prompted further investigations into the structure and ecology of the entire *E. coli* population including isolates from two geographically separated aquatic systems e.g. Roodeplaat and Rietvlei Dams. Population structure results indicated all isolates belonged to only one population (K=1). This result supported the notion that if sewage, and therefore humans, is the main source of *E. coli* into both dams then the *E. coli* population should be homogenous. Gene flow is thus expected to occur between individual bacteria, in order for them to form and maintain one population. To test this assumption it was important to investigate gene flow and genetic differentiation between individuals.

The analyses indicated that gene flow varied throughout the *E. coli* population. For the majority of populations, few or no restrictions on gene flow were observed providing further support for the Structure analysis, which indicated that it was most likely that all the isolates belonged to one population. The high level of gene flow between the majority of populations was expected as they were all isolated from an aquatic environment where the sediment and algal isolates are in constant contact with the water and populations are continuously mixed. In addition, treated effluent is released into both dams and therefore sewage isolates become mixed into the greater population. Although, gene flow was not restricted between many of the populations, there was still some level of genetic differentiation and population subdivision. The null hypothesis was rejected in those cases.

However, in terms of genetic differentiation, a number of sub-populations identified rejected the null hypothesis that sub-populations are not genetically distinct. Two of the clusters constantly observed for all four gene phylogenies, (Roodeplaat Dam water hyacinth (Q02H) and the Rietvlei decaying plant (DWWF) populations were found to be genetically distinct, with little or no genetic exchange between them and the rest of the population. These two groups of isolates also showed significant genetic differentiation at the nucleotide level and cluster together consistently with good bootstrap support (Chapter 2).

During the gene flow and genetic differentiation analyses, populations were also defined based on the geographical location and sample site from where the isolates were collected. Results for these populations appear to indicate that the exact location of the sampling site may have some effect on gene flow, especially in situations where populations have undergone some level of niche separation. However, defining the ecological niche of free-living bacteria is often difficult. Gray et al. (1999) discovered that the freshwater, sediment-dwelling bacterium, *Achromatium oxaliferum*, experienced adaptive radiation whereby the species has diversified into numerous forms that are capable of occupying different niches. Sub-populations of *A. oxaliferum* showed ecological differentiation within the same sediment environment by adapting to different redox conditions. Furthermore, Schauer et al. (2005) revealed that there was ecological niche separation at a sub-cluster level within the monophyletic SOL cluster of freshwater bacterioplankton. Adaptation to various water chemistries together with other abiotic and biotic conditions resulted in these bacteria adapting to different ecological niches within the freshwater environment.

Separate sub-populations were expected as *E. coli* isolates were not only collected from two different dams but also from different sources (sewage, water, sediment, algae and water plants). Multiple studies have showed that *E. coli* has the ability to survive and persist in soil and freshwater environments (Ishii et al., 2010, Walk et al., 2007, Power et al., 2005, Byappanahalli et al., 2006, Solo-Gabriele et al., 2000). Furthermore, there are is an increasing number of studies that indicate that environmental *E. coli* differ from their human and commensal counterparts on a genetic level (Luo et al., 2011, Oh et al., 2012), all supporting the possibility of environmental isolates within this population.

4.5 CONCLUSIONS

A number of environmental *E. coli* strains formed separate groups, based on their gene phylogenies, which suggested that they have become genetically separate from typical human isolates. This notion is also supported by the gene flow and genetic differentiation analyses, which confirmed that these plant-associated clusters showed some level of separation from the rest of the population. The effect of the environment and their unique niche clearly played an important role in the evolution of these *E. coli* strains. It is also possible that these strains may also be ecologically different from their gut-associated counterparts.

CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS

5.1 CONCLUSIONS

The results from this study indicated a highly diverse *E. coli* population present in aquatic environments. This was contrary to the normal assumption that due to human contamination being the primary source of *E. coli* in aquatic environments, a more homogenous *E. coli* population would be expected. The existence and diversity of environmental *E. coli* populations in other important aquatic environments such as aquifers was also not addressed during this study and should still be investigated.

The *rpoS* phylogeny, confirmed that all isolates belonged to *E. coli sensu stricto*. None of the isolates grouped with the five *E. coli sensu lato* clades previously described by Walk et al. (2009). It is unclear why strains belonging to these clades were not detected in any of the aquatic samples analysed during this study.

The presence of a number of possible environmental clusters was observed with strong support in all phylogenies for two of these clusters (Roodeplaat Dam water hyacinth and Rietvlei Dam decaying plant). Isolates from these clusters showed limited gene flow between them and the rest of the population. In addition, these two plant-associated populations showed higher levels of genetic differentiation than any of the other populations. This provided a strong indication that these two groups of isolates were, at some level, genetically distinct and have undergone significant population sub-division. These populations may have developed a close association with water plants and are therefore better adapted for growth and survival in the external environment outside of the primary host. The potential link between these isolates and other components of the aquatic ecosystem, such as aquatic invertebrates and water birds, should receive further attention.

Oh et al. (2012) showed that the environmental clades described by Walk et al. (2009) have undergone reductive evolution by the loss of genes associated with attachment, defence and stress response mechanisms. Their study was based on genomic comparisons between those environmental clades and strains of *E. coli sensu stricto,* isolated from humans. It would be interesting, for future work, to investigate if similar gene loss is observed within the environmental isolates found in this study, taking special note that the environmental isolates from this study all grouped within the “true” *E. coli*.

Many of the *E. coli* isolates obtained from the aquatic environment grouped with sewage isolates and would likely have the ability to circulate within the human population. This indicated that in most cases the presence of *E. coli* could still be used to evaluate the safety of water for human use. The current findings, however, still raise a number of specific questions related to the use of *E. coli* as a faecal indicator. Because these *E. coli* strains survive and proliferate in the secondary environment, it is important to know whether or not they can survive in the human gut and can pose a threat to water users by harbouring genes associated with the various types of pathogenic *E. coli*. On the other hand, if they are not circulating through the human population, it would be important to know how often and at what levels they are detected when the *E. coli* is used to evaluate the safety of water for human use.

5.2 RECOMMENDATIONS

Genome research and sequencing are currently the leading drivers in biological research. The genome sequence of an organism not only serves as the blue print of all the genes present but also provides the opportunity to expand our knowledge and understanding of the organism’s biology. Sequencing the

genomes of some of these environmental isolates will provide an ideal opportunity to address some of the questions raised above. Comparative genomics will be ideal to determine whether these “true” *E. coli* are also undergoing reductive evolution and have lost their ability to grow and cause disease in the human gut.

A genomics approach could also assist in developing more specific detection methods for *E. coli* associated with humans and warm-blooded animals. Based on the genome data, the unique metabolic capabilities of environmental strains could be detected and then used to differentiate them from the *E. coli* associated with human gastrointestinal tract.

REFERENCES

Anderson, K. L., Whitlock, J. T and Harwood, V. J. (2005). Persistence and differential survival of faecal indicator bacteria in subtropical waters and sediments. *Applied and Environmental Microbiology.* **71**(6): 3041-3048.

Anderson, M. A., Whitlock, J. T and Harwood, V. J. (2006). Diversity and distribution of *Escherichia coli* genotypes and antibiotic resistance phenotypes in faeces of humans, cattle, and horses. *Applied and Environmental Microbiology.* **72**(11): 6914-6922.

Baur, B., Hanselmann, K., Schlimme, W and Jenni, B. (1996). Genetic transformation in freshwater: *Escherichia coli* is able to develop natural competence. *Applied and Environmental Microbiology.* **62**(10): 3673-3678.

Bennett, A. F., Lenski, R. E and Mittler, J. E. (1992). Evolutionary adaptation to temperature. I. Fitness responses of *Escherichia coli* to changes in its thermal environment. *Evolution.* **46**(1): 16-30.

Bergholz, P. W., Noar, J. D and Buckley, D. H. (2011). Environmental patterns are imposed on the population structure of *Escherichia coli* after fecal deposition. *Applied and Environmental Microbiology.* **77**(1): 211-219.

Bergthorsson, U and Ochman, H. (1995). Heterogeneity of genome size among natural isolates of *Escherichia coli. Journal of Bacteriology.* **177**(20): 5784-5789.

Bodenstein, J.A., van Eeden P.H., Legadima, J. and Chaka, J. (2006). A preliminary assessment of the present ecological state of the major rivers and streams within the northern service delivery region of the Ekurhuleni metropolitan municipality. WISA 2006 conference paper.

Böhm, H and Karch, H. (1992). DNA fringerprinting of *Escherichia coli* O157:H7 strains by pulsed-field gel electrophoresis. *Journal of Clinical Microbiology.* **30**(8): 2169-2172.

Brady, C., Venter, S., Cleenwerck, I., Vancanneyt, M., Swings, J and Coutino, T. (2007). A FAFLP system for the improved identification of plant-pathogenic and plant-associated species of the genus *Pantoea*. *Systematic and Applied Microbiology.* **30**: 413-417.

Brennan, F. P., Abram, F., Chinalia, F. A., Richards, K. G and O'Flaherty, V. (2010). Characterisation of environmentally persistent *Escherichia coli* isolates leached from an Irish soil. *Applied and Environmental Microbiology*. **76**(7): 2175-2180.

Byappanahalli, M. and Fujioka, R. (2004). Indigenous soil bacteria and low moisture may limit but allow faecal bacteria to multiply and become a minor population in tropical soils. *Water Science and Technology.* **50** (1): 27–32.

Byappanahalli, M. N., Fowler, M., Shively, D. A and Whitman, R. L. (2003a). Ubiquity and persistence of *Escherichia coli* in a Midwestern coastal stream. *Applied and Environmental Microbiology.* **69**(8): 4549-4555.

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

Byappanahalli, M. N., Whitman, R. L., Shively, D. A., Ferguson, J., Ishii, S and Sadowsky, M. J. (2007). Population structure of *Cladophora* –borne *Escherichia coli* in nearshore water of lake Michigan. *Water Research*. **41**: 3649-3654.

Byappanahalli, M. N., Whitman, R. L., Shively, D. A., Sadowsky, M. J and Ishii, S. (2006). Population structure, persistence and seasonality of autochthonous *Echerichia coli* in temperate, costal forest soil from a Great Lakes watershed. *Environmental Microbiology.* **8**(3): 504-513.

Clermont, O., Bonacorsi, S and Bingen, E. (2000). Rapid and simple determination of the *Escherichia coli* phylogenetic group. *Applied and Environmental Microbiology*. **66**(10): 4555-4558.

Cohan, F. M. (2002). What are bacterial species? *Annual Review in Microbiology.* **56**:457-487.

Cooper, V. S and Lenski, R. E. (2000). The population genetics of ecological specialization in evolving *Escherichia coli* populations. *Nature*. **407**: 736-739.

Davis, S. A and Gordon, D. M. (2002). The influence of host dynamics on the clonal composition of *Escherichia coli* populations. *Environmental Microbiology.* **4**(5): 306-313.

Dijkshoorn, L., Towner, K. J and Struelens, M. New approaches for the generation and analysis of microbial typing data. Elsevier (2001). P 1-24, 178-205 and 299-334.

Doolittle, R.F., Feng, D.F., Tsang, S., Cho, G and Little, E. (1996). Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science.* **271:** 470–477.

Durso, L. M., Smith, D and Hutkins, R. W. (2004). Measurement of fitness and competition in commensal *Escherichia coli* and *E. coli* O157:H7 strains. *Applied and Environmental Microbiology.* **70**(11): 6466-6472.

Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S. R., Nelson, K. E and Relman, D. A. (2005). Diversity of the human intestinal microbial flora. *Science.* **308**: 1635-1638.

Elliot, E. L and Colwell, R. R. (1985). Indicator organisms for the estuarine and marine water. *FEMS Microbiology letters*. **32**(2): 61-79.

Escobar-Páramo, P., Giudicelli, C., Parsot, C and Denamur, E. (2003). The evolutionary history of *Shigella* and enteroinvasive *Escherichia coli* revised. *Journal of Molecular Evolution.* **57**: 140-148.

Falush, D., Stephens, M and Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes.* **7**(4): 574-578.

Farmer, J. J., III. 1995. Enterobacteriaceae: introduction and identification, p. 438–449. In P. R. Murray, E. J. Baron, M. A. Pfaller, F. C. Tenover, and R. H. Yolken (ed.), Manual of clinical microbiology, 6th ed. ASM Press, Washington, D.C.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**:368-376.

Gordon, D. M. (2004). The influence of ecological factors on the distribution and the genetic structure of *Escherichia coli.* In *Escherichia coli* and *Salmonella*: *Cellular and Molecular Biology.* Module 6.4.1. *American Society of Microbiology.* [Online] http://www.ecosal.org

Gordon, D. M. (2001). Geographical structure and host specificity in bacteria and the implications for tracing the source of coliform contamination. *Microbiology.* **147**: 1079-1085.

Gordon, D. M., Bauer, S and Johnson, J. R. (2002). The genetic structure of *Escherichia coli* populations in primary and secondary habitats. *Microbiology.* **148**: 1513-1522.

Gordon, D. M., Clermont, O., Tolley, H and Denamur, E. (2008). Assigning *Escherichia coli* strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method. *Environmental Microbiology.* **10**(10): 2484-2496.

Gray, N. D., Howarth, R., Rowan, A., Pickup, R. W., Gwyn Jones, J and Head, I. M. (1999). Natural communities of *Achromatium oxaliferum* comprise genetically, morphologically, and ecologically distinct populations. *Applied and Environmental Microbiology*. **65**(11): 5089-5099.

Grimont, F and Grimont, P. A. D. (1986). Ribosomal nuclaic acig ggene restriction patterns as potential taxonomic tools. *Annales de I'Institut Pasteur.* **137**(1): 165-175.

Guan, S., Xu, R., Chen, S., Odumeru, J and Gyles, C. (2002). Developement of a procedure for discriminating among *Escherichia coli* isolates from animal and human sources. *Applied and Environmental Microbiology.* **68**(6): 2690-2698.

Guindon S., Dufayard J.F., Lefort V., Anisimova M., Hordijk W., Gascuel O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*. **59**(3): 307-321.

Hahm, B., Maldonado, Y., Schreiber, E., Bhunai, A. K and Nakatsu, C. (2002). Subtyping foodborne and environmental isolates of *Escherichia coli* by multiplex-PCR, rep-PCR, ribotyping and AFLP. *Journal of Microbiological Methods*. **53**: 387-399.

Hahn, M. W. (2006). Microbial diversity of inland waters. *Current Opinion in Biotechnology.* **17**: 256-261.
Hall, T.A. (1999). BioEdit: A user- friendly biological sequence alignment editor and analysis program. *Nucleic Acids Symposium Series*. **41**: 95-98.

Hartl, D. L and Dykhuizen, D. E. (1984). The population genetics of *Escherichia coli*. *Annual Reviews in Genetics.* **18**: 31-68.
Hudson, R. R. (2000). A new statistic for detecting genetic differentiation. *Genetics*. **155**: 2011-2014.

Hudson, R. R., Boos, D. D and Kaplan, N. L. (1992b). A statistical test for detecting geographical subdivision. *Molecular Biology and Evolution*. **9**(1): 138-151.

Hudson, R. R., Slatkin, M and Madison, W. P. (1992a). Estimation of levels of gene flow from DNA sequence data. *Genetics*. **132**: 583-589.

Hulton, C. S. J., Higgins, C. F and Sharp, P. M. (1991). ERIC sequences: a novel family of repetitive elements in the genomes of *Escherichia coli, Salmonella typhimurium* and other enterobacteria. *Molecular Microbiology.* **5**(4): 825-834.

Ihssen, J., Grasselli, E., Bassin, C., Francois, P., Piffaretti, J., Köster, W., Schrenzel, J and Egli, T. (2007). Comparative genomic hybridisation and physiological characterisation of environmental isolates indicate that significant (eco-)physiological properties are highly conserved in *Escherichia coli*. *Microbiology.* **153**: 2052-2066.

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

Ishii, S and Sadowsky, M. J. (2009). Applications of rep-PCR fingerprinting technique to study microbial diversity, ecology and evolution. *Environmental Microbiology.* **11**(4): 733-740.

Ishii, S., Hansen, D. L., Hicks, R. E and Sadowsky, M. J. (2007). Beach sand and sediments are temporal sinks and sources of *Escherichia coli* in Lake Superior. *Environ. Sci. Technol.* **41**: 2203-2209.

Ishii, S., Ksoll, W. B., Hicks, R. E and Sadowsky, M. J. (2006). Presence and growth of naturalised *Escherichia coli* in temperate soils from Lake Superior watersheds. *Applied and Environmental Microbiology.* **72**(1): 612-621.

Ishii, S., Yan, H., Hansen, D. L., Hicks, R. E and Sadowsky, M. J. (2010). Factors controlling long-term survival and growth of naturalised *Escherichia coli* populations in temperate field soils. *Microbes Environment.* **25**(1): 8-14.

Jonas, D., Spitzmuller, B., Weist, K., Ruden, H and Daschner, F. D. (2003). Comparison of PCR-based methods for typing *Escherichia coli. Clinical Microbiology and Infection.* **9**(8): 823-831.

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

Lavigne, J and Blanc-Potard, A. (2008). Molecular evolution of *Salmonella enterica* serovar Typhimurium and pathogenic *Escherichia coli*: From pathogenesis to therapeutics. *Infection, Genetics and Evolution.* **8**: 217-226.

Leung, K. T., Mackereth, R., Tien, Y., Topp, E. (2004). A comparison of AFLP and ERIC-PCR analyses for discriminating *Escherichia coli* from cattle, pig and human sources. *FEMS Microbiology Ecology.* **47**: 111-119.

Luchi, S and Lin, E. C. C. (1993). Adaptation of *Escherichia coli* to redox environments by gene expression. *Molecular Microbiology.* **9**(1): 9-15.

Lukjancenko, O and Wassenaar, T. M. (2010). Comparison of 61 sequenced *Escherichia coli* genomes. *Microbial Ecology*. **60**: 708-720.

Luo, C., Walk, S. T., Gordon, D. M., Feldgarden, M., Tiedje, J. M and Konstantinidis, K. T. (2011). Genome sequencing of environmental *Escherichia coli* expands understanding of ecology and speciation of the model bacterial species. *PNAS*. **108**(17): 7200-7205.

Maiden, M. C. J., Bygraves, J. A., Feil, E., Morelli, G., Russel, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., Feavers, I. M., Achman, M and Spratt, B. G. (1998). Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *PNAS.* **95**: 3140-3145.

Malan, T. P and McClure, W. R. (1984). Dual promoter control of the *Lac* operon. *Cell.* **39**(1):173-180.
McLellan, S, L., Daniels, A. D and Salmore, A. K. (2003). Genetic characterisation of *Escherichia coli* populations from host sources of fecal pollution by using DNA fingerprinting. *Applied and Environmental Microbiology.* **69**(5): 2587-2594.

McNicholas, P., Salavati, R and Oliver, D. (1997). Dual regulation of *Escherichia coli secA* translation by distinct upstream elements. *Journal of Molecular Biology.* **265**: 128-141.

Nakamura, Y., Itoh, T., Matsuda, H and Gojobori, T. (2004). Biased biological functions of horizontally transferred genes in prokaryotes. *Nature Genetics.* **36**: 760-766.

Ochman, H and Selander, R. K. (1984). Standard reference strains of *Escherichia coli* from natural populations. *Journal of Bacteriology.* **157**(2): 690-693.

Oh, S., Buddenborg, S., Yoder-Himes, D. R., Tiedje, J and Konstantinidis, K. T. (2012). Genomic diversity of *Escherichia* isolates from diverse habitats. *PLOS one.* **7**(10): e47005.

Olive, D. M and Bean, P. (1999). Principles and applications of methods for DNA-based typing of microbial organisms. *Journal of Clinical Microbiology.* **37**(6): 1661-1669.

Paton, A. W and Paton, J. C. (1998). Detection and characterization of Shiga toxigenic *Escherichia coli* by using multiplex PCR assays for *stx1*, *stx2*, *eaeA*, enterohemorrhagic *E. coli hlyA*, rfb_{O111}, and rfb_{O157}. *Journal of Clinical Microbiology*. **36**: 598-602.

Posada D. (2008). jModelTest: Phylogenetic Model Averaging. *Molecular Biology and Evolution* **25**: 1253-1256.

Power, M. L., Littlefield-Wyer, J., Gordon, D. M., Veal, D. A and Slade, M. D. (2005). Phenotypic and genotypic characterisation of encapsulated *Escherichia coli* isolated from blooms in two Australian lakes. *Environ. Microbiology*. **7**(5): 631-640.

Prichard, J. K., Stephens, M and Donnelly, P. (2000). Inference of population structure using Multilocus genotype data. *Genetics.* **155**: 945-949.

Prosser, J. I., Bohannan, B. J. M., Curtis, T. P., Ellis, R. J., Firestone, M. K., Freckleton, R. P., Green, J. L., Green, L. E., Killam, K., Lennon, J. J., Osborn, A. M., Solan, M., van der Gast, C. J and Young, P. W. (2007). The role of ecological theory in microbial ecology. *Nature Perspectives.* **5**: 384-392.

Pupo, G. M., Lan, R and Reeves, P. R. (2006). Multiple independent origins of Shigella clones of *Escherichia coli* and convergent evolution of many of their characteristics. *PNAS.* **97**(19): 10567-10572.

Rademaker, J. L. W., Louws, F. J and de Bruijn, F. J. (1998). Characterization of the diversity of ecologically important microbes by rep- PCR genomic fingerprinting. In: Akkermans, A.D.L., van Elsas, J.D., de Bruijn, F.J. (Eds.). Molecular Microbial Ecology Manual. Kluwer Academic Publishers. Dordrecht. p. 3.4.3:1–3.4.3:27.

Rasko, D.A., Rosovitz,M. J., Myers, G. S. A., Mongodin, E. F., Fricke, W. F., Gajer, P., Crabtree, J., Sebaihia, M., Thomson, N. R., Chaudhuri, R., Henderson, I. R., Sperandio, V and Ravel, J. (2008). The pangenome structure of *Escherichia coli*: Comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *Journal of Bacteriology.* **190**(20): 6881-6893.

Ribot, E. M., Fair, M. A., Gautom, R., Cameron, D. N., Hunter, S. B., Swaminathan, B and Barret, T. J. (2006). Standardisation of pulsed-field gel electrophoresis protocols for the subtyping of *Eshcerichia coli* O157:H7, *Salmonella* and *Shigella* for PulseNet. *Foodborne Pathogens and Disease.* **3**(1): 59-67.

Sampson, R. W., Swiatnicki, S. A., Osinga, V. L., Supita, J. L., McDermott, C. M and Kleinheinz, G. T. (2006). Effects of temperature and sand on *E. coli* survival in northern lake water microcosm. *Journal of Water and Health.* **4**(3): 389-393.

Savageau, M. A. (1983). *Eschericia coli* habitats, cell types and molecular mechanisms of gene control. *The American Naturalist*. **122**(6): 732-744.

Schauer, M., Kamenik, C and Hahn, M. W. (2005). Ecological differentiation within a cosmopolitan group of planktonic freshwater bacteria (SOL cluster, *Saprospiraceae, Bacteroidetes*). *Applied and Environmental Microbiology*. **71**(10): 5900-5907.

Schwartz, D. C and Cantor, C. R. (1984). Separation of yeast chromosome-sized DNAs by pulsed field gel electrophoresis. *Cell*. **37**(1): 67-75.

Solo-Gabriele, H. M., Wolfert, M. A., Desmarais, T. R and Palmer, C. J. (2000) Sources of *Escherichia coli* in a costal subtropical environment. *Applied and Environmental Microbiology.* **7**(1): 230-237.

Spratt, B. G and Maiden, M. C. J. (1999). Bacterial population genetics, evolution and epidemiology. *Philos. Trans. R. Soc. Lond. B. Biol.* **345**: 701-710.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M and Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution.* **28**(10): 2731-2739.

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

Tenover, F. C., Arbeit, R. D., Goering, R. V., Mickelsel, P. A., Murry, B. E., Persing, D. H and Swaminathan, B. (1995). Interpreting chromosomal DNA restriction patterns produces by pulsed-field gel electrophoresis: criteria for bacterial strain typing. *Journal of Clinical Microbiology.* **33**(9): 2233-2239.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*. **22**: 4673–4680.

Trevors, J. T. (1998). Review: Bacterial population genetics. *World Journal of Microbiology and Biotechnology.* **14**: 1-5.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van Dalee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kulper, M and Zabeau, M. (1995). AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res.* **23**: 4407-4414.

Walk, S. T., Alm, E. W., Calhoun, L. M., Mladonicky, J. M and Whittman, T. S. (2007). Genetic diversity and population structure of *Escherichia coli* isolated from freshwater beaches. *Environ. Microbiology.* **9**(9): 2274-2288.

Walk, S. T., Alm, E. W., Gordon, D. M., Ram, J. L., Toranzos, G. A., Tiedjie, J. M and Whittam, T. S. (2009). Cryptic lineages of the genus *Escherichia. Applied and Environmental Microbiology.* **75**(20): 6534-6544.
Welch, R. A. (2006). The genus *Escherichia*. *Prokaryotes*. **6**: 60-71.

Wheeler Alm, E., Burke, J and Spain, A. (2003). Faecal indicator bacteria are abundant in wet sand at freshwater beaches. *Water Research*. **37**: 3978-3982.

Whitlock, M. C and McCauley D. E. (1999). Indirect measures of gene flow and migration: $F_{ST} \neq 1/(4Nm+1)$. *Heredity*. **82**:117-125.

Whitman, R. L., Nevers, M, B and Byappanahalli, M. N. (2006). Examination of the watershed-wide distribution of *Escherichia coli* along Southern Lake Michigan: an integrated approach. *Applied and Environmental Microbiology.* **72**(11): 7301-7310.

Whittam, T. S (1996). Genetic variation and evolutionary processes in natural populations of *Escherichia coli*. In: F. C. Neidhardt (ed.), *Escherichia coli* and *Salmonella*: cellular and molecular biology. *American Society for Microbiology,* Washington, D.C. p. 2708–2720.

Whittam, T. S. (1989) Clonal dynamics of *Escherichia coli* in its natural habitat. *Antonie Leeuwenhoek*. **55**: 23-32.

Winfield, M. D and Groisman, E. A. (2003). Role of nonhost environments in the lifestyle of *Salmonella* and *Escherichia coli. Applied and Environmental Microbiology.* **69**(7): 3687-3694.

Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., Wieler, L.H., Karch, H., Reeves, P.R., Maiden, M.C.J., Ochman, H. and Achtman, M. (2006). Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Molecular Microbiology.* **60**: 1136-1151.

Yang, H. H., Vinopal, R. T., Grasso, D and Smets, B. F. (2004). High diversity among environmental *Escherichia coli* isolates from a bovine feedlot. *Applied and Environmental Microbiology.* **70**(3): 1528-1536.

Yang, H. R., Wu, F. T., Tsai, J. L., Mu, J. J., Lin, L. F., Chen, K. L., Kuo, S. H., Chiang, C. S and Wu, H. S. (2007). Comparison between O serotyping method and multiplex real-time PCR to identify diarrheagenic *Escherichia coli* in Taiwan. *Journal of Clinical Microbiology.* **45**(11): 3620-2625.