The diversity and structure of *Escherichia coli* populations in fresh water environments

by

Sarah Catherine MacRae

Submitted in partial fulfilment of the

Requirements for the degree

Magister Scientiae

in the

Faculty of Natural and Agricultural Sciences

University of Pretoria

Pretoria

Supervisor: Prof S N Venter

Co Supervisor: Prof E T Steenkamp

Co Supervisor: Prof V S Brözel

DECLARATION

I declare that the dissertation, which I hereby submit for the degree **Magister Scientiae** at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at another university.

_____ _____

Sarah Catherine MacRae Date

The diversity and structure of *Escherichia coli* populations in fresh water environments

by

Sarah Catherine MacRae

Supervisor: Prof S N Venter

Co Supervisor: Prof E Steenkamp

Co Supervisor: Prof V S Brözel

Department: Microbiology and Plant Pathology

Degree: MSc (Microbiology)

SUMMARY

Escherichia coli is a well known commensal inhabitant of the gastrointestinal tract of both humans and animals and a highly diverse species. 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. It has been described that *E. coli* typically exists in two environments, the primary environment being the gastrointestinal tract of the host and the secondary environment being that environment outside of the host (water, soil and sediments). Upon introduction into the environment outside of the host, the numbers of *E. coli* steadily decline. Generally, where *E. coli* is present in the external environment and where its numbers are maintained it is due to a constant direct faecal input from the host. This short lifespan in the environment outside of the host forms the basis for the use of *E. coli* as an indicator organism for faecal contamination in water systems.

In contrast, multiple studies have shown that some *E. coli* strains have the ability to survive and persist in the external environment in the absence of faecal input from the host. With a large pan-genome and the possibility of horizontal gene transfer (HGT) of desirable traits, *E. coli* have the potential to adapt to a variety of different niches overcoming drastic changes in conditions in its new environment. In addition, adaptation to the secondary environment is facilitated by the presence of soils and sediments, where in an aquatic environment they provide a source of nutrients and protection from the drastic change in conditions. Here, *E. coli* has the ability to occupy a new niche and become naturalised within an aquatic environment.

The aim of this masters project was to examine and characterise the diversity of *E. coli* isolates collected from two South African freshwater environments namely, the Roodeplaat and Rietvlei Dams, Pretoria. Specific research questions addressed in this study include: (1) are their unique and genetically differentiated sub-populations within the aquatic environments sample? (2) Is there a link between the unique sub-populations and their sample site? (3) Finally, what is the relationship between sub-populations in terms of gene flow and population structure? Understanding *E. coli's* population structure and ecology may shed some light on its evolution and potential to adapt to new environments.

Following phylogrouping, AFLP and phylogenetic analysis of the *rpoS* and *uidA* genes, the results indicated that the population was highly diverse with the majority of strains grouping together with the sewage isolates. Furthermore, population structure analyses concentrating on gene flow and genetic differentiation revealed that possible environmental groups exist within the population. In particular, two groups of *E. coli* isolates associated with aquatic plants showed restricted gene flow and definite genetic differentiation. These two groups can also be observed in the *rpoS* and *uidA* phylogenetic analyses where they consistently group together in the absence of sewage isolates.

These findings demonstrate that some *E. coli* are not only able to survive outside of their host but have undergone some level of niche separation within the secondary environment. These results raise important questions into the accuracy of using *E. coli* as an indicator organism. In the long term, this study may aid in understanding the population dynamics of *E. coli* and the implications of environmental strains on using *E. coli* in assessing water quality.

ACKNOWLEDGEMENTS

- I would like to thank the Lord for blessing me with the opportunity and ability to further my studies, always listening to my prayers and giving me the faith that everything will be okay.

- Thank you to my wonderful parents for your unfaltering love and support, both emotionally and financially. To my mom Patricia, thank you for your guidance, friendship, calming words and always being there for me. To my dad Donald, my biggest fan and supporter, thank you for everything you did for me and always believing in me, without you none of this would have been possible. I am so sorry you are no longer here with us to see the finished product. I will never forget your words of encouragement and support and will always hold them close to my heart.

- To my brother Gregory and sister Claire, thank you for your love and support through the good times and bad. Thank you for your words of encouragement, all the laughs and believing in me. I feel incredibly blessed to not only call you my brother and sister but also my best friends.

- To my loving boyfriend Shane, thank you for being my shoulder to cry on and pillar of strength, for always making me laugh and putting a smile on my face. Thank you for allowing me to follow my dreams and being patient with me while I discover what they are.

- A huge thank you to my Supervisors. To Prof S N Venter, thank you for this opportunity and believing in me even when I did not believe in myself. To Prof E Steenkamp and Prof V Brözel thank you for your guidance and sharing your knowledge with me.

- To my lab mates and friends, David, Helen, Mathilde, Gaby, Ockert, Zander, Annie, Gina, Tarren, Adele, thank you for all the laughs and good times. Thank you for all your help, ideas and inspiring conversations. Special thanks to Tarren for your work on the Rietvlei Dam isolates.

- To the NRF and University of Pretoria, thank you for the financial support and giving me the opportunity to further my studies.

- I would like to dedicate this thesis to my wonderful father who passed away before I could complete this write up. I hope this will make you proud Pops, I miss you every day.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 2

CHAPTER 3

LIST OF TABLES

CHAPTER 2

CHAPTER 3

PREFACE

Most *Escherichia coli* strains naturally exist as either a commensal or a potential pathogen in the gastrointestinal tract of both humans and animals. For this reason, *E. coli* is widely used as an indicator for faecal contamination in water sources and general water quality. Its presence in water systems suggests the presence of other faecal pathogens and the possible risk to human health. Furthermore, using *E. coli* as an indicator organism is based on the assumptions that they are solely associated with the gut and are unable to survive outside of the host for a long period. Although, the majority of *E. coli* exist in the gut of the host (primary environment) and conform to the above-mentioned assumptions, several studies have shown that some strains have adapted and are capable of surviving and proliferating in the environment outside of the host (secondary environment). There is recent evidence of *E. coli* populations associated with several aquatic niches where they could be isolated from algae, sand and sediments on a regular basis.

Not only have these environmental strains adapted to the secondary environment but several studies have revealed that they are genetically distinct from their gut-associated counterparts. This possible niche separation may be due to the species having a large pan-genome, which allows *E. coli* to acquire new genes and diversify when it encounters a new environment. With the presence of possible unique naturalised *E. coli* in the secondary environment in the absence of faecal contamination, the continued use of *E. coli* as an indicator for water quality is questioned. If these strains have become adapted to the secondary environment, can they indicate recent faecal contamination?

This study focuses on the relationship between *E. coli* strains isolated from within the aquatic environments of the Roodeplaat and Rietvlei Dams situated in an urban environment (Pretoria, South Africa). In order to determine the relationship between strains the following questions will be addressed: (1) Are there unique environmental *E. coli* populations in the aquatic environments samples? (2) What is the diversity among the isolates? (3) What is the population structure among isolates? (4) Are isolates genetically differentiated from their commensal and pathogenic counterparts and (5) What level of population subdivision occurs within the population?

Firstly, sampling of the two aquatic environments in this study involved various sample types, representative of the ecology of each dam. Phylogrouping, Amplified Fragment Length Polymorphism (AFLP) and phylogenetic analysis of both the *rpoS* and *uidA* genes were used to determine the diversity within the *E. coli* population obtained. These methods were used to reveal if any possible unique environmental *E. coli* strains exist and to determine the relationship between isolates. Phylogenetic analyses were also used to verify whether any of these isolates belonged to the five novel *Escherichia* clades previously described.

Furthermore, to determine the relationship between *E. coli* strains, this study focused on population genetics analyses. Using computational analysis of the DNA sequence data of the two genes (*rpoS* and *uidA*), population structure, gene flow and population subdivision were determined. These analyses were again used to reveal the presence of unique environmental *E. coli* and suggest possible niche separation.

Based on multiple studies, it is apparent that *E. coli* has the ability to survive in the secondary environment. It is believed that this study may shed some light on the relationship between *E. coli* populations and aid in the more accurate use of *E. coli* as an indicator for recent faecal contamination. In the long term, the ability to distinguish between environmental and host-associated *E. coli* strains could allow for more precise methods to determine water quality and thereby improve risk assessments and corrective actions.

CHAPTER ONE

LITERATURE REVIEW

LITERATURE REVIEW

1.1 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 was 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* (Farmer, 1995). Within the Enterobacteriaceae *Escherichia* is most closely related to *Salmonella* and *Shigella*.

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 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 *Escherichia 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.*, 2000; Lan and Reeves, 2002). 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*., 2000). 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) (Hartl and Dykhuizen, 1984; Lan and Reeves, 2002; Escobar-Páramo *et al.*, 2003).

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 about 0.1% of the estimated 35 000 species of bacteria thought to make up the gut microbiota (Sekirov *et al.*, 2010). This is not surprising as *E. coli* is a facultative anaerobe existing in a predominately 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) in the more oxygenated mucosal lining of the gastrointestinal tract.

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 one time, an individual will be colonised by a predominant *E. coli* strain and over time that strain will most likely become the resident strain (Winfield and Groisman, 2003; Tenaillon *et al.*, 2010). This suggests that there is a strong relationship between the host and the 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 acquiring 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 I pili (Pratt and Kolter, 1998). 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). The majority of virulence factors are associated with plasmids and pathogenicity islands that non-pathogenic strains can acquire through horizontal gene transfer (HGT) (Touchon *et al.*, 2009; Moriel *et al.*, 2012).

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 diarrhea 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 diarrhea lasting up to two weeks or longer; enteropathogenic *E. coli* (EPEC) that causes of watery diarrhea in infants predominantly in developing countries; enterotoxigenic *E. coli* (ETEC) that is responsible for travellers' diarrhea; and lastly enteroinvasive *E. coli* (EIEC) that is genetically, biochemically and pathogenetically closely related to S*higella* and that causes invasive inflammatory colitis and dysentery by invading the intestinal epithelial tissue (Welch, 2006; Ishii and Sadowsky, 2008; Rasko *et al.*, 2008).

It is well known that upon entering the environment outside of the host, *E. coli* numbers show a steady decline over time. These studies show a negative growth rate of *E. coli* outside of the host implying that it does not survive outside of the host and that its presence is generally maintained in the secondary environment by the constant faecal input from the primary environment (Winfield and Groisman, 2003). This subsequently forms the basis of *E. coli's* use as an indicator organism. The general decline of *E. coli* in the environment is a consequence of factors such as temperature, moisture, low levels of nutrients, predation from protozoa, pH and UV. Therefore, the survival of *E. coli* in this environment requires it to overcome these factors and adapt in order to establish stable populations.

In this literature study, the genetic diversity within an *E. coli* population will be investigated, with special interest in its survival in an aquatic environment. Furthermore, it will be 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 being discussed. In addition, *E. coli* as an indicator organism will be discussed, along with the consequences of its existence in the external environment on its use as an indicator organism. This will be linked to the effects of the external environment on the overall genetics of *E. coli* as a population and multiple methods that have been used to characterise *E. coli* populations.

1.2 *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 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, facultatively 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 (WHO, Bartram and Pedly, 1996; 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 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 be true and in some cases unfounded. For example, we now know that *E. coli* is capable of proliferation in many environments and not only the gastrointestinal tract (see section 1.3). In addition, there is little evidence of a strict relationship neither 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 appears 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. It is likely then that significant changes in genetic diversity, occur during the transition between environments, through selection, and that these changes will be more significant within more complex environments (Gordon, 2001).

1.3 *E. coli* in the environment outside of the host

According to Gordon (2001) some *E. coli* find themselves in an environment outside of their host at some stage in their lifecycle 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, the primary and the secondary environment. The primary environment is represented by the gastrointestinal tract of the human or animal host, whereas 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).

It is well known that upon introduction into the external environment, there is substantial die-off of feacal bacteria over time. Multiple factors contribute to this decline in survival time including, sunlight and exposure to UV, temperature, limited nutrients, and predation, as mentioned above. Since the 1960s, the negative effect of sunlight on *E. coli* survival has been proven. Fujioka *et al*. (1981) and Davies and Evison (1991) both observed a dramatic decline of *E. coli* numbers as a result of exposure to sunlight with the rate of decline differing depending on summer verses winter periods (Noble *et al.*, 2004). Similar results were observed where the level of *E. coli* in the external environment is directly influenced by temperature (Noble *et al*., 2004). In addition, moisture content has also been shown to have an effect on *E. coli* levels. Beverdorf *et al.* (2006) showed that *E. coli* levels were significantly higher in sand with high moisture content as opposed to sands with low moisture content. Lastly, levels of *E. coli* in the external environment are influenced by the presence of indigenous microorganisms, either directly through predation or indirectly through competition for nutrients (Davies and Bavor, 2000).

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 (Solo-Gabriele *et al.*, 2000; Gordon *et al.*, 2002; Anderson *et al.*, 2005; Power *et al.*, 2005; Ishii *et al.*, 2006; 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* was always part of the microbiota in the external environment and that some strains acquired the ability to cause disease to 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).

1.3.1 *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. pneumoniae*, 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 agreed 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 could often be the main source of *E. coli* in freshwater systems (Solo-Gabriele *et al.*, 2000; Byappanahalli *et al.*, 2003a; Ishii *et al.*, 2007). Here, where *E. coli* strains survive, their 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 soil can persist in high numbers and following contact with water in high tide or rain, it acts as a constant source of *E. coli* into neighbouring water sources. According to Solo-Gabriele *et al.* (2000) *E. coli* are capable of multiplying to high cell concentrations upon drying of the soil. Therefore, when the water encounters the dry soil, large numbers of *E. coli* cells are released into the water. A similar scenario may also be implicated in the *E. coli* blooms in certain freshwater lakes, where depending on the season and rainfall, high rainfall may result in the flushing of *E. coli* from the soil banks into the water (Solo-Gabriele *et al.*, 2000; Ishii *et al.*, 2007).

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.

1.3.2 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 (Solo-Gabriele *et al.*, 2000; Sampson *et al.*, 2006). 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 (Wheeler Alm *et al.*, 2003; Brennan *et al.*, 2010). 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). By being a heterotrophic organism, *E. coli* is able to survive on low levels of carbon and nitrogen sources, in addition to other trace elements such 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, that it is able to integrate into the microbial communities in different environments (Bennett *et al.*, 1992; Ishii and Sadowsky, 2008).

1.3.3 Genetic diversity of naturalised *E. coli*

The persistence and proliferation of *E. coli* in the secondary environment raises the question: 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, 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. Here genes with products in low demand are under negative control and genes with products in high demand are under positive control and depending on the demand, these control systems alternate in the different environments (Savageau, 1983). 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). The expression of catabolic operons in *E. coli* is tightly regulated in all aspects in order to direct and control cellular activities.

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 the soil environment, 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 increase the health risks associated with recreational water, if they maintain the ability to circulate within the GI tract of the host and retain pathogenicity factors.

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 out compete 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.

1.4 *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 (Durso *et al.*, 2004; Yang *et al.*, 2004; Anderson *et al.*, 2006). 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 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).

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). However, *E. coli* populations do show a high level of genetic diversity (Touchon *et al.*, 2009). The population structure of *E. coli* is often defined as a balance between mutation and recombination. Here *E. coli* has the potential to change 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. 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).

1.5 *E coli* genomic diversity

E. coli strains show a significant difference at a genomic level in regards to 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 650kb. The results of this study suggested that the acquisition and removal of genetic information is not evolutionary 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 5020 genes. The conserved core genome size was calculated to consist of approximately 2200 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 13000 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 natural 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.

1.6 Characterisation of *E. coli* populations

1.6.1 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 was 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; Walk *et al*., 2007; Tenaillon *et al.*, 2010). MLEE has some drawbacks including problems with band resolution and mainly that it determines phenotypes rather than genotypes. The phenotype of an enzyme can easily change in response to a change in environment and therefore have a positive or negative effect on the MLEE results, making it difficult for standardisation.

1.6.2 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, B_1, B_2, D and E, although the members of group E are less common (Gordon, 2004; Gordon *et al.*, 2008). Groups A and B_1 are considered to be sister groups with group B_1 believed to represent the "ancestral lineage" of *E. coli* (Gordon, 2004; Gordon *et al.*, 2008). Group B_2 strains are monophyletic whereas strains belonging to group D are not and possibly represent two or more clades. In addition, strains belonging to groups B_2 and D have larger genomes than A and B_1 strains and the presence or absence of virulence factors involved in causing extra-intestinal disease may 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 B_2 and some to group D. Strains associated with environmental sources belong mostly to the B_1 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 B_1 appear to be generalists as they appear to cover a larger range of environments. In contrast, strains belonging to groups B_2 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 its product is responsible for 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, B_1, B_2 and D. The presence of *chuA* gene designates strains to either phylogroup B_2 or D. The presence of the *yjaA* gene then differentiates phylogroup B_2 from D and is present in most strains belonging to phylogroup A. Lastly, the presence of the TSPE4.C2 fragment differentiates phylogroup B_1 from A, being present in all B_1 strains (Clermont *et al.*, 2000; 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 compared 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 B_1, suggesting that specific genotypes were favoured by natural selection. Therefore, this B_1 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.

1.6.3 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 ERIC-PCR (McLellan *et al.*, 2003). PFGE was originally developed by Schwartz and Cantor (1984) where DNA molecules up to 2000 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).

1.6.4 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 (Dijkshoorn *et al*., 2001; Hahm *et al*., 2003; Leung *et al.*, 2004; Brady *et al*., 2007).

In a study by Guan *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 enterobacterial repetitive intergenic consensus polymerase chain reaction (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.

1.6.5 Multilocus sequence analysis (MLSA)

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 unequivocal (Dijkshoorn *et al.*, 2001; Walk *et al.*, 2007; Tenaillon *et al.*, 2010). 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 the closely related *Escherichia* species that is, *E. albetii* and *E. fergusonii* and *Shigella flexneri* as another representative *E. coli* strain. In addition, isolates collected from human, animal and, different from most previous studies, various environmental sources were included. 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 enterica* formed monophyletic clades. The remaining monophyletic clusters were named CladeI to CladeV 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* 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*, 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 novel clades are hard to differentiate from *E. coli* based on traditional phenotypic analysis and 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 (Solo-Gabriele *et al.,* 2000; Gordon *et al.,* 2002; Power *et al.,* 2005; Ishii *et al.,* 2006; Walk *et al.,* 2007).

1.7 Conclusions and 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. 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. However, 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 (Solo-Gabriele *et al.* 2000; Gordon *et al.*, 2002; Power *et al.*, 2005; Walk *et al*., 2007).

Further indication of the existence of environmental *E. coli* stains comes 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, that 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.

1.8 References

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

2. 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 Environmenal Microbiology*. **72**(11): 6914-6922.

3. 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.

4. 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.

5. 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.

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

7. Beversdorf, L. J., Bornstein-Frost, S. M and McLellan, S. L. (2006). The potential for beach sand to serve as a reservoir for *Escherichia coli* and the physical influences of cell die-off. *Journal of Applied Microbiology*. **102**: 1372-1381.

8. 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.

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

10. 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.

11. 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.

12. 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.

13. 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.

14. 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.

15. 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.

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

17. Davies, C. M and Evison, L. M. (1991). Sunlight and the survival of enteric bacteria in natural waters. *Journal of Applied Microbiology*. **70**(3): 265-274.

18. Davies, C. M and Bavor, H. J. (2000). The fate of stormwater-associated bacteria in constructed wetland and water pollution control pond systems. *Journal of Applied Microbiology*. **89**: 349-360.

19. 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.

20. 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.

21. 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.

22. 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.

23. 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.

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

25. 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.

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

27. Fujioka, R. S., Hashimoto, H. H., Siwak, E. B and Young, R. H. (1981). Effect of sunlight on survival of indicator bacteria in seawater. *Applied and Environmental Microbiology*. **41**(3): 690-696.

28. 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.

29. 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.

30. 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

31. 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.

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

33. 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.

34. Hahm, B., Maldonado, Y., Schreiber, E., Bhunai, A. K and Nakatsu, C. (2003). Subtyping foodborne and environmental isolates of *Escherichia coli* by multiplex-PCR, rep-PCR, ribotyping and AFLP. *Journal of Microbiological Methods*. **53**: 387-399.

35. Hartl, D. L and Dykhuizen, D. E. (1984). The population genetics of *Escherichia coli*. *Annual Reviews in Genetics.* **18**: 31-68.

36. 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.

37. 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.

38. 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. *Environmental Scence and Technology.* **41**: 2203-2209.

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

40. 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.

41. 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.

42. Lan, R and Reeves, P. R. (2002). *Escherichia coli* in disguise: molecular origin of Shigella. *Microbes and Infection.* **4**: 1125-1132.

43. 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.

44. 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.

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

46. 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.

47. Malan, T. P and McClure, W. R. (1984). Dual promoter control of the *Lac* operon. *Cell*. **39**(1):173-180.

48. 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.

49. 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.

50. Moriel, D. G., Rosini, R., Sieb, K . L., Serino,, L., Pizza, M and Rappuoli, R. (2012). *Escherichia coli*: Great diversity around a common core. *mBio*. **3**(3): e00118-12.

51. Noble, R. T., Lee, I. M and Schiff, K. C. (2004). Inactivation of indicator micro-organisms from various sources of faecal contamination in seawater and freshwater. *Journal of Applied Microbiology*. **96**: 464-472.

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

53. 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.

54. 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.

55. 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. *Environmental Microbiology*. **7**(5): 631-640.

56. Pratt, L. A and Kolter, R. (1998). Genetic analysis of *Escherichia coli* biofilm formation: roles of flagella, motility, chemotaxis and type I pili. *Molecular Microbiology*. **30**(2): 285-293.

57. Pupo, G. M., Lan, R and Reeves, P. R. (2000). Multiple independent origins of Shigella clones of *Escherichia coli* and convergent evolution of many of their characteristics. *PNAS.* **97**(19): 10567-10572.

58. 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.

59. 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.

60. 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.

61. 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.

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

63. Sekirov, I., Russell, S. L., Antunes, L. C. M and Finlay, B. (2010) Gut Microbiota in Health and Disease. *Physiological Reviews.* **90**: 859-904.

64. 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.

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

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

67. 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.

68. Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., Bouvet, O., Calteau, A., Chiapello, H., Clermont, O., Cruveiller, S., Danchin, A., Diard, M., Dossat, C., El Karoui, M., Frapy, E., Garry, L., Ghigo, J. M., Gilles, A. M., Johnson, J., Le Bougue´nec, C., Lescat, M., Mangenot, S., Martinez-Je´hanne, V., Matic, I., Nassif, X., Oztas, S., Petit, M. A., Pichon, C., Rouy, Z., Saint Ruf, C., Schneider, D., Tourret, J., Vacherie, B., Vallenet, D., Me´digue, C., Rocha, E. P. C and Denamur, E. (2009). Organised genome dynamics in the *Escherichia coli* species results un highly diverse adaptive paths. *PLOS Genetics.* **5**(1): e1000344.

69. 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.

70. 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. *Enviromental Microbiology.* **9**(9): 2274-2288.

71. 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.

72. Welch, R. A. (2006). The genus *Escherichia*. *Prokaryotes*. **6**: 60-71.

73. 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.

74. 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.

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

76. 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.

77. WHO: Bartram, J and Pedley, S. (1996). Chapter 10: Microbial analysis. Water Quality Monitoring - A Practical Guide to the Design and Implementation of Freshwater Quality Studies and Monitoring Programmes. Edited by Bartram, J and Ballance, R. United Nations Environment Program and World Health Organisation.

78. 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.

79. 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.

80. 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). Conparison between O serotyping method and multiplex real-time PCR to identify diarrheagenic *Escherichia coli* in Taiwan. *Journal of Clinical Microbiology*. **45**(11): 3620-2625.

CHAPTER TWO

THE DIVERSITY OF *E. COLI* WITHIN THE ROODEPLAAT DAM, SOUTH AFRICA

THE DIVERSITY OF *E. COLI* WITHIN THE ROODEPLAAT DAM, SOUTH AFRICA

2.1 Abstract

Escherichia coli is a highly diverse species, existing as both commensals and pathogens in the gastrointestinal tracts of humans and warm-blooded animals. It is this association with the gastrointestinal tract, which has led to its use as an indicator organism. This is based mainly on the assumption that its replication and growth is restricted to the GI tract. This may be the case for the majority of *E. coli* strains, although recent studies have indicated that some strains have become naturalised and are capable of diversifying and finding a new niche in the environment outside of the host. The aim of this study is to determine the diversity and dynamics within the *E. coli* population collected from the aquatic ecosystem in and around Roodeplaat dam. Samples types included water, sediment, algae and water hyacinths. As representatives of the primary environment, *E. coli* was also isolated from raw sewage obtained from sewage treatment works, which discharge their treated effluent into the dam. *rpoS* (RNA polymerase sigma factor S, sigma 38) gene sequencing was performed on the 194 isolates collected and revealed a high level of diversity within the population. Similar results were also observed with the AFLP analysis performed. In addition, a PCR-based phylogrouping technique was used to group *E. coli* isolates into one of the four common phylogenetic groups A, B_1, B_2 or D. The majority of the *E. coli* isolates grouped within the A and B_1 groups. The *rpoS* gene phylogeny also revealed possible environmental groups, where isolates clustered in the absence of sewage isolates. These results revealed a high level of diversity within the population with the indication that possible environmental groups may exist.

2.2 Introduction

The majority of *E. coli* occur as commensals in the gastrointestinal tracts of humans and animals of which a small proportion are pathogenic (Tenaillon *et al*., 2010). Based on decades of research, a number of assumptions are made concerning the survival of *E. coli* in the environment outside of the gastrointestinal tract of the host. These typically suggest that *E. coli* survives for only short periods and is unable to multiply outside of the host. In addition, it is believed that *E. coli* demonstrates a clonal composition, in that populations in the external environment are direct descendents of the original contaminant from the gut (Brennan *et al.,* 2010). Lastly, it is assumed that the amount of *E. coli* in the external environment is directly proportional to the faecal input from the host (Winfield and Groisman, 2003; Power *et al.,* 2005). All these assumptions have contributed to the use of *E. coli* as an indicator to monitor the faecal pollution of water.

The preferred or dominant environment of *E. coli* is increasingly being questioned. Savageau *et al.* (1983) was the first to report that *E. coli* exists in two distinct habitats. These are the primary environment, that consists of the gut of the host and the secondary environment, which refers to the external environment, outside of the host. It was estimated that *E. coli* spend up to half their life cycle in the secondary environment. Furthermore, *E. coli* surviving and proliferating in the secondary environment, in the absence of faecal contamination, has also been reported on several occasions. Sand and sediment has been shown to be a reservoir for *E. coli* and other faecal associated bacteria (Whitman *et al.,* 2006). Multiple studies have shown that *E. coli* persists in freshwater beaches and are responsible for the re-introduction of *E. coli* into freshwater systems (Solo-Gabriele *et al.,* 2000; Byappanahalli *et al.,* 2003a; Wheeler Alm *et al.* 2003; Byappanahalli and Fujioka 2004; Ishii *et al*., 2007). In addition, a study by Power *et al.* (2005) found that *E. coli* responsible for bloom events in a freshwater lake in Sydney, Australia had adapted to the freshwater environment by developing an outer capsule, which apparently allowed increased survival in the external environment. Lastly, *E. coli* strains have also been found to be associated with green algae in the genus *Cladophora,* suggesting that these eukaryotes provide a suitable environment for the survival and growth of *E. coli* (Byappanahalli *et al.,* 2003b; Badgley *et al.,* 2011).

E. coli populations associated with secondary environments sometimes appear to be genetically distinct from their gut-associated counterparts. Multiple studies have shown that selection plays a role in favouring certain genotypes more suited to the external environment, thus resulting in the development of possible niche-adapted populations (Whittam, 1989; Gordon *et al.,* 2002; Walk *et al.,* 2007). At the intraspecific level, *E. coli* is also known to be diverse with strains existing in a multitude of different environments and the presence of sub-species has been accepted (Ochman

and Selander, 1984). Therefore, a need exists to capture and quantify the diversity within the species and link it to their ecology.

Initially, studies involving intraspecific differentiation of *E. coli* strains were based on phenotypic variation (Selander and Levin, 1980; Tenaillon *et al.*, 2010). An example of this is serotyping, which is known to play an important role in determining sub-species, especially in terms of identifying possible pathogenic strains (Orskov and Orskov, 1992). Although, serotyping is valuable in disease detection and epidemiology, this method does not necessarily give an accurate representation of the total diversity of the species, and like other phenotypic markers do not correspond well with the evolution of the species (Lan and Reeves, 2002). Contemporary studies on the diversity of *E. coli* and other bacteria thus employ DNA-based markers (Clermont *et.*, 2000; Walk *et al.*, 2009). For example, Amplified Fragment Length Polymorphism (AFLP) fingerprint analysis has widely been used in identification, classification and differentiation of bacteria belonging to *Enterbacteriaceae*, including *E. coli* (Arnold *et al.*, 1999; Jonas *et al.*, 2003). Because it reflects polymorphisms from across the genome, AFLP gives a comprehensive view of the diversity between strains due to its high discriminatory power (Hahm *et al.*, 2003; Jonas *et al.*, 2003). More recently, sequence analysis has become an important method in determining the relationship between strains of the same species and the variation among strains at a population level (Tenaillon *et al.*, 2010).

In the year 2000 Clermont *et al.* introduced a rapid method for grouping *E. coli* strains based on their phylogenetic relationships. These so-called phylogroups have been suggested to give an indication as to the origin of the strain, as well as their potential to cause disease (Tenaillon *et al.*, 2010). The four main phylogroups that are currently accepted are groups A, B_1, B_2 and D, which can be distinguished based on the presence or absence of two genes (*chuA* and *yjaA*) and the DNA fragment TspE4.C2. Strains of the four groups differ from one another in terms of phenotypic characteristics, genome size and ecological niches. Strains assigned to phylogroups A and B_1 are considered to be associated with the environment whereas strains assigned to phylogroups B_2 and D are thought to be more often associated with the pathogenic strains and those associated with mammals (Gordon *et al.*, 2008; Tenaillon *et al.*, 2010). In addition, there are two other groups namely, groups E and F. Group E mostly consist of a small number of unassigned strains of which *E. coli* O157:H7 is a member whereas group F consists of strains closely related to group B_2 (Clermont *et al.*, 2013).

In light of the increasing amount of evidence for *E. coli* in the external environment, the diversity within the genus *Escherichia* was addressed in an extended Multi Locus Sequence Typing (MLST) study by Walk *et al.* (2009). Here they discovered five novel *Escherichia* clades, CI to CV. All five clades grouped more closely to *E. coli* than to *Salmonella enterica*. Of the five clades CIII, CIV and CV were identified as possible environmental groups. Isolates belonging to these groups were isolated from various environmental sources such as surface water and freshwater systems. This study suggested a higher level of diversity within the genus *Escherichia* than previously assumed, and specifically within the *E. coli* species, as the five clades were hard to differentiate phenotypically from *E. coli*. This study also depicted a high level of genotypic variation and complex evolutionary histories within *E. coli*.

Understanding the diversity, ecology and relationship between *E. coli* strains is important when addressing the suitability of *E. coli* as indicator for determining water quality. To study the diversity and dynamics of an *E. coli* population in an aquatic ecosystem the Roodeplaat dam was selected as study site. It had all the characteristics typically associated with such impoundments where habitats are continuously inter-mixed and not that well delineated. For example, sediments, algae and water plants form unique niches within the aquatic environment but are also in continuous contact with the overall water body, with no physical barriers between them. This dam not only serves as a recreational water source but also as an important drinking water source to parts of the city of Tshwane, South Africa. It also houses two sewage treatment works (Zeekoegat and Baviaanspoort), from which treated effluent is released back into the dam. The aim of the work reported in this chapter was to determine the diversity of *E. coli* within the Roodeplaat Dam and to determine whether the *E. coli* diversity in the aquatic environment mirrors the diversity of *E. coli* found in humans and warm-blooded animals. Further objectives were to determine if there are indications of unique environmental strains associated with the novel *Escherichia* clades described by Walk *et al.* (2009). To achieve these aims, AFLP and phylogroup analyses, as well as phylogenetic analysis based on the DNA sequence of the *rpoS* and 16S rRNA genes were used.

2.3 Materials and methods

2.3.1 Site description and sampling

The Roodeplaat Dam, previously named the Pienaars River Dam, was constructed in 1956 as a water supply for the surrounding landowners. It is situated on the Pienaars River (also known in some parts as the Moretele River), which is a tributary of the Crocodile River. The dam was originally constructed as an irrigation dam and but has also become popular for recreational use. It is an important source of water for the City of Tshwane. The water treatment plant on the dam supplements the water supply to the northern areas of Tshwane, namely Doornpoort, Montana, Wonderboom and Magaliesberg. Roodeplaat Dam forms part of the 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*).

Samples were collected in the month of October after the first rains of spring. Samples were then collected from eight sites around the Roodeplaat dam (Figure 2.1). Water, algae, water hyacinth and sediment samples were collected at the various depths around the dam. Algal, sediment and water samples were also collected from the Hartebeesspruit 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 to represent the *E. coli* strains circulating in the human and animal populations.

2.3.2 Comparing methods for *E. coli* isolations and Most Probable Number counts

To determine the most appropriate medium for the isolation of *E. coli* isolates an initial comparison was performed between McConkey (Oxoid), mFC (Merck, Biolab) and Membrane Lactose Glucuronide Agar (MLGA) (Oxoid). For this purpose, nutrient broth was inoculated with a previously isolated *E. coli* strain and allowed to grow over night at 37 °C. Serial dilutions were then performed and 100 µl of each dilution plated on McConkey, mFC and MLGA media using spread plate method. Plates were incubated for 24 hours at 25 °C, 37 °C and 42 °C for McConkey and mFC, and at 37 °C for MLGA.

Serial dilutions were prepared for the sewage samples expected to have a high colony count. Dilutions were prepared using Ringers (1/4 strength) solution up to 10^{-6}. Hundred µl of each dilution was plated out on to MLGA and incubated at 37 °C for 24 hours. Green colonies were

then randomly selected from the serial dilutions, streaked out for single colonies on MLGA, and again incubated at 37 °C for 24 hours. Each of the colonies selected were verified using Colilert® (IDEXX Laboratories). A single colony was picked and added to 5 ml of Colilert® solution in a sterile test tube and incubated at 37 °C for 24 hours. Colonies included in the study were those that turned the broth yellow and fluoresced under UV.

Environmental samples were processed by centrifuging 150 ml of each sediment sample at approximately 3000 g and re-suspending the pellet in 150 ml sterile distilled water. The sample was then sonicated for approximately 30 seconds to dislodge any bacteria from the algae or plant material into the surrounding solution. Serial dilutions were then prepared with the resulting sonicated solution. Algae were removed from approximately 1 L of dam water by filtration using a 0.45 µm filter (Whatman®) as it provides a pore size small enough to trap algal and bacterial cells on the membrane. The algal filtrate was then re-suspended in distilled water and sonicated for approximately 30 seconds after which the water was used for serial dilution. Processing of the water hyacinth samples involved placing the plant matter under aseptic conditions in sterile distilled water. The sample was then sonicated for ± 30 seconds and used for serial dilution.

Colilert® media was added to 100 ml of both processed samples (sediment, algal and water hyacinth) and dam water samples to determine the level of *E. coli* in the samples as well as for the isolation of specific strains. The Colilert® solution was poured into a Quanti-tray® and incubated at 37 °C for 18 hours. After incubation, Most Probable Number (MPN) counts were performed. For each sample type, 15 *E. coli* positive (fluorescent) wells were then selected at random. The fluorescing wells were cut open with a scalpel, under aseptic conditions, and samples streaked out on MLGA and incubated at 37 °C for 24 hours. Single colonies, confirmed to be *E. coli*, were selected for further DNA-based studies.

2.3.3 Determining phylogroups

Genomic DNA was extracted from the selected strains using Zymo Research Genomic DNA II kit™ (Fermentas), following manufacturer's instructions. Phylogroups were determined following the protocol described by Clermont *et al.* (2000), but with various modifications. DNA amplification was performed using three separate primer pairs shown in Table 2.1, as opposed to the original triplex method described by Clermont *et al.* (2000). 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 (QIAGEN). Amplification was performed using a Veriti™ Thermal Cycler (Applied Biosystems). The PCR products were mixed with gel red (Biotium) in a 1:5 volume ratio and subjected to electrophoresis (3.2 V/cm) for 30 min, using 1 % agarose (WhiteSci) gel and 1 X TAE buffer (40mM Tris-acetate and 1mM EDTA, pH 8.0) (Sambrook *et al.*, 1989; Brody and Kern, 2004). DNA was visualised under UV excitation and fragment sizes were estimated using a 1000 bp marker (Fermentas).

2.3.4 Amplified fragment length polymorphism (AFLP)

The AFLP method described by Vos *et al.* (1995) was followed, with various modifications. Fifty to 100 ng of genomic DNA of 104 isolates were digested with both *EcoRI* and *TruI* restriction enzymes. These 104 isolates were selected from amongst the total 194 to represent all sampling areas, namely, water, sewage, sediment, algae and water hyacinth along with known *E. coli* isolates (labelled *E. coli* culture in Figure 2.4). Each 15 µl restriction enzyme digestion reaction included 5 X Restriction/Ligation buffer (50 mM Tris-HAc (acetic acid), pH 7.5, 50 mM MgAc, 250 mM KAc (potassium acetate) and 25 mM DTT (dithiothreitol), 5 U *EcoRI* (Fermentas) and 4 U *TruI* (Fermentas). The digestion reaction was incubated at 37 °C for 2 hours, followed by a heating step at 70 °C for 15 minutes. To each 15 µl digestion reaction mixture, 5 pmol of *EcoRI* and 50 pmol *TruI* double stranded adaptors (see Table 2.2 for adaptor sequence) were added along with 5 X Restriction/Ligation, 0.3 mM ATP and 2.5 U T4 DNA Ligase (Fermentas). The resulting 20 µl ligation reaction mixture was incubated at 20 °C for 2 hours and then diluted 1:10 with nuclease free water (QIAGEN).

Pre-amplification involved a 25 µl PCR reaction containing the following: 2.5 mM $MgCl_2$, 250 µM of each nucleotide (dATP, dCTP, dGTP and dTTP), 10 µM of each EcoRI-O (5'-GAC TGC GTA CCA ATT C-3') and TruI-O (5'-GAT GAG TCC TGA CTA A-3') (Inqaba Biotechnologies), 1.5 U of Taq DNA polymerase and 10 X reaction buffer (Southern Cross Technologies), as well as 2 µl of diluted ligation reaction mixture. PCR conditions involved an initial denaturation at 94 °C for 3 minutes followed by 20 cycles of denaturation of 94 °C for 30 seconds, annealing at 56 °C for 1 minute and extension at 72 °C for 1 minute, and lastly a final extension of 72 °C for 5 minutes. Amplification was performed using a Veriti™ Thermal Cycler (Applied Biosystems). Each amplified product was then diluted 1:50 with nuclease free water (QIAGEN).

Multiple selective primer combinations were tried in order to achieve the number of bands sufficient to provide adequate information. Primer combinations included EcoRI-C/TruI-GC, EcoRI-C/TruI-TA, EcoRI-C/TruI-CG, EcoRI-G/TruI-GC, EcoRI-G/TruI-TA and lastly EcoRI-G/TruI-CG (Figure 2.3). Either the EcoRI-C or the EcoRI-G primer was fluorescently labelled. Selective amplification was carried out with each 20 µl reaction containing 2.5 mM $MgCl_2$, 250 µM of each nucleotide (dATP, dCTP, dGTP and dTTP), 100 µM of fluorescently labelled EcoRI-C (5'-GAC TGC GTA CCA ATT CC-3') and TruI-TA (5'-GAT GAG TCC TGA CTA ATA-3') (Inqaba Biotechnologies), 0.5 U of Taq DNA polymerase and 10 X reaction buffer (Southern Cross Technologies), as well as 5 µl of diluted pre-amplification product. Selective PCR conditions involved an initial denaturation at 94 °C for 5 minutes followed by 9 cycles of denaturation at 94 °C for 30 seconds, annealing of primers at 65 °C for 30 seconds and extension at 72 °C for 1 minute, and lastly a final extension of 72 °C for 5 minutes, decreasing the annealing temperature by 1 °C until an annealing temperature of 56 °C is reached. This was followed by 23 cycles of denaturation at 94 °C for 30 seconds, annealing of primers at 56 °C for 30 seconds and extension at 72 °C for 1 minute.

An 8% polyacrylamide gel was prepared consisting of 20 ml Long Ranger gel solution (LI-COR Biosciences) 7 M urea, 10 X TBE Buffer (890 mM Tris, 890 mM Boric acid, 20 mM EDTA, pH 8.0) (Brody and Kern, 2004), 150 µl 10% ammonium persulfate (APS) and 15 µl TEMED (Tetramethylethylenediamine) for polymerisation. The gel solution was poured into the LI-COR gel casting apparatus and left to polymerise for 60 minutes. The gel was pre-run for 30 minutes at 1500 V and 35 W to equilibrate the ions in the gel with the addition of 1 X TBE buffer. The selective amplification products were loaded with equal volumes of formamide loading buffer (95% formamide, 20 mM EDTA, pH8.0) and bromophenol blue. The mixture was heated for 3 minutes at 90 °C and cooled down on ice for 10 minutes. Approximately 0.5 to 0.8 µl of the mixture was loaded into the wells along with an IRD – 700 standardised marker in every tenth well. Gels were run on a LI-COR IR^2 automated sequencer (LI-COR Biosciences) with 0.8 X TBE running buffer, for 4 hours at 1500 V and 42 W.

After electrophoresis, banding patterns were analysed with BioNumerics software Version 6.1 (Applied Maths). The gel was normalised using the 700 bp standardised marker and the area between 50 bp and 700 bp was analysed. Using the UPGMA algorithm a dendrogram was constructed by applying Pearsons correlation coefficient with an optimisation value of 0 % and curve smoothing of 0 %.

2.3.5 *rpoS* gene sequencing

The *rpoS* gene encoding the RNA polymerase sub-unit sigma factor 38, was amplified using the protocol described by the online MLST database EcMLST (www.shigatox.net), with the following exceptions. Each isolate was amplified using the primers described by Walk *et al.* (2009), rpoS-F (5'-CGC CGG ATG ATC GAG AGT AA-3') and rpoS-R (5'-GAG GCC AAT TTC ACG ACC TA-3') (Inqaba Biotechnologies). The total reaction volume was reduced to 25 µl with 2.5 mM $MgCl_2$, 250 µM of each nucleotide (dATP, dCTP, dGTP and dTTP), 5 pmol of each primer pair, 2.5 U of Taq DNA polymerase, 10 X reaction buffer (Southern Cross Technologies), as well as and 4 ng/µl of genomic DNA.

Amplification was performed using a Veriti™ Thermal Cycler (Applied Biosystems). The PCR conditions involved an initial denaturation at 94 °C for 10 minutes followed by 35 cycles of denaturation of 92 °C for 1 minute, primer annealing at 58 °C for 1 minute and extension at 72 °C for 30 seconds followed by 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 (QIAGEN). The PCR products were subjected to agarose gel electrophoresis and visualized as described above.

PCR products were purified using 20 U/µl Exonuclease (Fermentas) and 1 U/µl Alkaline phosphotase (Fermentas), after which amplicons were sequenced with the forward primer rpoS-F (5'-CGC CGG ATG ATC GAG AGT AA-3') (Inqaba Biotechnologies) (Walk *et al.*, 2009). Each 12 µl sequencing reaction mixture included 0.5 µl ABI PRISM® BigDye® v3.1 sequencing reaction mix and 1 X sequencing buffer, 100 µM undiluted primer and approximately 150 ng of amplified PCR product. Sequencing amplification involved an initial denaturation at 96 °C for 5 seconds followed by 25 cycles of denaturation of 96 °C for 10 seconds, primer annealing at 55 °C for 5 seconds and extension at 60 °C for 4 minutes. Products were then sequenced using an ABI 3130 Prism DNA Automated Sequencer (Perkin-Elmer).

ABI sequence files were examined and edited where needed using BioEdit Sequence Alignment Editor V 7.0.9.0 (Hall, 1999). Sequences were compared to those in the the NCBI GenBank database using BLAST. Reference sequences of the *rpoS* gene of *E. coli*, *E. fergusonii* and *E. alberti* were obtained from the GenBank database in addition to *rpoS* gene sequences of cryptic *E. coli* Clades obtained from Walk *et al.* (2009). All sequences were then aligned using MAFFT (Version 6) online alignment tool (Katoh *et al.*, 2002) and trimmed again using BioEdit Sequence Alignment Editor V 7.0.9.0 (Hall, 1999). jModeltest program (Posada, 2008) was used for the selection of an appropriate model. Maximum Likelihood trees (Felsenstein, 1981) were constructed using the phylogeny software PhyML 3.0 (Guindon *et al.*, 2010).

2.3.6 16S rRNA gene sequencing

Amplification of the 16S rRNA gene was performed on those isolates giving irregular bands after *rpoS* gene amplification. The 16S rRNA gene was amplified using the universal primers 16F27 (5'-AGA GTT TGA TCC TGG CTC AG-3') and 16R1522 (5'- AAG GAG GTC ATC CAG CCG CA - 3') (Inqaba Biotechnologies) (Coenye *et al.,* 1999). Each 25 µl reaction mixture contained the following: 2.5 mM $MgCl_2$, 250 µM of each nucleotide (dATP, dCTP, dGTP and dTTP), 10 uM of both the forward and reverse primers, 2.5 U of Taq DNA polymerase and 10 X reaction buffer (Southern Cross Technologies) and 50 – 100 ng of genomic DNA.

Amplification was performed using a Veriti™ Thermal Cycler (Applied Biosystems). PCR conditions involved an initial denaturation at 94 °C for 10 minutes followed by 30 cycles of denaturation of 94 °C for 1 minute, primer annealing at 58 °C for 1 minute and extension at 72 °C for 1 minute followed by a final extension of 72 °C for 5 minutes. A negative control of nuclease free water was included, as previously described. The PCR products were subjected to agarose gel electrophoresis and purification as described above.

Purified amplicons were then sequenced as before, by making use the forward primer 16F27 (5'-AGA GTT TGA TCC TGG CTC AG-3'), the reverse primer, 16R1522 (5'- AAG GAG GTC ATC CAG CCG CA - 3') and an internal primer 16F536 (5'-CAG CAG CCG CGG TAA TAC-3') (Inqaba Biotechnologies) (Coenye *et al.*,1999). Using the resulting sequence information a consensus sequence was constructed for each isolate using CLC Main Workbench software version 5.5 (CLC Bio). Homology search was performed for each sequenced isolate using the BLAST program provided by the NCBI GenBank database.

2.4 Results

2.4.1 *E. coli* isolations and Most Probable Number counts

In comparison to McConkey (Oxoid) and mFC (Merck, Biolab) agars, MLGA (Oxoid) proved to be the most effective in isolating *E. coli*. Green colonies were clearly visible and identified as *E. coli* at 37°C. Isolates were labelled according to sample sites corresponding to numerical codes used by the Water affairs stationed at the Roodeplaat Dam (e.g. Q01, Q02, Q07 etc.) (Figure 2.1). In addition, Colilert® was a successful verification tool as *E. coli* isolates were easily identified as the isolates fluorescing under UV light. Most Probable Number counts for all sample types are listed below in Table 2.3 and Table 2.4.

Water isolated from the intermittent stream, Hartbeesspruit, clearly had the highest MPN of *E. coli* with 198.9 cfu/100ml. The sample site Q10 also had a high MPN of *E. coli* with 116.9 cfu/100ml. Conversely, water obtained from a 25m depth at sample site Q01 had low levels of bacteria as indicated by the MPN of <1 cfu/100ml and as a result no *E. coli* isolates were collected from this sample.

For algal, sediment and water hyacinth samples, MPN counts were generally higher than for those of the water samples. Sediment collected from the Hartbeesspruit River leading into the dam indicated the highest MPN of all sample types with $9.6x10^4$ cfu/100ml. In addition, algae collected from the both the Hartbeesspruit and the Jetty also indicated high MPN counts of 416.0 cfu/100ml and 6.6 x 10^3 cfu/100ml, respectively. Also taking note of samples collected from the site Q02, the water hyacinth sample indicates a much higher MPN count of 191.8 cfu/100ml as opposed to the 9.7 cfu/100ml obtained from the water collected from the same site.

From all samples collected, 194 *E. coli* isolates were obtained. Of the 194 isolates, 52 were from the sewage samples, 66 from the water samples, 13 from the water hyacinth, 45 from the algal samples and lastly 18 from the sediment sample. Sample names and type are listed in Table 2.5.

2.4.2 Determining phylogroups

Following PCR, isolates were assigned to phylogroups based on the scheme presented in Figure 2.2. Placing the isolates in phylogroups gave an indication of their origin and potential pathogenicity. The presence or absence of the two genes *chuA* and *yjaA* and the DNA fragment TspE4.C2 were determined and the phylogroup of each strain was determined. The numbers of isolates of each sample type belonging to the four main phylogroups are listed in Table 2.6.

The majority of sewage isolates belonged to phylogroups B_2 and A. In other sample types, water, sediment and algae the vast majority of isolates belong to phylogroup B_1. The *E. coli* isolated from the water hyacinth were all identified as belonging to phylogroup B_2. Overall, the majority of isolates belonged to phylogroup B_1 constituting 41.80% of the 194 isolates. Phylogroup B_2 constituted 20.10% of the total number of isolates, phylogroup D constituted 15.50% and phylogroup A constituted 9.80%. Those isolates giving no amplicons for either of the two genes, *chuA* and *yjaA*, or the DNA fragment TspE4.C2 were assigned as unknowns. Although, Clermont *et al.* (2000) states that isolates yielding no amplicons should be assigned to group A, Gordon *et al.* (2008) later stated that those strains failing to yield product rarely fall within Group A and should not be assigned to a phylogroup. Unknown isolates constituted 12.80% of the total number of isolates.

2.4.3 AFLP and LI-COR analysis

AFLP analysis was performed initially to determine if any unique environmental *E. coli* clusters were present. EcoRI-C and TruI-TA selective primer combination gave the suitable number of bands, averaging approximately 60 bands per isolate. This primer combination gave well-defined bands with high resolution when compared to the other selective primer combinations (Figure 2.3.). EcoRI-C/TruI-TA was therefore chosen as the preferred primer combination for the AFLP analysis of obtained *E. coli* isolates.

Overall, the UPGMA analysis of the AFLP data indicated a high level of diversity among the isolates (Figure 2.4). One possible environmental cluster (Cluster 1) was observed, which included no sewage samples. Cluster 1 consisted of five isolates including strains isolated from algae, water hyacinth and water samples. However, this cluster showed little resolution with a similarity value of only 38%. A second cluster (cluster 2) was observed, consisting of strains isolated from algae and water samples along with an *E. coli* culture from an unknown water source. The similarity value among isolates was, however, also low at approximately 40%.

2.4.4 *rpoS* Sequence analysis

The *rpoS* gene was sequenced to determine if any of the *E. coli* isolates grouped within the cryptic clades defined by Walk *et al.* (2009). The *rpoS* gene was selected, as it was one of the few genes that grouped monophyletically in all observed clades in the MLST study, performed by Walk *et al*. (2009). Following sequencing of the *rpoS* gene for all isolates, two Maximum Likelihood trees were constructed. The first tree was constructed from aligned *rpoS* sequences of all *E. coli* isolates with the addition of *rpoS* sequence information of cryptic *E. coli* clades obtained from Walk *et al.* (2009) (Figure 2.5). The second tree was constructed based on only the *E. coli* isolates identified as true *E. coli* and excluding the *rpoS* sequence information of cryptic *E. coli* clades obtained from Walk *et al.* (2009) (Figure 2.6).

Within the tree containing the sequences for the cryptic clades, two main groups were visible (Figure 2.5). The first consisted solely of the Roodeplaat Dam *E. coli* isolates, whereas the second group contained the five cryptic *Echerichia* clades, defined by Walk *et al.* (2009), which were well defined throughout the tree with strong bootstrap support. None of the Roodeplaat Dam *E. coli* isolates grouped with any of the five cryptic species clades. All the *E. coli* isolates grouped with the true *E. coli* strains (E667 and B692).

Longer *rpoS* sequences were available for the construction of this second tree, which provided better resolution among the true *E. coli* isolates. Resolution among the strains improved although, a large number of the isolates still formed one poorly defined group of which the phylogroups are mixed. However, as resolution improves, phylogroups become more apparent with isolates sharing the same phylogroup clustering together. An environmental group (Cluster 1) was visible which included only strains isolated from water, algae and sediment samples and the majority of which belonged to phylogroup B_1. The strains isolated from the water hyacinth formed a clearly defined group in the absence of any sewage isolates indicating another possible environmental group (Cluster 2) and all belonging to phylogroup B_2.

2.4.5 16S rRNA gene sequence analysis

Multiple isolates produced irregular bands or no amplicons at all, during amplifications of the *rpoS* gene. A 16S rRNA gene sequence analysis was performed to confirm whether these isolate were in fact true *E. coli.* A region of approximately 1300 bp was used for each sequenced isolate in the BLAST searches (Table 2.7). These analyses confirmed that the majority of the isolates tested were identified as *Citrobacter* sp. with high similarity values. In addition, one isolate Q081 was identified as *Moellerella wisconsensis*, another, KW2 was identified as *Plesiomonas shigelloides*, and lastly Q074 was identified as *Enterobacter amnigenus* all with similarity values of 99 %. These isolates were excluded from further analysis.

A Maximum Likelihood tree (Figure 2.7) constructed from the 16S rRNA gene squences corresponded with the BLAST results. The majority of isolates suggested by BLAST to represent *Citrobacter* grouped with *Citrobacter freundii*, *Citrobacter werkmani* and *Citrobacter murliniae*. However, two isolates Q104 and Q108 grouped more closely with *Citrobacter braakii* despite BLAST results indicating they were *Citrobacter freundii.* As predicted from the BLAST results, isolate Q074 grouped closely with *Enterobacter amnigenus* and isolate KW2 with *Plesiomonas shigelloides.* The isolate Q081 grouped closely with the outgroup, *Moellerella wisconsensis*.

2.5 Discussion and conclusions

E. coli strains were successfully isolated using MLGA and verified using Colilert® media. MPN counts indicated that sample types associated with sediment, algae and water hyacinth produced higher numbers of *E. coli* on average than the water samples. These results correlate to multiple studies, which state that soil and sediments are often a main source of *E. coli* in freshwater systems (Solo-Gabriele et al., 2000; Byappanahalli et al., 2003a; Ishii et al., 2007). In addition, Byappanahalli *et al*, (2003b) states that the macro-algae *Cladophora*, supports the growth of *E. coli* in freshwater systems. Water hyacinth may have the same effect on the growth of *E. coli*.

E. coli counts were also higher in the Hartebeesspruit water than in the dam water. This may be a consequence of the flow of water in the river disrupting the sediment and resulting in increased *E. coli* counts (Whitman *et al.*, 2006).

False positives were recognised as MLGA isolates producing no amplicons or irregular bands for both phylogrouping and *rpoS* sequences analyses were observed. The observed false positive results are comparable to those seen by Eccles *et al.* (2004) where a false positive rate of 3 % was observed when using MLGA. These isolates were clearly identified as *Citrobacter* species and were excluded from further analysis. Isolation methods allowed for the presence of false positives but 16S rRNA gene sequence analysis successfully identified those isolates as species of *Citrobacter, Enterobacter amnigenus, Moellerella wisconsensis* and *Plesiomomas shigelloides.* The isolation methods of Colilert® and MLGA media are designed for not only the isolation of *E. coli* but also other coliforms such as *Citrobacter*. It is therefore not a surprise that the majority of false positives were identified as *Citrobacter* species. They are also associated with the gastrointestinal tracts of humans and are found in almost everywhere in water, wastewater, soils and sediments (Gauthier and Archibald, 2001).

Strains belonging to the four phylogroups of *E. coli* were observed in this study. These groups are believed to 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 B_1 (Gordon *et al.*, 2008; Tenaillon *et al.*, 2010). Consistent with this, 51.6% of the Roodeplaat *E. coli* isolates belong to phylogroups A and B_1. More specifically, 41.8% of the isolates belonged to phylogroup B_1, which is more than any other phylogroup. Walk *et al.* (2007) stated that despite recombination events in nature, the phylogroup B_1 is favoured by natural selection in the secondary environment. The results of the current study thus suggest that the secondary environment plays an important role in shaping the observed diversity of *E. coli* in the Roodeplaat Dam.

Interestingly, all strains isolated from the water hyacinth (Q02H) were assigned to phylogroup B_2. Strains belonging to phylogroup B_2 typically are associated with mammalian and pathogenic strains (Clermont *et al.*, 2000; Walk *et al.*, 2007). In addition, the majority of sewage isolates also belong to phylogroup B_2 however, *rpoS* gene sequence analysis indicates that the water hyacinth isolates were genetically distinct from the isolates representing those present in the human population. It would therefore be interesting, for future work, to investigate their pathogenic properties.

The phylogroups to which isolates belong correlated to some extent with the *rpoS* clusters that were recovered. In the parts of the *rpoS* phylogeny where resolution between strains was low, a mixture of all phylogroups was observed. However, as resolution improved and more distinct groups became apparent, some groups corresponding to or containing phylogroups were better supported. The tree was dominated by phylogroups B_1, B_2 and D. A similar relationship between phylogroups was observed in the study by Gordon *et al.* (2008) where phylogroups clustered together in other MLST analyses. Clermont *et al.* (2000) states that isolates yielding no PCR products for either the two genes (*chuA* and *yjaA*) or the DNA fragment (TspE4.C2) should be assigned to phylogroup A. However, Gordon *et al.* (2008) stated that based on MLST analysis, isolates yielding no PCR products, rarely belong to phylogroup A and should be assigned as unknowns. This could explain the low level of correlation in some parts of the tree.

AFLP was performed, on a subset of representative isolates, in the hope that the data would clearly reveal the level of diversity among the *E. coli* isolates and if any potential separate environmental clusters were present without any sewage isolates being part of the group. Two possible environmental clusters were observed, however resolution was low and there was no clear correlation between the two clusters produced by AFLP analysis and the clusters produced by the *rpoS* gene sequence analysis. Overall AFLP fingerprint analysis indicated a high level of diversity among *E. coli* strains but no clear clusters are visible or well supported. In addition, there was poor correlation between phylogroups and the groupings visible in the AFLP dendrogram. These inconclusive results may be a consequence of the method being based on the whole genome. *E. coli* is known to have a mosaic genome that is highly diverse in terms of genome size and composition, depending on the specific strain. This may account for the low resolution and high level of diversity observed in the AFLP dendrogram. The choice to move forward with sequence analysis of a maintained housekeeping gene (*rpoS*) allowed for the better differentiation and grouping of strains. In addition, correlation between phylogroups and the phylogenetic clustering improved as opposed to the correlation observed with the AFLP data.

In this study, a portion of the *rpoS* gene was used to show the phylogenetic relationship between isolates. This 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). Using a MLSA approach of 22 housekeeping genes, Walk *et al.* (2009) discovered the presence of five novel clades (Clades I –V) within the cryptic *E. coli* species. The *rpoS* gene along with two others (*fumC* and *lysP*) indicated a possible monophyletic origin for all clades. Based on their *rpoS* sequence, all of the 194 *E. coli* strains isolated in the current study grouped with the true *E. coli* and none of them could be linked to the cryptic species described by Walk *et al.* (2009). This suggests that the Roodeplaat Dam aquatic environment is dominated by true *E. coli*.

Two possible environmental clusters were detected among the set of *E. coli* isolates from the Roodeplaat Dam. The first consisted of water, algae and sediment isolates and the second of only water hyacinth isolates. These clusters were evident in the *rpoS* phylogenies despite the generally low resolving power of this gene. However, a large number of isolates group together with poor resolution and they consist of mixtures of both environmental isolates and sewage isolates. This mixture of isolates may be because sewage strains are released into the dam and are therefore indistinguishable from water isolates. Because there is little or no physical barrier preventing strains from coming into contact with each other, it might thus be misleading to separate isolates based on sampling site. Nevertheless, as the resolution improves clusters become more apparent and the diversity of strains becomes more evident. The absence of sewage isolates within some of the better supported clusters indicate the possibility of these clusters being environmental groups, which have adapted to the environment outside of the host. The possibility exists that these unique environmental clusters may have adapted to the environment outside of the primary host and found a unique niche within the aquatic system.

The presence of possible environmental groups in this study correlates to findings in a study by Byappanahalli *et al*. (2006). They 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 environment consistently grouped separately in the maximum Likelihood trees. All of the water hyacinth isolates consistently grouped together in one cluster in the *rpoS* phylogenetic analyses and in the absence of faecal strains. None of these isolates were found anywhere else in the trees.

A number of isolates from this study could not be differentiated from *Shigella flexneri* and *Shigella dysenteriae*. Because *Shigella* forms part of the species *E. coli* (Pupo *et al.*, 2000; Lan and Reeves, 2002), this result was expected. *Shigella* species are generally regarded as the more pathogenic strains within the *E. coli* species complex. It would be interesting to investigate the potential pathogenicity of those true *E. coli* strains that group closely with the known *Shigella* reference strains.

It is well known that *E. coli* is a highly diverse species. However, the majority of diversity studies are based on human, clinical and pathogenic strains (Touchon *et al.*, 2009; Walk *et al.*, 2009; Tenaillon *et al*., 2010). This chapter reveals that there is also a high level of diversity within the *E. coli* populations in aquatic environments such as the Roodeplaat Dam. AFLP analysis initially revealed a high level of diversity among *E. coli* strains and *rpoS* sequence analysis 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.

These clusters may represent populations that have adapted to a niche within the aquatic environment, especially those associated with water hyacinth. The high level of diversity among the Roodeplaat Dam 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 with the addition of *E. coli* isolates collected from the Rietvlei Dam and sequence analysis of an alternative gene. It would be interesting to investigate whether the diversity and apparent environmental groups are maintained when combined with isolates from another aquatic environment and if their phylogenetic position is determined based on more variable gene regions. In addition, future work should possibly include an attempt to correlate the *E. coli* found in this water system to those *E. coli* populations associated with water birds.

2.6 References

1. Arnold, C., Metherell, L., Willshaw, G., Maggs, A and Stanley, J. (1999). Predictive fluorescent amplified-fragment length polymorphism analysis of *Escherichia coli*: High-resolution typing method with phylogenetic significance. *Journal of Clinical Microbiology*. **37**: 1274-1279.

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

3. 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.

4. Brody, J. R and Kern. S. E. (2004). History and principles of conductive media for standard DNA electrophoresis. *Analytical Biochemistry*. **333**: 1-13.

5. 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.

6. 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.

7. 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.

8. 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 Micobiology*. **8**(3): 504-513.

9. 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.

10. Clermont, O., Christenson, J. K., Denamur, E and Gordon, D. M. (2013). The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environmental Microbiology Reports.* 5(1): 58-65.

11. Coenye, T., Falsen, E., Vancanneyt, M., Hoste, B., Govan, J.R.W., Kersters, K and Vandamme, P. (1999) Classification of *Alcaligenes faecalis*-like isolates from the environment and human clinical samples as *Ralstonia gilardii* sp. nov. *International Journal of Systematic and Evolutionary Microbiology*. **49**: 405-413.

12. Eccles, J. P., Searle, R., Holt, D and Dennis, P. J. (2004). A comparison of methods used to enumerate *Escherichia coli* in conventionally treated sewage sludge. *Journal of Applied Microbiology*. **96**(2): 375-383.

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

14. Gauthier, F and Archibald F. (2001). The ecology of "fecal indicator" bacteria commonly found in pulp and paper mill water systems. *Water research.* **35**(9): 2207-2218.

15. 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.

16. 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.

17. 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):307-21.

18. Hall, T.A. (1999) BioEdit: A user- friendly biological sequence alignment editor and analysis program. *Nucleic acids Symposium Series*. **41**: 95-98.

19. Hahm, B., Maldonado, Y., Schreiber, E., Bhunai, A. K and Nakatsu, C. (2003). Subtyping foodborne and environmental isolates of *Escherichia coli* by multiplex-PCR, rep-PCR, ribotyping and AFLP. *Journal of Microbiological Methods*. **53**: 387-399.

20. 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.

21. 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.

22. 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.

23. Lan, R and Reeves, P. R. (2002). *Escherichia coli* in disguise: molecular origin of Shigella. *Microbes and Infection.* **4**: 1125-1132.

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

25. Orskov, F and Orskov, I. (1992). *Escherichia coli* serotyping and disease in man and animals. *Canadian Journal of Microbiology.* 38: 699–704.

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

27. 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.

28. Pupo, G. M., Lan, R and Reeves, P. R. (2000). Multiple independent origins of Shigella clones of *Escherichia coli* and convergent evolution of many of their characteristics. *PNAS.* **97**(19): 10567-10572.

29. Sambrook, J., Fritsch, E. F and Maniatis, T. (1989). Molecular cloning: A laboratory manual. Cold Springs Harbour Press. Cold Spring Habour. New York.

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

31. Selander, R. K and Levin, B. R. (1980). Genetic diversity and structure in *Escherichia coli* populations. *Science*. **210**: 545–547

32. 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.

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

34. Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., Bouvet, O., Calteau, A., Chiapello, H., Clermont, O., Cruveiller, S., Danchin, A., Diard, M., Dossat, C., El Karoui, M., Frapy, E., Garry, L., Ghigo, J. M., Gilles, A. M., Johnson, J., Le Bougue´nec, C., Lescat, M., Mangenot, S., Martinez-Je´hanne, V., Matic, I., Nassif, X., Oztas, S., Petit, M. A., Pichon, C., Rouy, Z., Saint Ruf, C., Schneider, D., Tourret, J., Vacherie, B., Vallenet, D., Me´digue, C., Rocha, E. P. C and Denamur, E. (2009). Organised genome dynamics in the *Escherichia coli* species results un highly diverse adaptive paths. *PLOS Genetics*. **5**(1): e1000344.

35. 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.

36. 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.

37. 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.

38. 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.

39. 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.

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

41. 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.

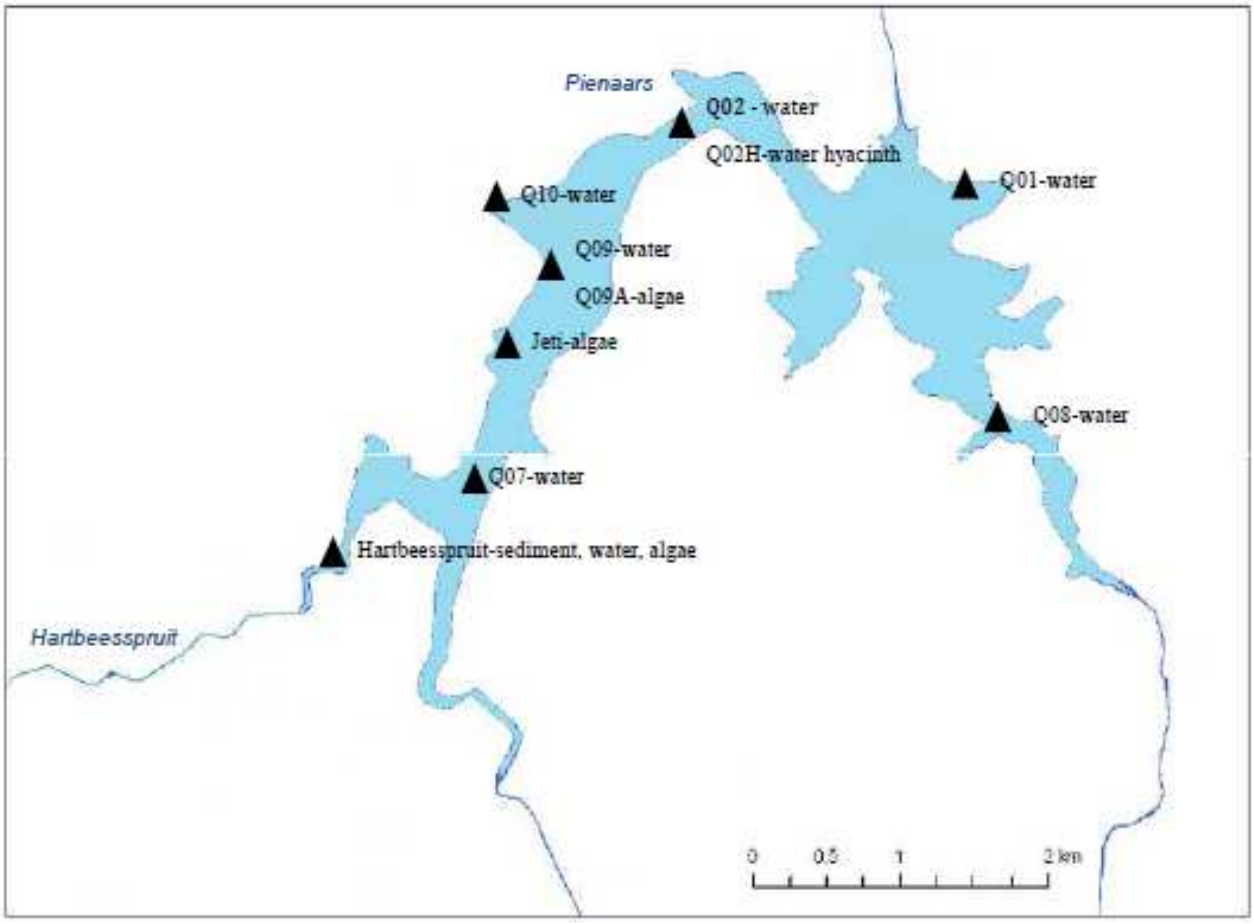

Figure 2.1: Map showing the Roodeplaat dam, Pretoria and sites of sample collection in and around the dam (represented by ▲). Numerical codes (Q01 – Q10) indicate the sampling points used by the department of water affairs housed at the dam. Also indicated are the sample types obtained at each sampling point.

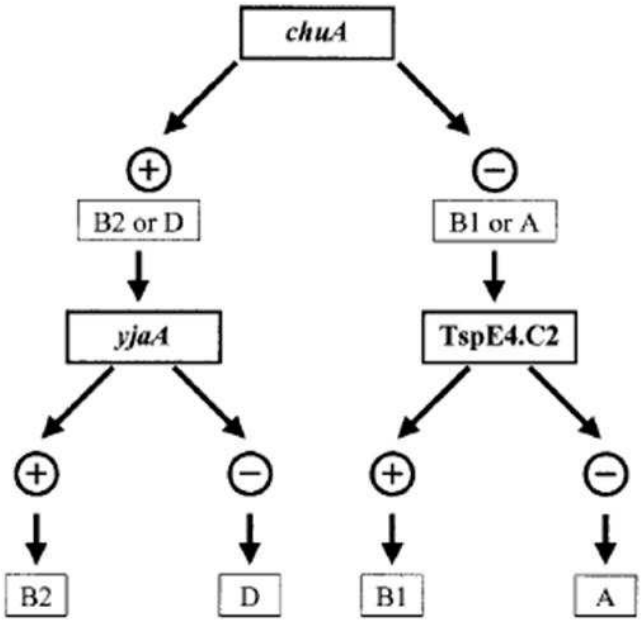

Figure 2.2: Dichotomous decision tree to determine the phylogenetic group of an *E. coli* strain by using the result of PCR amplification of the *chuA* and *yjaA* genes and the DNA fragment TspE4.C2 (from Clermont *et al.*, 2000).

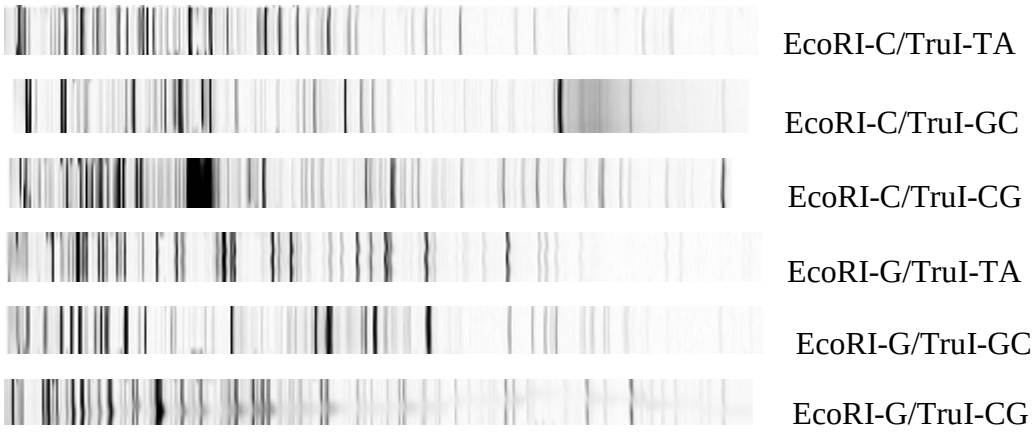

Figure 2.3: AFLP fingerprint of an isolated *E. coli* strain comparing the banding patterns generated from the different selective primer combinations EcoRI-C/TruI-TA, EcoRI-C/TruI-GC, EcoRI-C/TruI-CG, EcoRI-G/TruI-TA, EcoRI-G/TruI-GC and EcoRI-G/TruI-CG.

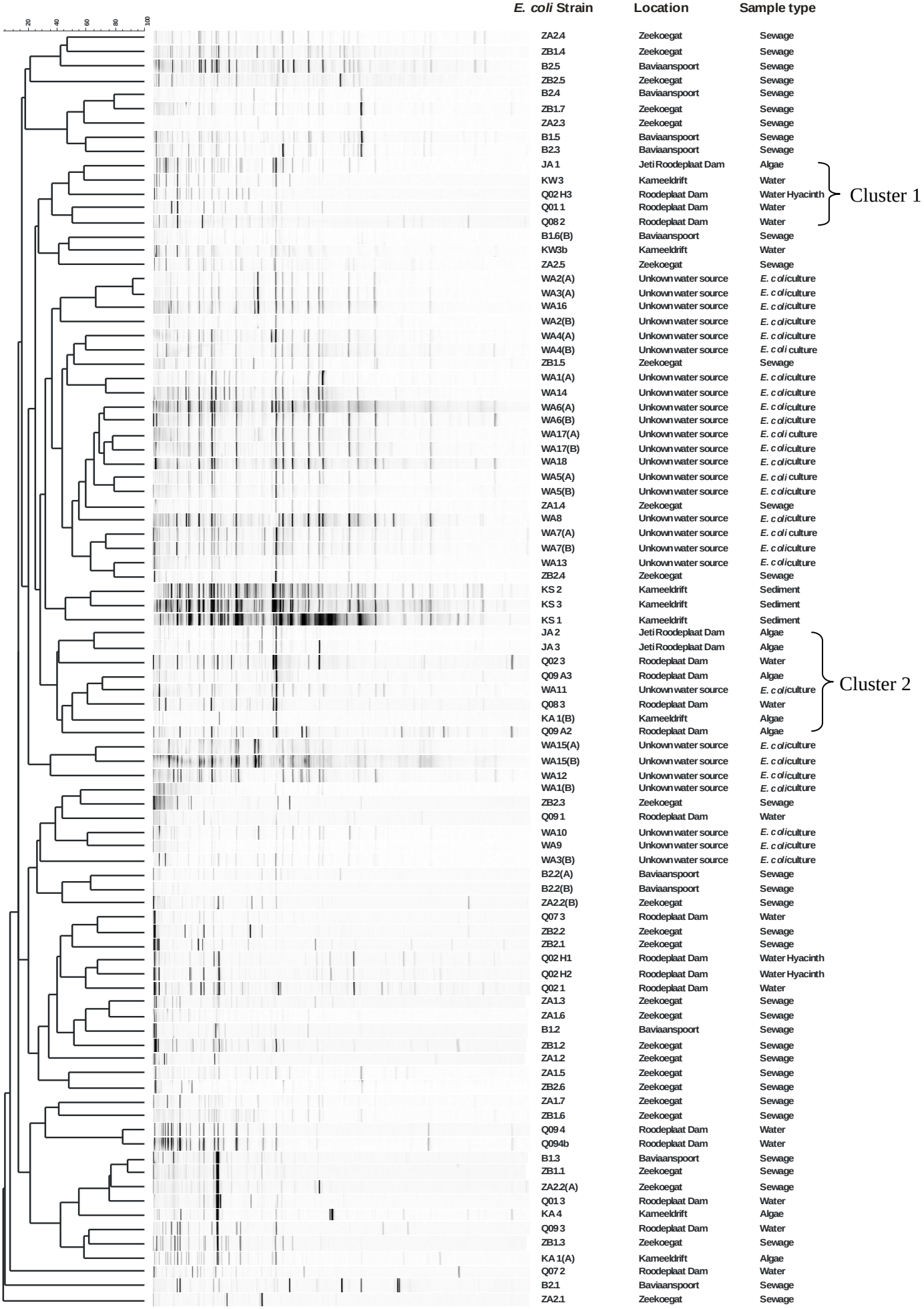

Figure 2.4: UPGMA dendrogram based on the AFLP fingerprint analysis of selected *E. coli* isolates from each sample type using the primer combination EcoRI-C/TruI-TA. The levels of similarity representing the Pearsons co-efficient, are expressed as percentages. The banding patterns adjacent to each branch were normalised and the background subtracted and processed using BioNumerics software version 6.1 (Applied Maths).

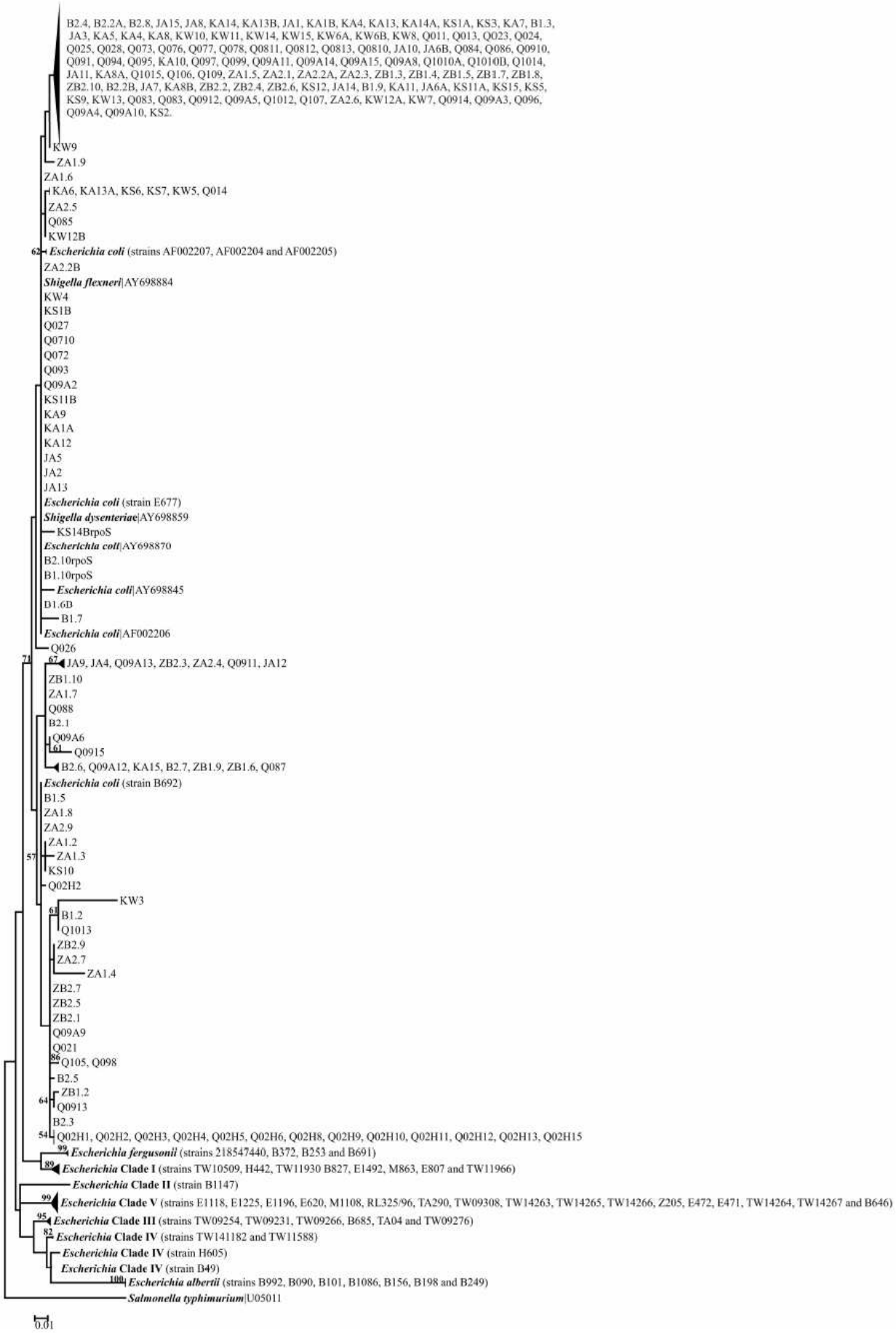

Figure 2.5: Maximum Likelihood phylogenetic tree showing the relatedness between *E. coli* isolates isolated from the Roodeplaat Dam and adjacent sewage treatment works. The tree is based on the *rpoS* sequence information of all *E. coli* isolates including *rpoS* sequence information of cryptic *E. coli* clades obtained from Walk *et al.* (2009). With *Salmonella enterica* as the outgroup and bootstrap analysis of 1000 replicates (Bootstrap values are indicated as percentages).

Figure 2.6: Maximum Likelihood phylogenetic tree showing the relatedness between *E. coli* isolates, isolated from the Roodeplaat Dam. The tree is based on the *rpoS* sequence information of all true *E. coli* isolates. With *Escherichia fegusonii* as the outgroup and bootstrap analysis of 1000 replicates (Bootstrap values are indicated as percentages). Phylogroups are represented in colour with Group A in red, Group B_1 in blue, Group B_2 in green, Group D in purple and the unknowns left in black.

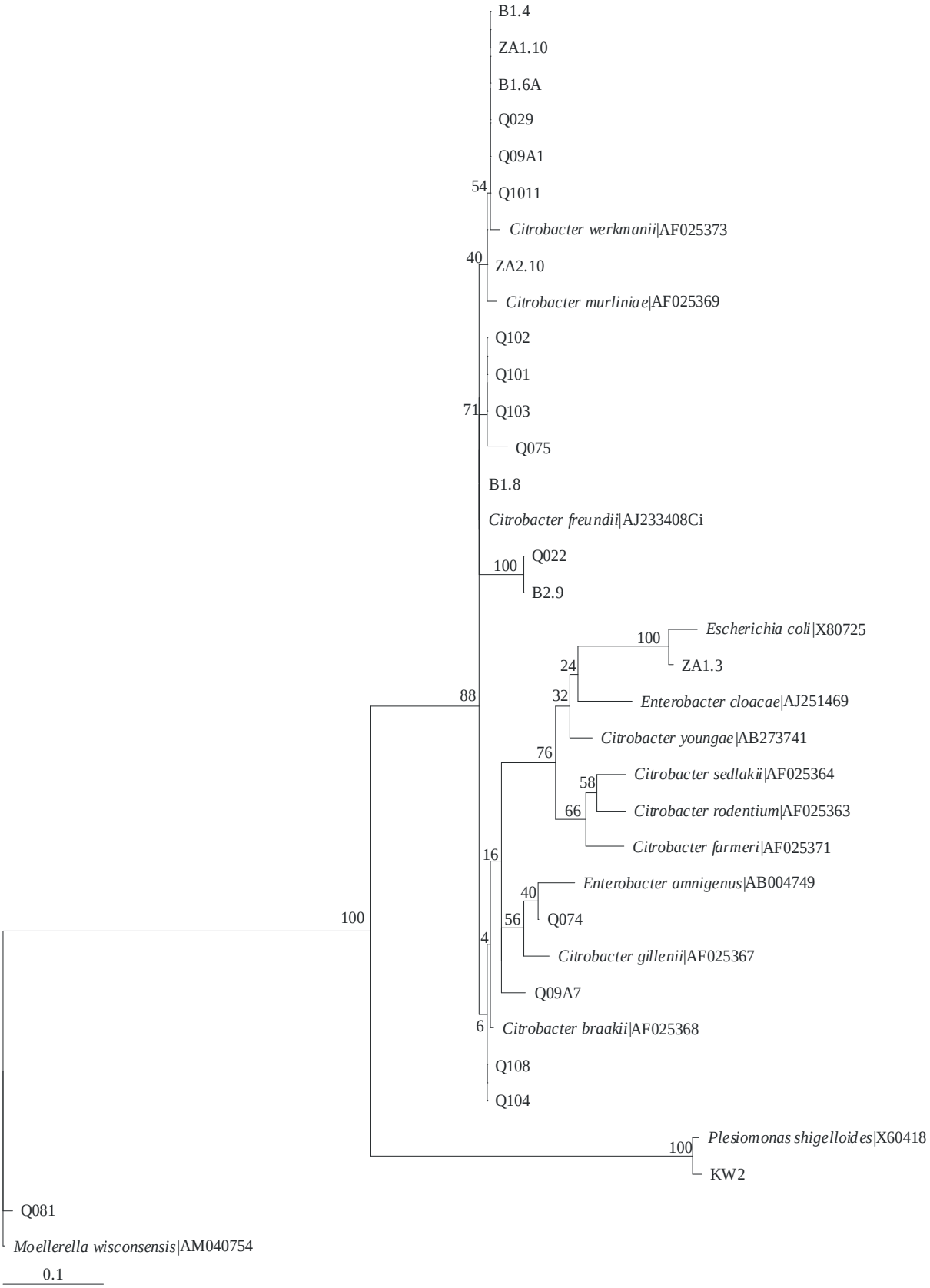

Figure 2.7: Maximum Likelihood phylogenetic tree based on the 16S rRNA gene sequence of those isolates from the Roodeplaat Dam producing irregular *rpoS* PCR results. With *Moellerella wisconsensis* as the outgroup and bootstrap analysis of 1000 replicates (Bootstrap values are indicated as percentages).

Table 2.1: Primer pairs used in the determination of phylogenetic groups and expected amplicon sizes, described by Clermont *et al.* (2000)

Primer name	**Primer sequence**	**Expected amplicon size**
chuA.1	5'-GACGAACCAACGGTCAGGAT-3'	279 bp
chuA.2	5'-TGCCGCCAGTACCAAAGACA-3'	
yjaA.1	5'-TGAAGTGTCAGGAGACGCTG-3'	211 bp
yjaA.2	5'-ATGGAGAATGCGAACCTCAAC-3'	
TspE4C2.1	5'-GAGTAATGTCGGGGCATTCA-3'	152 bp
TspE4C2.2	5'-CGCGCCAACAAAGTATTACG-3'	

Table 2.2: Structure of AFLP adaptors (Vos *et al.*, 1995)

Adaptor name	**Sequence structure**
EcoRI-adaptor	5-CTCGTAGACTGCGTACC ATCTGACGCATGGTTAA-5
MseI-adapter	5-GACGATGAGTCCTGAG TACTCAGGACTCAT-5

Table 2.3: The Most Probable Number counts for water samples, resulting from Quanti-tray® 2000 Colilert®

Water sample	**MPN (cfu/100 ml)**
Hartbeesspruit water	198.9
Q01	4.1
Q01 25m	<1
Q02	9.7
Q07	10.9
Q08	14.4
Q09	41.7
Q10	116.9

Table 2.4: The Most Probable Number counts for algal, sediment and water hyacinth samples, resulting from Quanti-tray® 2000 Colilert®

Sample type	**MPN (cfu/100 ml)**
Hartbeesspruit algae	416.0
Hartbeesspruit sediment	9.6×10^4
Jetty algae	6.6×10^3
Q02 Hyacinth	191.8
Q09 algae	55.6

Table 2.5: List of sample names, sample types and sample points of origin from the Roodeplaat Dam

Isolate Name	Sample type	Sample point
Roodeplaat Dam		
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	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

Table 2.5 continued: List of sample names, sample types and sample points of origin from the Roodeplaat Dam

KA1A, KA1B, KA4, KA5, KA6, KA7, KA8A, KA8B, KA9, KA10, KA11, KA12, KA13A, KA13B, KA14, KA15	Algae	Hartbeesspruit

Table 2.6: Results of the number of isolates, of each sample type, belonging to the four phylogroups

	A	**B_1**	**B_2**	**D**	**Unknown**	**Total**
Sewage	12	4	17	11	9	53
Water	5	34	6	9	12	66
Sediment	1	14	1	2	0	18
Algae	1	29	2	8	4	47
Water Hyacinth	0	0	13	0	0	13
Total	19	81	39	30	25	194
Total %	9.80%	41.80%	20.10%	15.50%	12.80%	

Table 2.7: The BLAST search results based on 16SrRNA gene sequence of unknown isolates obtained from some sewage, water and algal sample types

Sample name	**BLAST species**	**% similarity**
ZA1.10	*Citrobacter freundii*	99%
B1.4	*Citrobacter freundii*	99%
B1.6A	*Citrobacter freundii*	99%
B1.8	*Citrobacter freundii*	100%
ZA2.10	*Citrobacter freundii*	100%
B2.9	*Citrobacter freundii*	100%
	Uncultured *Citrobacter* sp	100%
Q022	*Citrobacter freundii*	99%
Q029	*Citrobacter freundii*	99%
Q074	*Enterobacter amnigenus*	99%
Q075	*Citrobacter freundii*	99%
Q081	*Moellerella wisconsensis*	99%
Q09A1	*Citrobacter freundii*	100%
Q09A7	*Citrobacter freundii*	99%
Q101	*Citrobacter freundii*	100%
Q102	*Citrobacter freundii*	99%
Q103	*Citrobacter freundii*	99%
Q104	*Citrobacter freundii*	100%
	Citrobacter braakii	99%
Q108	*Citrobacter freundii*	99%
Q1011	*Citrobacter freundii*	99%
KW2	Uncultured *bacterium*	99%
	Plesiomonas shigelloides	99%

CHAPTER 3

POPULATION STRUCTURE AND ECOLOGY OF *E. COLI* ISOLATED FROM FRESHWATER ENVIRONMENTS IN SOUTH AFRICA

POPULATION STRUCTURE AND ECOLOGY OF *E. COLI* ISOLATED FROM FRESHWATER ENVIRONMENTS IN SOUTH AFRICA

3.1 Abstract

It is well known that *E. coli* is a highly diverse species and due to its large pan-genome, it has the potential to occupy various ecological niches. Multiple studies have reported not only the survival and proliferation of *E. coli* outside of the host, but also the adaptation of naturalised strains results in some level of genetic differentiation. Therefore, the aim of this chapter is to establish the population structure and genetic relatedness of *E. coli* strains obtained from aquatic environments. This was done in order to determine whether these *E. coli* belong to separate populations that are genetically different from their commensal and pathogenic counterparts as a consequence of their isolation and adaptation to the external environment. Isolates collected from the Rietvlei Dam were added to those obtained from the Roodeplaat Dam (chapter 2) giving a total of 293 isolates. *uidA* (β-D-Glucuronidase) and *rpoS* (RNA polymerase sigma factor S, sigma 38) gene sequences were used to determine the relationship between the *E. coli* isolates. Phylogenetic analyses of these genes generated similar clustering patterns and revealed two possible environmental groups. The overall population structure was then determined using the software Structure. All isolates were shown to belong to the same population (K=1). In addition, gene flow and population subdivision were statistically determined using the program dnaSP. Results showed some level of genetic differentiation of the two clusters associated with water hyacinth and an aquatic plant. These two groups also correlated to the two possible environmental clusters observed in the *uidA* and *rpoS* phylogenies. These results support the hypothesis that unique environmental *E. coli* populations do exist and that they are in fact genetically distinct from the rest of the primary population.

3.2 Introduction

Commensal (non-pathogenic) and pathogenic *Escherichia coli* strains are commonly associated with the gastrointestinal tracts of warm-blooded animals. A vast body of studies have concentrated on the pathogenicity of this species, as it is responsible for a variety of diarrhoea-associated illnesses (Lavigne and Blanc-Potard, 2008). These bacteria also spend a considerable part of their life in an environment outside of their primary host (Gordon, 2001). Savageau (1983) suggested that *E. coli* inevitably has two habitats, that is, the gastrointestinal tract of the host and the external environment (water, soil and sediment) forming the primary and secondary environment respectively. Because of its close association with the gastrointestinal tract of humans and animals, it was assumed that *E. coli* does not multiply or survive for long periods in external environments (Burton *et al.*, 1986). The presence of *E. coli* in these other environments is widely believed to be maintained by the constant input of isolates from the gastrointestinal tract of the mammalian host. For this reason, the use of *E. coli* as an indicator of recent faecal contamination was introduced and has remained in use for over 100 years (Winfield and Groisman, 2003).

The evolution and ecology of *E. coli* have recently received much attention, not only because of its importance as pathogen and indicator, but also because of its use as an model organism. These studies have shown that some *E. coli* strains are capable of surviving in soil and water for long periods, even in the absence of any obvious faecal contamination (Solo-Gabriele *et al.*, 2000; Gordon *et al.*, 2002; Power *et al.*, 2005; 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 environmental *E. coli* strains may be genetically different from their commensal and pathogenic counterparts. Therefore, if these environmental strains could be effectively characterised or identified and their ecology and risk to humans better understood, the use of *E. coli* as an indicator organism could be improved.

The detection of genetic differences among individuals within a specific environment is best accomplished using a population genetic approach (Sunnucks, 2000). Such studies can also shed light on the distribution and interaction of genes within a population, which in turn is important for understanding how processes such as natural selection, genetic drift, gene flow and recombination affect the overall evolution and ecology of a species (Hartl and Clark 1997; Spratt and Maiden, 1999; Didelot and Maiden, 2010; Tenaillon *et al.*, 2010; Andam

and Gogarten, 2011). However, bacterial populations rarely conform to the simplicity of some of the models commonly used to describe eukaryotic populations (Spratt and Maiden, 1999; Prosser *et al.*, 2007; Didelot and Maiden, 2010). Although their short generation times allow for the detection of evolutionary changes on feasible time scales (Dobrindt and Chowdary, 2010; Brockhurst *et al.*, 2011), their relative ease of dispersal complicates population genetic inferences (Tenaillon *et al.*, 2010). This is further exacerbated by the partitioning of their genetic information on core and accessory genomes of which the latter is particularly prone to horizontal gene transfer (HGT) (Lawrence and Hendrickson, 2005; Lavigne and Blanc-Potard, 2008; Touchon *et al.*, 2009; Mira *et al.*, 2010). The combined effects of high accessory gene diversity and HGT generally allow (many) bacterial species to occupy a range of different environments resulting in populations that are complex and difficult to define (Nakamura *et al.*, 2004; Touchon *et al.*, 2009; Lukjancenko and Wassenaar, 2010).

The potential 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 differential adaptation to water chemistry and the geological location 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.

Recent genomic studies have revealed that environmental and human isolates of *E. coli* have numerous genes specific to each set of strains (Luo *et al.*, 2011). 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 (Oh *et al*., 2012). These *E. coli*-like strains are thought to originate from true *E. coli* through a type of reductive evolution by losing genes that were no longer needed in a specific niche (Oh *et al*., 2012).

Despite the discovery of unique *E. coli*-like lineages that potentially represent cryptic species of *Escherichia*, a large proportion of the strains obtained from environmental sources represent true *E. coli* (Walk *et al.*, 2009). In a recent study on the diversity of *E. coli* in a South African freshwater body (the Roodeplaat Dam) (Chapter 2 of this dissertation), all strains represented true members of this species, regardless of their source (e.g., sediments, dam walls, aquatic plants, open water). In contrast to initial expectations, considerable genetic variation was observed among the various *E. coli sensu stricto* strains. The main source of *E. coli* in the two aquatic environments (both Roodeplaat and Rietvlei Dams) was believed to be faecal material from humans and animals originating from adjacent sewage treatment works. 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 overall objective of the current study was to examine and characterize the diversity observed in a collection of *E. coli sensu stricto* isolated from freshwater environments in South Africa, by making use of phylogenetic and population genetic approaches. The DNA sequence information for two gene regions (*rpoS* and *uidA*) were used to infer gene trees and to calculate population genetic parameters reflecting gene flow and genetic differentiation. The specific questions addressed were as follows: (i) Are there unique and genetically differentiated subpopulations of *E. coli* in the aquatic environments sampled? (ii) If present, are those unique populations linked to the ecology of their sample site? (iii) Finally, what is the extent of the "connectedness" (*i.e.*, gene flow) among the subpopulations? Being a highly diverse species, understanding its population structure and ecology may improve our understanding of *E. coli* as a species and also shed light on how it evolves and adapts to new environments. In the long term, this study may have an impact on refining the use of *E. coli* as an indicator organism and its role in accurate water quality assessment.

3.3 Materials and methods

3.3.1 *E. coli* isolate collections

In this study, two collections of *E. coli sensu stricto* were used. The one collection included strains that were obtained from the Roodeplaat Dam (Chapter 2; Sections 2.2.1 and 2.2.2). The Roodeplaat Dam is situated on the Pienaars River and is an important water source for the city of Tshwane, especially to the northern areas of Tshwane, namely Doornpoort, Montana, Wonderboom and Magaliesberg. The Dam forms part of the catchment area draining a large part of the City of Tshwane, and houses a water treatment plant. There are also two upstream sewage treatment works (Zeekoegat and Baviaanspoort), which release their treated effluent into the dam.

The second set included isolates from a previous study (Seale, 2010) that were obtained from the Rietvlei Dam, situated in the Rietvlei nature reserve (Pretoria) on the Hennops River. The Rietvlei Dam Water Treatment Works, operated by the Tshwane Metropolitan Municipality, supplies approximately 10 % of the water demand of Pretoria (Bodenstein *et al*., 2006). The dam is also important for recreation as it houses a yacht and canoe club. Being situated in a nature reserve, the water of the Rietvlei Dam is relatively unpolluted with low nitrate-nitrogen levels. In addition, the Dam has limited influx of treated sewage and urban drainage.

E. coli was isolated from samples collected from the banks of the dam in addition to water and sediment samples from varying depths throughout the dam, as described in Chapter 2, Section 2.2.2. Additional *E. coli* isolates were obtained from aquatic plant samples and from the Tshwane drinking water distribution network, which was isolated at the laboratory of the Rietvlei Dam Water Treatment Works. Strains of *E. coli* were also obtained from sewage samples as representative of those strains dominant in the human population. Sewage samples were collected from the inflow and final effluent of the treatment plant that drains into the dam.

3.3.2 *uidA* PCR and sequencing

The *uidA* gene encoding the β-glucuronidase enzyme was previously shown to display greater sequence variation among *E. coli* strains than *rpoS* (Rice *et al.,* 1991, Walk *et al.,* 2009). The *uidA* gene was amplified using the protocol, described by the online MLST database EcMLST (www.shigatox.net), with the following modifications. Each isolate was amplified using the primers uidA-F (5'-CAT TAC GGC AAA GTG TGG GTC AAT-3') and uidA-R (5'- TCA GCG TAA GGG TAA TGC GAG GTA-3') described by Walk *et al.* (2009). Each 25µl reaction volume included 10 X Reaction buffer, 2.5 mM $MgCl_2$, 250 µM of each nucleotide (dATP, dCTP, dGTP and dTTP), 5 pmol of each primer pair, 2.5 U of Taq DNA polymerase (Southern Cross Technologies) and 4 ng/ul of genomic DNA. The PCR cycling conditions consisted of an initial denaturation at 94 °C for 10 minutes followed by 35 cycles of denaturation of 92 °C for 1 minute, primer annealing at 58 °C for 1 minute and extension at 72 °C for 30 seconds followed by a final extension of 72 °C for 5 minutes. These reactions were performed using a Veriti™ Thermal Cycler (Applied Biosystems). PCR products were subjected to agarose gel electrophoresis, previously described above (section 2.3.3) (Sambrook *et al.,* 1989), and visualised using Gel red (Biotium) according to manufacturer's instructions.

For sequencing, all PCR products were purified using 20 U/µl Exonuclease (Fermentas) and 1 U/µl Alkaline phosphatase (Fermentas). The *uidA* amplicons were then sequenced with the forward primer uidA-F by making use of the ABI PRISM® BigDye® v3.1 (Life Technologies), and an ABI 3130 Prism DNA Automated Sequencer (Perkin-Elmer) (See section 2.2.5). The resulting sequences were visualized and corrected where needed with BioEdit Sequence Alignment Editor V 7.0.9.0 (Hall,1999), after which the identity of all sequenced products were verified using similarity searches against the NCBI GenBank database (Altschul *et al.*, 1990; www.blast.ncbi.nlm.nih.gov). Additionally, to identify potential mistakes in the sequences, a six frame translation was performed for each sequence using the online tool provided by Baylor College of Medicine HGSC. The Wise2 (version 2.1.20) (Birney *et al.,* 1996) online tool was used for those isolates that could not be translated in frame. This allowed identification of mistakes in the DNA sequences, which were corrected by referring back to the original chromatograms.

3.3.3 Phylogenetic analyses

For phylogenetic analyses, the *uidA* sequences determined in this study and *rpoS* sequences obtained from Rietvlei Dam (Seale, 2010) and Roodeplaat Dam isolates (Chapter 2, Section 2.2.5) were aligned using MAFFT (Version 6) multiple alignment tool with the FFT-NS-i iterative refinement method (mafft.-cbrc.-jp/-alignment/-server/; Katoh *et al.*, 2002). In addition to these sequences, both data sets also included reference sequences obtained from GenBank. The *rpoS* dataset also included sequences for *E. fergusonii*, *Shigella flexneri* and *S. dysenteriae*, while the *uidA* dataset included sequences for *E. coli* and *S. dysenteriae.*

The aligned datasets were subjected to Maximum Likelihood (ML) analysis (Felsenstein., 1981) in PhyML 3.0 (Guindon *et al.*, 2010) using the best-fit substitution models as indicated by jModeltest (Posada., 2008). The *rpoS* dataset utilized the TPM2 model (Kimura, 1981) with gamma correction to account for site rate variation, while the *uidA* data used the TVMef model (Posada, 2003) with gamma correction and a proportion of invariable sites. Branch support was estimated using non-parametric bootstrap analyses based on 1000 pseudoreplicates under the same model parameters.

3.3.4 Population genetic analyses

To determine whether the collection of *E. coli* has structured into distinct populations, Structure (Version 2.3) (Prichard *et al.*, 2000) was used. Based on genotype and allele frequencies, this software employs a Bayesian model-based clustering method that uses a Marcov Chain Monte Carlo (MCMC) methodology for inferring population structure (Falush *et al.*, 2007). In this study, 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 (Prichard *et al.*, 2000). The admixture model is said to be a flexible model when dealing with the complications of real populations, and is recommended as a good starting point for population analyses (Prichard *et al.*, 2010). These analyses included the aligned sequence information for the *uidA* and *rpoS* genes for all *E. coli* isolates collected from both the Roodeplaat and Rietvlei dams. Each nucleotide of the gene sequence was recognised as an individual locus and the analysis was run assuming the presence of 1 to 20 populations (K=1 to K=20), with a burnin length of 200 000 and a run length of 2 000 000. K was then selected through comparisons of penalized log likelihood scores over independent runs with differing numbers of K-clusters.

DnaSP Version 5.10.01 (Librado and Razos, 2009) was employed to calculate the statistics for gene flow and population subdivision from the aligned *rpoS* and *uidA* gene sequences. For these analyses, test populations were defined based on the geographical location and sample type. Gene flow, as measured by the effective number of migrants per generation (Nm) was estimated from the sequence-based statistic *Fst* described by Hudson *et al.* (1992a). *Fst* reflects the proportion of total genetic variance contained within a population relative to the total genetic variance. An *Fst* value of zero indicates a high level of gene flow, while an *Fst* value of one suggests no gene flow. In addition, the Nm value 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.

Population subdivision was estimated using the sequence-based K_{ST}^{*} statistic, the statistical significance of which was assessed using permutation tests with 10 000 replicates (Hudson *et al.*, 1992b). Based on the number of nucleotide differences within each sequence type, this test determines the likelihood of population differentiation under a null hypothesis of no subdivision. Therefore, high K_{ST}^{*} values with a significant probability (*P*)-values indicate rejection of the null hypothesis and significant levels of populations subdivision (Hudson *et al.*, 1992b), which suggests genetic differentiation where sub-populations may ultimately become fixed and completely isolated.

3.4 Results

3.4.1 *E. coli* isolate collections

A total of 293 strains of *E. coli* were included in this study. Of these, 99 originated from the Rietvlei Dam (Table 3.1), and 194 from the Roodeplaat Dam. Of the Rietvlei Dam strains, 35 were isolated from sediment, 14 from aquatic plants, 16 from both raw and treated sewage, 32 from the dam water and 2 from algal samples. Of the Roodeplaat Dam strains 18 were isolated from sediment, 13 from aquatic plants, 52 from sewage (raw and treated), 66 from the dam water and 45 from algal samples.

3.4.2 Phylogenetic analyses

Phylogenetic relationships among the strains, were inferred using two gene regions (*uidA* and *rpoS*). ML analyses of these two datasets revealed a high level of diversity among the various *E. coli* strains (Figures 3.1 and 3.2). However, numerous strains originating from the two water bodies clustered together. A similar trend was seen for many of the strains from the sewage samples, which grouped with strains isolated from other sites in both water bodies.

Compared to the *uidA* sequences, the *rpoS* sequences were generally less variable, as reflected in the branch lengths in the two gene trees (Figures 3.1 and Figure 3.2). A number of unique *uidA* sequences or haplotypes shared the same *rpoS* sequence type. For example, 109 *uidA* sequence types were observed among the 293 strains examined, with only 71 *rpoS* sequence types found. In addition, compared to the *rpoS* data, analysis of the *uidA* sequences separated the strains into a larger number of defined clusters.

Comparison between the *uidA* and *rpoS* gene trees revealed four common clusters of strains (Figures 3.1 and 3.2). Of the four common groups, clusters 1 and 3 grouped consistently in both gene trees with good bootstrap values supporting their topologies. Cluster 1 consisted of thirteen water hyacinth isolates collected from the Roodeplaat Dam, while cluster 3 included isolates from an aquatic plant in the Rietvlei Dam. The majority of isolates in the *uidA* gene tree are maintained in the *rpoS* gene tree, with the exception of an additional isolate in the *uidA* tree (JA14) and 3 additional isolates in the *rpoS* tree (Q0911, DWWF14G, DWWF28G). The remaining two clusters (clusters 2 and 4) are more variable, consisting of algal, water and sediment isolates. Although these clusters are not maintained across the two gene trees as well as clusters 1 and 3, and are not well supported, they do not include sewage isolates. These four clusters thus potentially represent groups of environmental *E. coli*. Lastly, the *rpoS* gene tree includes an additional possible environmental cluster (cluster 5) consisting of water, sediment and algal isolates. However, it is not maintained in the *uidA* gene tree, nor is it well supported.

3.4.3 Population genetic analyses

To determine whether the collection of *E. coli* strains included in this study represent members of distinct populations, the *rpoS* and *uidA* sequence data were subjected to population analyses with the program Structure (Prichard *et al.,* 2000). During these analyses, allele frequencies were used to probabilistically assign strains to one, two, three, through to 20 (*i.e.*, K=1, 2, 3 … 20) populations (Table 3.2). Based on these analyses, the estimated Ln probability values and the variance of Ln likelihood scores for K=1 were the lowest for both the *rpoS* and *uidA* data. Therefore, these results indicated that all isolates probably represent members of the same population.

To determine whether this population of *E. coli* isolates was subdivided based on geographic origin, sample site or sample type, the *rpoS* and *uidA* datasets were subjected to analyses of population differentiation and gene flow. Highly significant K_{ST}^{*}-values were obtained for almost all data partitions, especially when the *uidA* data were used (see Tables 3.3 and 3.4), which allowed confident rejection of the null hypothesis of no population subdivision. For example, both datasets suggested that the genetic makeup of *E. coli* in the Rietvlei Dam is markedly different from that in the Roodeplaat Dam. This was also true when the strains from water, algae (*uidA* only) and sewage were compared to those in the rest of the collection. The null hypothesis could not be rejected only for two data partitions (*i.e.*, sediment isolates vs. other isolates for both datasets and the algae isolates vs. other isolates for *rpoS* only).

Despite the high level of population differentiation observed within the collection of *E. coli* isolates, gene flow among subdivisions, was wide spread in many cases (Tables 3.3 and 3.4). Gene flow between the collections from the Roodeplaat and Rietvlei Dams for both genes were high (*rpoS* F_{st} = 0.02286; Nm = 21.37 and *uidA* F_{st} = 0.05478; Nm = 8.63). Similarly, relatively high gene flow was also detected between the respective sewage and the rest of strains (*rpoS* F_{st} = 0.02489; Nm = 19.59 and *uidA* F_{st} = 0.02034; Nm = 24.08) and water strains and the rest of strains (*rpoS* F_{st} = 0.05832; Nm = 8.07 and *uidA* F_{st} = 0.02263; Nm = 21.60). This was also true when comparing sediment isolates with the remaining isolates from both the Roodeplaat and Rietvlei Dams (e.g. *rpoS* F_{st} = 0.00873; Nm = 56.79 and *uidA* F_{st} = 0.00739; Nm = 67.14, respectively).

Gene flow only appeared to be limited in the comparisons involving the water hyacinth strains and the aquatic plant strains. There is little gene flow observed between water hyacinth isolates (cluster 1) and the remaining isolates from both dams (*rpoS* F_{st} = 0.66040; Nm = 0.26 and *uidA* F_{st} = 0.71104; Nm = 0.20). Similar results are observed between Rietvlei Dam aquatic plant isolates (cluster 3) and the remaining isolates from both dams (*rpoS* F_{st} = 0.47855; Nm = 0.54 and *uidA* F_{st} = 0.17537; Nm = 2.35). In addition, there is also little or no gene flow between the Roodeplaat Dam water hyacinth isolates and the Rietvlei Dam aquatic plant isolates (*rpoS* F_{st} = 0.71420; Nm = 0.20 and *uidA* F_{st} = 0.82611; Nm = 0.17) indicating that these clusters may be becoming isolated groups.

3.5 Discussion and conclusions

A high diversity of *E. coli* was observed in this study with many of the isolates grouping with isolates assumed to be associated with humans. In addition, environmental isolates were found to be distinct but along with those the human isolates, none clustered with the environmental clades described by Walk *et al,* (2009). This is different from what has been observed for some bacterial populations that conform to a clonal model where there is little or no genetic exchange or recombination and divergence is solely through the accumulation of mutations (Spratt and Maiden, 1999). Lineages within such populations will then arise or fade out as a consequence of selection or other stochastic events (*i.e.,* a random or unpredictable event such as genetic bottle-necking) (Wahl and Gerrish, 2001). In contrast, other populations may undergo frequent genetic exchange via HGT and if individuals are in frequent contact with each other, there may be no restriction on gene exchange between individuals (*i.e.*, gene flow), resulting in a population with little or no structure. The findings of the current study show that *E. coli* fall somewhere between these two extremes, exhibiting some clonal structure due to recent clonal descent disrupted by varying degrees of HGT (Desjardins *et al.,* 1995; Ihssen *et al.,* 2007; Touchon *et al.,* 2009).

The results of this study show that the population of *E. coli* obtained from the freshwater environments examined, is highly diverse. The population was thought to be mainly homogenous as the primary source was treated sewage. *E. coli* is however known to be a highly diverse species, as observed in chapter 2 of this study. Phylogroup, AFLP and *rpoS* gene sequence analyses all revealed a high level of diversity, where the majority of strains could not be separated from those isolated from sewage. The high level of diversity within

the *E. coli* species has been observed in multiple studies. Walk *et al.* (2007) revealed a high level of diversity within *E. coli* isolated from freshwater beaches. Using MLST analysis, they were able to identify 130 sequence types. McLellan (2004) showed that although the level of diversity among *E. coli* isolated from environmental sources is less than that of *E. coli* isolated from host sources, it is still extensive. Furthermore, Byappanahalli *et al.* (2006) demonstrated that soil is a potential habitat for *E. coli* resulting in a high level of diversity within soil-borne *E. coli.* Although, the majority of environmental stains clustered with commensals and pathogens of the gastrointestinal tract, these studies have showed that *E. coli* has the ability to diversify and adapt to soil and freshwater environments (Solo-Gabriele *et al.,* 2000; Power *et al.,* 2005; Byappanahalli *et al.,* 2006; Walk *et al.,* 2007; Ishii *et al.,* 2010). The presence of *E. coli* in various environments outside of the host and its differentiation on a genetic level (Luo *et al.,* 2011, Oh *et al.,* 2012), all support the notion of *E. coli* being a highly diverse species. These findings in turn support the observed diversity and the possible emergence of environmental lineages within the population in this study.

Several genetically distinct subpopulations were found, as the null hypothesis that subpopulations are not genetically distinct was rejected. However, the diversity observed within this collection of *E. coli* strains obtained from the two water bodies, was not homogenously distributed. In fact, the genetic makeup of the strains associated with the water plants were markedly different from the majority of strains, observed specifically in gene flow and genetic differentiation analyses. Significant genetic differentiation was also observed between the isolates from the two dams. These results correlated with those observed in the previous study (chapter 2) where over half of the isolates grouped within environmental phylogroups A and B_1. Such differentiation among different populations is reminiscent of the grouping reported by Hoffmann *et al.* (2001) who observed that 70% of strains isolated from rivers and surface waters in and around Munich, Germany, belonged to groups A and B_1. Walk *et al.* (2007) as well as Power *et al.* (2005) also showed that the majority of strains isolated from freshwater sources belonged to phylogroup B_1. These findings suggest that these groups with distinct genetic makeup possibly represent subpopulations and have undergone some level of niche separation/adaptation.

These results of possible niche separation are supported by those of Gray *et al.* (1999) and Schauer *et al.* (2005). Restricted gene flow and high levels of genetic differentiation indicate that the *E. coli* strains examined in the current study may have adapted to environments associated with water plants. They have thus become both genetically and ecologically different to their gut associated counterparts. These plants may provide a site for attachment and protection in an ever-fluctuating aquatic environment, leading to ecological differentiation and niche separation (Schauer *et al.*, 2005).

The observed diversity within the population of *E. coli* examined might be attributable to various factors. 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 (Whitby *et al.*, 1999; Schauer *et al.*, 2005; Hahn, 2006). Consequently, environmental niche is one of the main driving forces responsible for shaping populations (Schauer *et al.*, 2005). Horizontal gene transfer of beneficial genes within and between species in such an environment also contributes to deviation from the clonal model and possible genetic differentiation in adapting to new niches (Ihssen *et al.*, 2007). This is especially true for a highly diverse species such as *E. coli*, which with its flexible accessory genome, has the potential to obtain specific genes in order to adapt to a various environments (Ihssen *et al.*, 2007, Lukjancenko and Wassenaar, 2010).

Population genetic analysis procedures are usually dependent on the definition of sub-populations (*i.e.*, a priori knowledge is used to partition the strain collection). However, defining the ecological niche of free-living bacteria and accurate definition of subsequent data partitions are often difficult. In a freshwater environment it is expected that there is little or no restriction on the gene flow within the population as there are no physical barriers preventing individuals from encountering each other (Hahn, 2006). To some extent, this was also evident in our data, although the exact location of the sampling site (e.g., Roodeplaat Dam water hyacinth and Rietvlei Dam aquatic plant) may have some effect on gene flow, especially in situations where populations have undergone some level of niche separation. Likewise, 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 cluster of a freshwater bacterioplankton. This cluster is a cluster of filamentous bacterioplankton, which is widely distributed in freshwater systems. Adaptation to various water chemistries together with other abiotic and biotic conditions resulted in this cluster of bacteria adapting to different ecological niches within the freshwater environment. These results may explain the presence of the plant-associated clusters within the *E. coli* population collected in this study. In light of the studies mentioned above, the observed restricted gene flow within the plant-associated clusters in this study could possibly be the start of niche separation within the *E. coli* population.

The restricted level of gene flow and genetic differentiation observed in this study between the plant-associated sub-populations indicates that those isolates may have become adapted to the secondary environment. These strains may even be naturalised without the ability to establish themselves when in contact with the primary hosts. 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. This may be the case with the plant-associated isolates observed in this study. 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 (Lukjancenko and Wassenaar, 2010). 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.

The results presented in this study raise important questions regarding the use of *E. coli* as an indicator organism. *E. coli* clearly possess the ability to survive and proliferate in the secondary environment. This occurrence is observed within this study by the presence of unique environmental clusters, specifically those associated with water plants. The presence

of these environmental groups may complicate the process of determining water quality and lead to incorrect risk assessment of a water body. In addition, it may also be important to understand if these possible environmental groups maintain the ability to circulate through humans, and therefore impact human health. Although, *E. coli* in the secondary environment is an important consideration, this study specifically deals with the *E. coli* population and the structure within.

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 comparisons between those environmental clades and the true *E. coli,* described as originating from humans. Therefore, 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 these environmental isolates group within the true *E. coli*.

3.6 References

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W and Lipman, D. J. (1990) "Basic local alignment search tool." *Journal of. Molecular. Biology*. **215**:403-410.

2. Andam, C. P and Gogarten, J. P. (2011). Biased gene transfer in microbial evolution. *Nature Reviews Microbiology.* **12**: 543-555.

3. Birney, E., Thompson, J. D and Gibson, T. J. (1996). Pairwise and searchwise: Comparison of a protein profile to all three translation frames simultaneously. *Nucleic Acids Research.* **24**: 2730-2739.

4. 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.

5. Brockhurst, M. A., Colegrave, N and Rozen, D. E. (2011). Next-generation sequencing as a tool to study microbial evolution. *Molecular Ecology.* **20**(5): 972-980.

6. Burton, G. A., Gunnison, D and Lanza, G. R. (1986). Survival of pathogenic bacteria in various freshwater sediments. *Applied and Environmental Microbiology* **53**(4): 633-638.

7. 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 Micobiology.* **8**(3): 504-513.

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

9. Desjardins, P., Picard, B., Kaltenbock, B., Elion, J and Denamur, E. (1995). Sex in *Escherichia coli* does not disrupt the clonal structure of the population: evidence from random amplified polymorphic DNA and restriction-fragment-length polymorphism. *Journal of MolecularEvolution*. **41**: 440-448.

10. Didelot, X and Maiden, M. C. J. (2010). Impact of recombination on bacterial evolution. *Trends in Microbiology*. **18**: 315-322.

11. Dobrindt, U and Chowdary, M. G. (2010). Genome dynamics and its impact on evolution of *Escherichia coli. Medical Microbiology and Immunology.* **199**:145-154.

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

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

14. 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.

15. 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.

16. 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 Mocrobiology*. **65**(11): 5089-5099.

17. 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.

18. Hahn, M. W. (2006). Microbial diversity of inland waters. *Current Opinion in Biotechnology*. **17**:256-261.

19. Hall, T.A. (1999). BioEdit: A user- friendly biological sequence alignment editor and analysis program. *Nucleic Acids Symposium Series*. **41**: 95-98.

20. Hartl, D. L and Dykhuizen, D. E. (1984). The population genetics of *Escherichia coli*. *Annual Reviews in Genetics.* **18**: 31-68.

21. Hartl, D. L and Clark, A. G. (1997). *Principles of population genetics*. Vol. 116. Sunderland: Sinauer associates.

22. Hoffmann, H., Hornef, M. W., Schubert, S and Roggenkamp, A. (2001). Distribution of the outer membrane haem receptor protein ChuA in environmental and human isolates of *Escherichia coli*. *International Journal of Medical Microbiology*. **291**:227-230.

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

24. 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.

25. Hudson, R. R. (2000). A new statistic for detecting genetic differentiation. *Genetics*. **155**: 2011-2014.

26. 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.

27. 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.

28. 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.

29. Kimura, M. (1981). Estimation of evolutionary distances between homologous nucleotide sequences. Proceedings of the National Academy of Sciences, USA. **78**: 454-458.

30. 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.

31. Lawrence, J. G and Hendrickson, H. (2005). Genome evolution in bacteria: order beneath the chaos. *Current Opinion in Microbiology*. **8**: 572-578.

32. Librado, P and Razos, J. (2009). DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics: **25**:* 1451-1452.

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

34. 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.

35. 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.

36. Mira, A., Martin-Cuadrado, A. B., D'Auria, G and Rodriguez-Valera, F. (2010). The bacterial pan-genome: a new paradigm in microbiology. *International Microbiology.* **13**:45-57.

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

38. 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.

39. Posada, D. (2003). Using Modeltest and PAUP* to select a model of nucleotide substitution. Pp 6.5.1-6.5.14 in A. D, Baxevanis., D. B. Davison., R. D. M. Page., G. A. Petsko., L. D. Stein and G. D. Stormo, eds.

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

41. 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.

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

43. Prichard, J. K., Wen, X and Falush, D. (2010). Documentation of *structure* software: version 2.3. http://pritch.bsd.uchicago.edu/structure.html.

44. 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**:3 84-392.

45. Rice, E. W., Allen, M. J, Brenner, D. J and Edberf, S. C. (1991). Assay for beta-glucuronidase in species of the genus *Escherichia* and its applications for drinking-water analysis. *Applied and Environmental Microbiology.* **57**(2): 592-593.

46. Sambrook, J., Fritsch, E. F and Maniatis, T. (1989). Molecular cloning: A laboratory manual. Cold Springs Harbour Press. Cold Spring Habour. New York.

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

48. 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.

49. Seale, T. (2010). Unpublished data.

50. 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.

51. 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.

52. Sunnucks, P. (2000). Efficient population markers for population biology. *Trends in Ecology and Evolution.* **15**(5): 199-203.

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

54. Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., Bouvet, O., Calteau, A., Chiapello, H., Clermont, O., Cruveiller, S., Danchin, A., Diard, M., Dossat, C., El Karoui, M., Frapy, E., Garry, L., Ghigo, J. M., Gilles, A. M., Johnson, J., Le Bougue´nec, C., Lescat, M., Mangenot, S., Martinez-Je´hanne, V., Matic, I., Nassif, X., Oztas, S., Petit, M. A., Pichon, C., Rouy, Z., Saint Ruf, C., Schneider, D., Tourret, J., Vacherie, B., Vallenet, D., Me´digue, C., Rocha, E. P. C and Denamur, E. (2009). Organised genome dynamics in the *Escherichia coli* species results un highly diverse adaptive paths. *PLOS Genetics.* **5**(1): e1000344.

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

56. Wahl, L. M and Gerrish, P. J. (2001). The probability that beneficial mutations are lost in populations with periodic bottlenecks. *Evolution.* **55**(12): 2606-2610.

57. 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.

58. 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.

59. Whitby, C. B., Saunders, J. R., Rodriguez, J., Pickup, R. W and McCarthy, A. (1999). Phylogenetic differentiation of two closely related *Nitrosomonas* spp. that inhabit

different environments in an oligotrophic freshwater lake. *Applied and Environmental Microbiology*. **65**(11): 4855-4862.

60. 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.

61. 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.

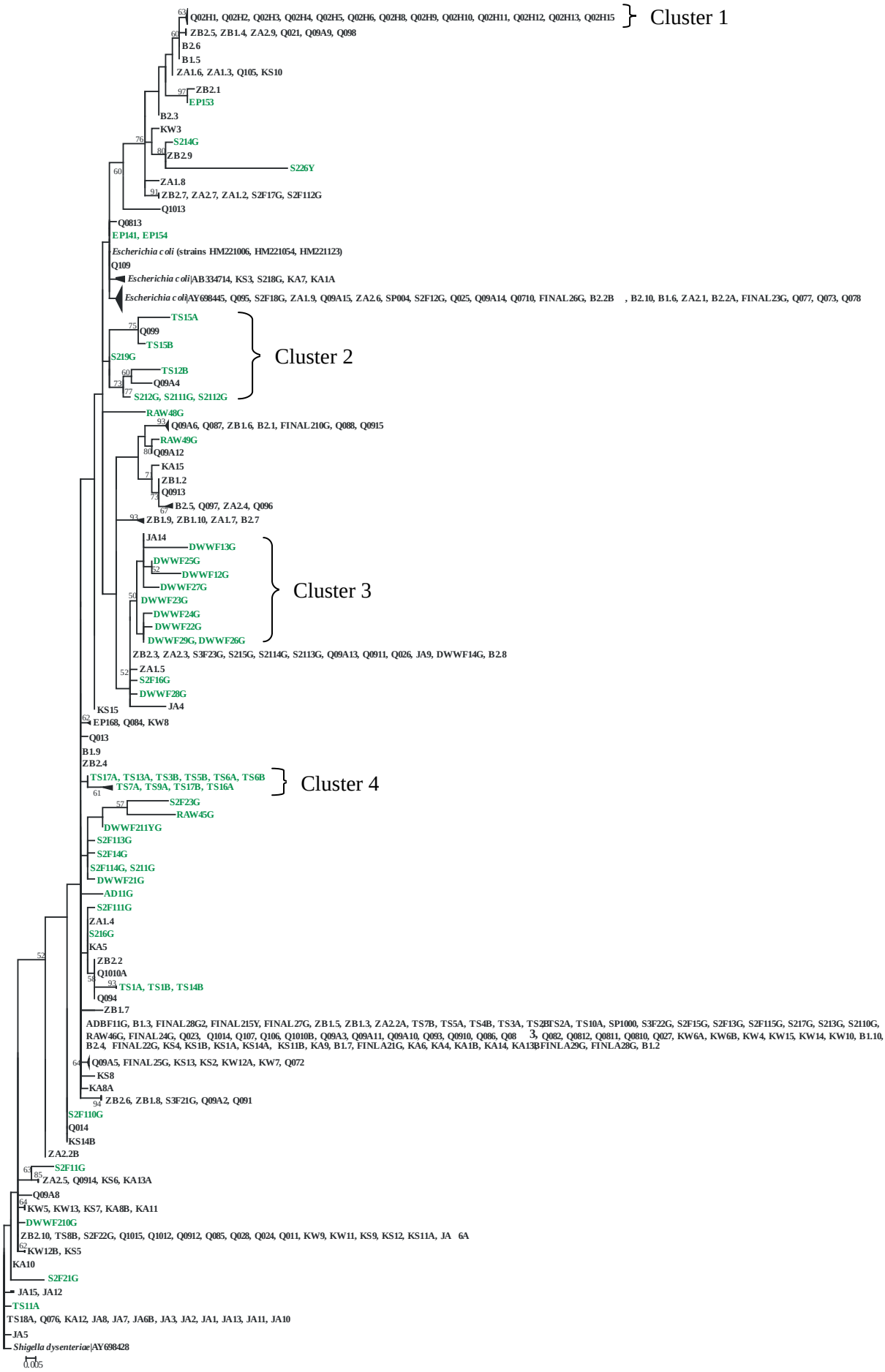

Figure 3.1: Maximum Likelihood phylogenetic tree showing the relationship between *E. coli* isolates, isolated from both the Roodeplaat and Rietvlei Dams. The tree is based on the *uidA* sequence information of all true *E. coli* isolates. Rietvlei Dam isolates are indicated in green. The tree was rooted with *Shigella dysenteriae* as the outgroup and bootstrap analysis of 1000 replicates. Bootstrap values are indicated as percentages and values below 50 were excluded.

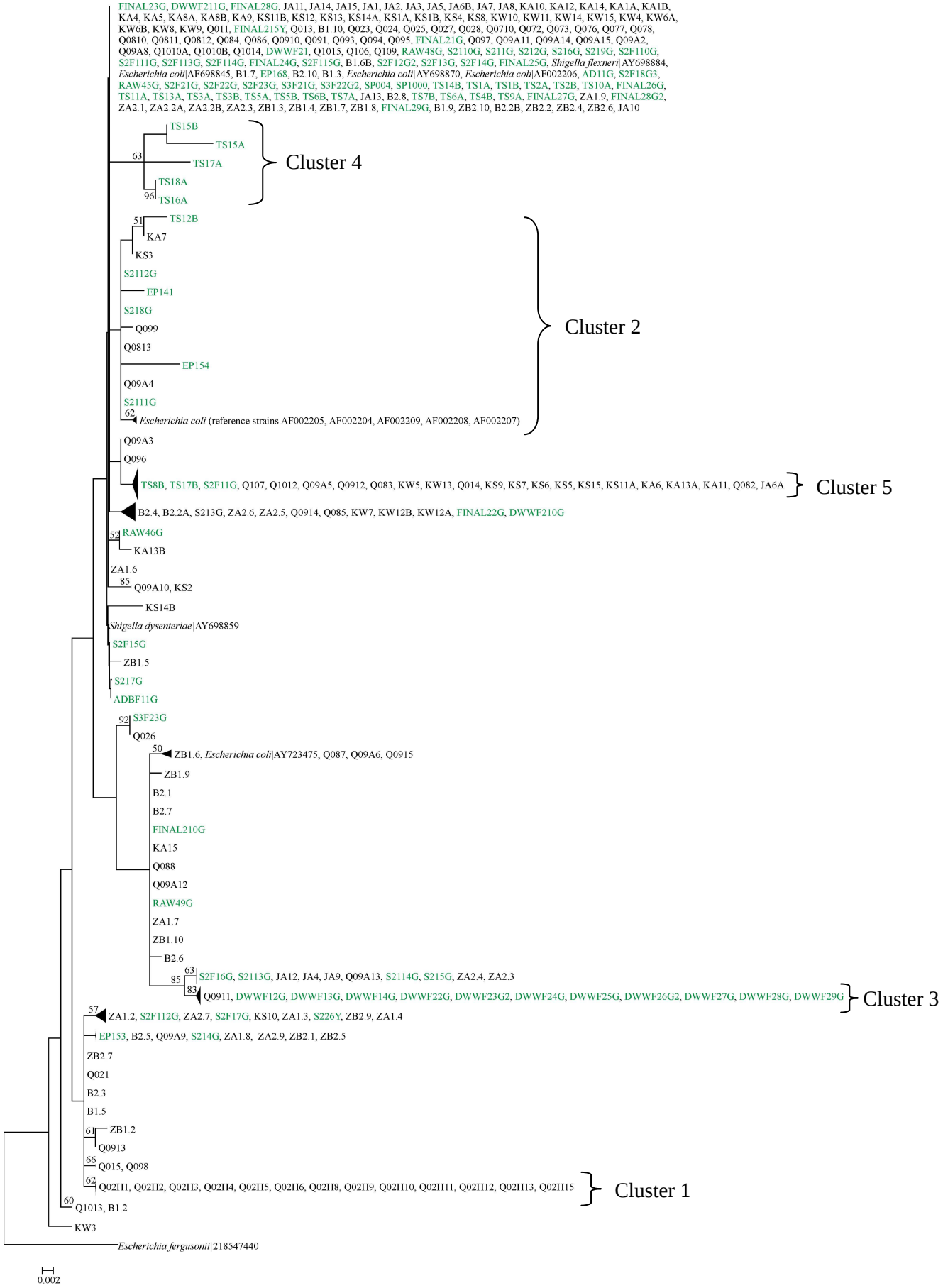

Figure 3.2: Maximum Likelihood phylogenetic tree showing the relationship between *E. coli* isolates, obtained from both the Roodeplaat and Rietvlei Dam. The tree is based on the *rpoS* sequence information of all true *E. coli* isolates. A number of *E. coli* reference sequences were included. Rietvlei Dam isolates are indicated in green. The tree was rooted with *Escherichia fergusonii* as the outgroup and bootstrap analysis of 1000 replicates. Bootstrap values are indicated as percentages and values below 50 were excluded.

Table 3.1: List of sample names, sample types and sample points from the Rietvlei Dam (for Roodeplaat sample names and types included in this chapter, see Table 2.5

Isolate Name	Sample type	Sample point
Rietvlei Dam		
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	Aquatic plant	Rietvlei Dam
RAW42G, RAW45G, RAW46G, RAW48G, RAW49G, FINAL21G, FINAL22G, FINAL23G, FINAL24G, FINAL25G, FINAL26G, FINAL27G, FINAL28G, FINAL29G, FINAL210G, FINAL15Y	Sewage	Hartebeesfontein 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, EP153, EP154, EP168, SP004, SP1000, AD11G, ADBF11G.	Dam water	Rietvlei Dam

Table 3.2: Structure results showing estimated Ln probability of data and the variance of Ln likelihood for K=1 to K=20 for the *rpoS* and *uidA* genes

Number of populations (K)[a]	*rpoS* gene		*uidA* gene	
	Ln P(D)[b]	Var[LnP(D)][c]	Ln P(D)[b]	Var[LnP(D)][c]
1	-2423.2	52.2	-4849.8	65.4
2	-1534.6	110	-3852.6	145.8
3	-1481.9	223.6	-2984	340.6
4	-1357.7	163	-2306.7	242.2
5	-1373.6	260.6	-2208.1	576.6
6	-1150.3	330.3	-2041.4	291.8
7	-1069.1	244.8	-1912.5	250.1
8	-1068.7	253.6	-1944	374.8
9	-1078.1	268.9	-1712.7	382
10	-1079.6	283.6	-1696.2	446.2
11	-1087.5	296.9	-1654.9	370.1
12	-1122.9	359.4	-1926.7	921.5
13	-1185.9	455.3	-2245.7	1649.4
14	-2013.2	2209.9	-1739.2	572.3
15	-1114.9	478.1	-2444.3	2040.8
16	-1111.3	452.1	-2075.8	1258
17	-1069.3	422	-2216.5	1558.3
18	-1144.1	496.5	-1727.9	709.8
19	-1129.8	512.1	-2249	1719.7
20	-1241.1	709.1	-2026.2	1471.9

[a] The number of populations estimated. K=1 to K=20, aiming for the smallest value of K that captures the main structure of the data.

[b] Ln probability of data

[c] Variance of Ln likelihood

Table 3.3: Gene flow and genetic differentiation estimates based on *rpoS* sequence data of isolates from the Roodeplaat and Rietvlei Dams

Populations compared [a]	Gene flow [b]		Genetic differentiation [c]
	F_{ST}	*Nm*	K_{ST}^{*}
All Roodeplaat Dam isolates vs. all Rietvlei Dam isolates	0.02286	21.37	0.00787 (0.0070^{**})
Roodeplaat Dam water hyacinth vs. remaining Roodeplaat and Rietvlei Dam isolates	0.66040	0.26	0.07520 (0.0000^{***})
Rietvlei Dam aquatic plant vs. remaining Roodeplaat and Rietvlei Dam isolates	0.47855	0.54	0.05924 (0.0000^{***})
All sediment isolates vs. remaining Roodeplaat and Rietvlei Dam isolates	0.00873	56.79	0.00139 (0.1860^{ns})
All algae isolates vs. remaining Roodeplaat and Rietvlei Dam isolates	0.01295	38.11	0.00259 (0.0910^{ns})
All sewage isolates vs. remaining Roodeplaat and Rietvlei Dam isolates	0.02489	19.59	0.00702 (0.0120^{**})
All water isolates vs. remaining Roodeplaat and Rietvlei Dam isolates	0.05832	8.07	0.01332 (0.0010^{**})
Rietvlei Dam aquatic plant isolates vs. Roodeplaat Dam Water Hyacinth isolates	0.71420	0.20	0.56207 (0.0000^{**})

[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). Parentheses indicate probability (*P*)-values. High K_{ST}^{*} and low (P)-values indicate rejection of the null hypothesis and significant levels of populations subdivision.

ns = not significant

$^{*}0.01 < P > 0.05$

$^{**}0.001 < P > 0.01$

$^{***}P < 0.001$

Table 3.4: Gene flow and genetic differentiation estimates based on *uidA* sequence data of isolates from the Roodeplaat and Rietvlei Dams

Populations compared [a]	Gene flow [b]		Genetic differentiation [c]
	F_{ST}	*Nm*	K_{ST}^{*}
All Roodeplaat Dam isolates vs. all Rietvlei Dam isolates	0.05478	8.63	0.01720 (0.0000^{***})
Roodeplaat Dam water hyacinth vs. remaining Roodeplaat and Rietvlei Dam isolates	0.71104	0.20	0.06416 (0.0000^{***})
Rietvlei Dam aquatic plant vs. remaining Roodeplaat and Rietvlei Dam isolates	0.17537	2.35	0.01313 (0.0000^{***})
All sediment isolates vs. remaining Roodeplaat and Rietvlei Dam isolates	0.00739	67.14	0.00126 (0.1460^{ns})
All algae isolates vs. remaining Roodeplaat and Rietvlei Dam isolates	0.07637	6.05	0.01373 (0.0000^{***})
All sewage isolates vs. remaining Roodeplaat and Rietvlei Dam isolates	0.02034	24.08	0.00532 (0.0040^{**})
All water isolates vs. remaining Roodeplaat and Rietvlei Dam isolates	0.02263	21.60	0.00837 (0.0010^{**})
Rietvlei Dam aquatic plant isolates vs. Roodeplaat Dam Water Hyacinth isolates	0.82611	0.17	0.51297 (0.0000^{***})

[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). Parentheses indicate probability (*P*)-values. High K_{ST}^{*} and low (P)-values indicate rejection of the null hypothesis and significant levels of populations subdivision.

ns = not significant

$^{*}0.01 < P > 0.05$

$^{**}0.001 < P > 0.01$

$^{***}P < 0.001$

CHAPTER 4

CONCLUSIONS

CONCLUSIONS

This study was based on the growing body of evidence indicating that *E. coli* not only exists, but also multiplies and proliferates in the aquatic environment, outside of the host. In addition, not only are these strains present in the secondary environment but several studies have revealed that they have on some level, become genetically distinct. The presence of *E. coli* in the secondary environment in the absence of faecal material, questions the sustainability of *E. coli* as an indicator organism.

The overall goal of this project was to investigate the presence of possible unique environmental *E. coli* strains in an aquatic environment. Strains were isolated from different niches within two dams within the larger Pretoria area. These isolates were used to determine if they were genetically different to their commensal and pathogenic counterparts. This study also attempted to reveal the genetic diversity and population structure of *E. coli* isolated from this secondary environment.

Initial analyses including phylogrouping, AFLP and *rpoS* gene sequencing, performed on the isolates obtained from the Roodeplaat Dam revealed a high level of diversity. Concerning phylogrouping, approximately half of the isolates belonged to Groups A and B_1. It has previously been demonstrated that environmental isolates are more likely to fall into these two groups. The phylogenetic analysis of the *rpoS* sequences was used to determine whether unique environmental populations were present amongst the isolates. The overall resolution among the *E. coli* isolates increased showing that the majority of isolates formed one large cluster. Amongst the possible environmental clusters, the group of water hyacinth associated isolates had the best support. The analysis was also used to determine whether any strains isolated from the Roodeplaat Dam belonged to the 5 novel clades of *E. coli*. Phylogenetic analysis of the sequence data revealed that all *E. coli* isolates grouped within *E. coli sensu stricto*. It was unclear why strains belonging to the clades were undetected.

To get a better idea of the population structure of *E. coli* in aquatic environments, additional isolates obtained from another impoundment (Rietvlei Dam) were included in the next part of the study. Inferring phylogenetic relationships within a population is difficult with only one gene and therefore the *uidA* gene was also sequenced for all the isolates. Phylogenetic analysis of this gene sequence data revealed that the high level of diversity observed previously. In addition, the majority of the clusters present in the *rpoS* phylogeny were

maintained in the *uidA* phylogeny. Again, possible environmental groups identified in the *rpoS* phylogeny were observed in the *uidA* phylogeny, and the isolates associated with Roodeplaat Dam water hyacinth (Q02H) and Rietvlei Dam aquatic plant (DWW) formed well-supported clusters.

The population structure of the larger set of isolates was then determined in order to get insight into the relationship between isolates. The program Structure revealed that all the isolates belonged to one population (K=1). In addition, isolates collected from the Roodeplaat Dam could not be separated from those collected from the Rietvlei Dam and the majority of environmental isolates could not be separated from sewage isolates. However, this was an interesting result, based on the phylogenetic analysis some population separation would have been expected.

The high level of gene flow between isolates from the different niches was expected, as they were all isolated from an aquatic environment and formed part of a single large population. This was therefore not surprizing that little or no restriction on gene flow was observed between the majority of isolates. However, two groups, the Roodeplaat Dam water hyacinth (Q02H) and the Rietvlei Dam aquatic plant (DWW) isolates, showed little or no gene flow between them or the rest of the population. This result is supported by the two gene phylogenies where both these plant-associated groups consistently grouped together in well supported clusters. This indicated some degree of population subdivision within the larger population suggesting possible niche separation.

These results challenge the idea that the *E. coli* population should be homogenous, based on the assumption that the primary source of *E. coli* into the aquatic environment is believed to be human or animal contamination. The two gene phylogenies together with the population structure analysis revealed that possible unique environmental *E. coli* may exist within the Roodeplaat and Rietvlei aquatic environments. This study supports the idea that genetically distinct populations of naturalised *E. coli* may exist. The results also indicated that the exact location of the sampling site may have some effect on gene flow, although, defining the ecological niche of free-living bacteria in aquatic environments is difficult.

The presence of possible environmental *E. coli* in the secondary environment raises questions about their ability to continue to circulate within the human population as well as whether or not they maintain the genes associated with pathogenicity. Conversely, if they no longer circulate within the human population, how often and at what levels are they detected. Future research to address these questions should involve genome-based studies where the possibility of reductive evolution, can be investigated. Sequencing of environmental *E. coli* strains will aid in answering some of the questions mentioned above. In addition, genome sequences may shed light on possible markers, specific to environmental strains, which would help in differentiating environmental from pathogenic strains when determining water quality. This information will certainly assist in improving the use of *E. coli* in evaluating the safety of water for human use.