Characterization of *Escherichia coli* diversity in a freshwater environment

by

Tarren Seale

Submitted in partial fulfilment of the requirements

for the degree

Master Scientiae (MSc)

In

The Faculty of Natural and Agricultural Sciences

University of Pretoria

Pretoria

Supervisor: Prof. S.N. Venter

Co-supervisor: Prof. E.T. Steenkamp

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Declaration

I, the undersigned, hereby declare that the thesis submitted herewith for the degree **Magister Scientiae** to the University of Pretoria contains my own independent work. This work has not previously been submitted for any other degree at any other University.

_____

Tarren Seale

Characterization of *Escherichia coli* diversity in a freshwater environment

by

Tarren Seale

Supervisor: Prof. S.N. Venter

Co-supervisor: Prof. E.T. Steenkamp

Department: Microbiology and Plant Pathology

Degree: MSc (Microbiology)

SUMMARY

Escherichia coli (*E. coli*) is an indicator of faecal contamination as it is assumed that faecal contamination is the main source of these bacteria in the environment. Recent studies have, however, shown that *E. coli* can be found in the environment without any apparent link to faecal contamination. These environmental *E. coli* isolates multiply and survive in niches including soil, sand, sediment and water. Environmental *E. coli* are usually associated with phylogroups A and B1, two of the 7 phylogroups (A, B1, B2, C, D, E and F) typically used to group *E. coli* isolates. Some environmental isolates have also been linked to Clades III-V, which are novel undescribed *Escherichia* species. In this study *E. coli* was isolated from different niches within various freshwater dams. To represent the *E. coli* circulating in the human population, *E. coli* was isolated from sewage samples. To determine the diversity within the *Escherichia coli* population and to identify possible environmental *E. coli*, the sigma factor S (*rpoS*), β-D-glucuronidase (*uidA*), methyl-directed mismatch repair (*mutS*) and fatty acyl-CoA synthetase (*fadD*) genes were sequenced and Maximum Likelihood trees were drawn using the individual and concatenated datasets. The phylogenetic trees were also used to determine which

phylogroup the isolates are associated with and if any of the isolates belonged to the undescribed species or clades. The population dynamics was determined using TCS and SplitsTree analysis. The phylogenetic trees showed that the diversity amongst these isolates was high. None of the isolates were part of the clades and most of the isolates group with phylogroups A and B1 as expected. Three possible unique clusters of environmental isolates were observed which respectively formed part of phylogroups B2, D and no specific phylogroup. Phylogroups B2 and D are usually associated with isolates that cause extra-intestinal infection and was not expected to be represented by environmental isolates. Population structure analyses indicated that these clusters could be part of sub-populations within the larger *E. coli* population and may be genetically separate from the rest of the isolates.

ACKNOWLEDGEMENTS

- I would like to thank God for being there for me not only during this degree, but every step of my life and for listening to my prayers.
- Thank you to my mom, Michelle Seale for being the strongest person I know and for being my example. Thank you for all the support and love during my studies.
- Thank you to my brother, Darren Seale and his family Gerda, Cameron and Marcel for their love and words of encouragement.
- Thank you to my uncle Mika Moerat and his family Rooshaan, Tashriq and Mya for all their help during the tough times and for helping me pursue my dreams.
- To my supervisors Prof. S.N. Venter and Prof. E.T. Steenkamp, thank you for all the discussions and the meetings we had about my work. Thank you for the opportunity and for believing in me.
- To my lab mates from lab 9-35 (Gaby, Zander, Gina, Eric, Pieter, Divine, Palesa, Solize, Barry and Sisanda) and lab 9-40 (Annie, Chrizelle, Karabo and Juanita) thank you for all the help with protocols and chemicals. Thank you for all the brainstorming sessions.
- Thank you to Sarah MacRae for all the help and technical support, for the isolates and the critical discussions.
- Thank you to the people that helped me with sample collection as well as sending me isolates during my degree.
- The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.
- Thank you to the Water Research Commission (WRC) and the University of Pretoria (UP) for financial support during this project. The opinions expressed are those of the author and are not attributed to the funding bodies.
- This work is based on the research supported by the National Research Foundation (NRF) of South Africa.

TABLE OF CONTENTS

Chapter 2

***Escherichia coli* diversity in subtropical catchments**

Chapter 3

Population dynamics of *Escherichia coli*

LIST OF TABLES

CHAPTER 2

CHAPTER 3

LIST OF FIGURES

CHAPTER 2

CHAPTER 3

PREFACE

Escherichia coli (*E. coli*) is widely used as an indicator of recent faecal contamination as it is believed that this bacterium is primarily associated with the gastro-intestinal tract of warm-blooded animals and humans . The presence of *E. coli* is therefore linked to the possibility that other intestinal pathogens could potentially also be present in the tested sample (Ashbolt et al. 2001). The use of *E. coli* as an indicator is based on the assumption that *E. coli* cannot survive for long periods in the environment outside the primary host (Byappanahalli and Fujioka, 2004). A number of studies have, however, shown that *E. coli* can multiply and survive for long periods in the environment (Byappanahalli and Fujioka, 2004).

Initially studies focused on the presence and survival of *E. coli* in the environment for long periods. These studies found that *E. coli* could regularly be isolated from the environment and that *E. coli* can survive for extended periods at environmental temperatures (Byappanahalli and Fujioka, 2004, Ishii et al. 2006). This work was followed by studies showing that *E. coli* could also be isolated from soil, algae, sediment, water and plants (Fisher et al. 1998, Whitman and Nevers, 2003, Byappanahalli and Fujioka, 2004, Méric et al. 2013). Several of these *E. coli* isolates associated with the secondary environment were part of phylogroup B1 (Walk et al. 2007). Other studies reported that some environmental isolates grouped separately from the faecal isolates in a phylogenetic tree (Whittam, 1989, Byappanahalli et al. 2006) and a number even formed separate clades representing undescribed species within the genus *Escherichia* (Walk et al. 2009).

The fact that *E. coli* can multiply and exist outside the host, questions the use of *E. coli* as an indicator organism. The implications of environmental *E. coli* can, however, not be addressed without a clear understanding of how the population of *E. coli* in the environment is structured and how isolates relate to the *E. coli*

associated with the intestines of mammals. Although several studies has addressed some aspects the main question that still needs to be answered is whether separate and genetically distinct environmental *E. coli* populations exist in aquatic habitats.

To address this question, *E. coli* was isolated from different environmental sources during the initial phase of the study. The environments targeted included water plants (to focus on the association of *E. coli* with plants) as well as other presumably distinct freshwater niches using MLGA (Membrane Lactose Glucuronide Agar). *E. coli* was also isolated from sewerage samples to represent *E. coli* circulating within the human population. The *rpoS* (sigma factor) and *uidA* (β-D-glucuronidase) genes were sequenced for all of the *E. coli* isolates. The gene sequences were compared using phylogenetic analysis, firstly to determine if any of the *E. coli* isolates form part of the separate clades observed by Walk et al. (2009) and secondly the phylogenetic trees were analysed to determine whether any genetically separate clusters that contained only environmental isolates were observed.

During the second part of the study two additional genes, the *mutS* (methyl-directed mismatch repair) and *fadD* (fatty-acyl CoA synthetase) genes were sequenced for each of the isolates. This was to determine if the unique clusters found based on the *rpoS* and *uidA* phylogenetic analyses remained intact when more variable genes were examined. The more variable genes could also allow for the elucidation of new unique clusters. This data was also used to investigate the overall population structure of the isolates and how different clusters relate to each other. This was done by analyzing the concatenated gene sequence dataset using phylogenetic, SplitsTree and TCS analyses. All this data was then used to develop a better understanding of the diversity and population structure of *E. coli* in aquatic environments.

REFERENCES

Ashbolt, N.J., Grabow, W.O.K. and Snozzi, M. (2001) Indicators of microbial water quality. In: Fewtrell L, Bartram J, eds. *Water quality: Guidelines, standards and health. Assessment of risk and risk management for water-related infectious disease*. WHO Water Series. London, IWA Publishing, pp. 289–315.

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* Volume 50 Number 1 pp. 27–32.

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

Fisher, M.M., Wilcox, L.W. and Graham, L.E. (1998) Molecular characterization of epiphytic bacterial communities on Charophycean green algae. *Applied and Environmental Microbiology* Volume 64 Number 11 pp. 4384-4389.

Ishii, S., Ksoll, W.B., Hicks, R.E. and Sadowsky, M.J. (2006) Presence and growth of naturalized *Escherichia coli* in temperate soils from Lake Superior Watersheds. *Applied and Environmental Microbiology* Volume 72 Number 1 pp. 612–621.

Méric, G., Kemsley, E.K., Falush, D., Saggersm E.J. and Lucchini, S. (2013) Phylogenetic distribution of traits associated with plant colonization in *Escherichia coli. Environmental Microbiology* Volume 15 Number 2 pp. 487-501.

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

Walk, S.T., Alm, E.W., Gordon, D.M., Ram, J.R., Toranzos, G. A., Tiedje, J.M. and Whittam, T. S. (2009) Cryptic Lineages of the Genus *Escherichia. Applied and Environmental Microbiology* Volume 75 Number 20 pp. 6534–6544.

Whitman, R.L. and Nevers, M.B. (2003) Foreshore sand as a source of *Escherichia coli* in nearshore water of a Lake Michigan Beach. *Applied and Environmental Microbiology* Volume 69 Number 9 pp. 5555–5562.

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

CHAPTER 1

SURVIVAL OF *ESCHERICHIA COLI* IN DIVERSE ENVIRONMENTS AND GENETIC CHARACTERIZATIONS OF THESE POPULATIONS

SURVIVAL OF *ESCHERICHIA COLI* IN DIVERSE ENVIRONMENTS AND GENETIC CHARACTERIZATIONS OF THESE POPULATIONS

1.1 *ESCHERICHIA COLI*: DIVERSITY AND ECOLOGY

1.1.1 Taxonomic status of *Escherichia coli*

Escherichia coli (*E. coli*) forms part of the genus *Escherichia* in the family *Enterobacteriaceae.* Although *E. coli* and the *Shigella* species are valid and separately described species, research strongly support the notion that *Shigella* species are only unique pathotypes and should be considered as belonging to *E. coli* (Alm et al. 2011). The *Escherichia* genus is phylogenetically not yet well defined. Currently the species included in this genus are *E. hermannii, E. fergusonii, E. vulneris and E. albertii* (Welch, 2006, Alm et al. 2011). *E. blattae* and *E. adecarboxylata* were recently moved to the genera *Shimwellia* and *Leclercia,* respectively, (Gyllenberg et al. 1997, Priest and Barker, 2010) and additional species within the genus may still need to be re-classified (Welch, 2006, Alm et al. 2011). Furthermore, Walk et al. believes that within the genus *Escherichia,* there are five cryptic clades (CI-CV) (Walk et al. 2009). These are clusters of isolates that cannot be phenotypically distinguished from known *E. coli sensu stricto*, however they form genetically distinct clusters (Walk et al. 2009) (see below).

1.1.2 Occurrence of *Escherichia coli*

E. coli can occupy a number of different habitats and are traditionally known to include commensal strains that cause no disease in their primary host (warm-

blooded animals and humans) as well as pathogenic strains responsible for gastroenteritis, sepsis and urinary tract infections (Lavigne and Blanc-Potard, 2008). The strains of *E. coli* that are pathogenic and cause diarrhea include enterohemorrhagic *E. coli* (EHEC). This virotype is associated with bloody diarrhea and secretes a toxin that causes hemolytic-uremic syndrome which affects the kidneys (Rasko et al. 2008, Leimbach et al. 2013). Another virotype, enterotoxigenic *E. coli* (ETEC) produces heat-labile and heat stable toxins that disturb the water balance in the gut resulting in watery diarrhea (Rasko et al. 2008, Leimbach et al. 2013). Enteroinvasive *E. coli* (EIEC) invades the colon via the enterocytes (Rasko et al. 2008, Leimbach et al. 2013). Enteropathogenic *E. coli* (EPEC) uses the Type 3 secretion system (T3SS) to colonize the small intestine and enteroaggregative *E. coli* (EAEC) causes disease by biofilm formation and subsequent toxin and cytolitic factor production (Rasko et al. 2008, Leimbach et al. 2013). These strains are usually transferred by the faecal-oral route through the consumption of contaminated food and water and may in severe cases result in the death of the patient (Walk et al. 2009).

1.1.3 *Escherichia coli* genome structure

E. coli has a dynamic or mosaic genome structure (Welch et al. 2002, Leimbach et al. 2013). The overall genome for the species consists of a core genome and a large accessory genome that together contribute to an extensive pangenome (Rasko et al. 2008). The core genome constitutes the genes that are shared by all *E. coli* strains, while the accessory genome consists of genes and elements that are unique to a specific *E. coli* strain or group of strains (Rasko et al. 2008). This packaging of the genetic material of the bacterium is linked to its ability to adapt to diverse habitats. When studying 17 *E. coli* genomes, Rasko et al. found that amongst the estimated 5000 genes found, the core genome of *E. coli* constitutes an average of 2 200 genes. The remaining accessory genes codes for genes required for niche adaptation and include a number of pathogenicity genes in the case of the virotypes (Rasko et al. 2008).

These authors also showed that the genome of *E. coli* is continuously evolving and diversifying because the addition of the genome sequence for a new strain typically increases the known pangenome of the bacterium (Rasko et al. 2008).

In *E. coli*, pathogenicity elements are expected to be located on the accessory genome as they are usually acquired by horizontal gene transfer (Leimbach et al. 2013). These elements mediate many functions such as replication of the bacterium within the host, attachment to the host and break down of the host defences (Lavigne and Blanc-Potard, 2008). These functions are facilitated by effectors which include toxins, adhesins, iron uptake systems and proteases (Lavigne and Blanc-Potard, 2008, Leimbach et al, 2013). The T3SS is also associated with these elements and plays an important role in pathogenicity by transferring effectors from the bacterium straight into host cells (Lavigne and Blanc-Potard, 2008).

1.1.4 *Escherichia coli* in other primary hosts

Domestic animals

Commensal *E. coli* is associated with domestic animals (Beutin et al. 1993). They have been isolated from animals when studying drug resistant *E. coli* (Bywater et al. 2004, Khachatryan et al. 2004). However, the presence of *E. coli* virotypes associated with domestic animals has also been shown. Beutin et al. (1993) studied seven domestic animals: cattle, sheep, goats, chickens, dogs, cats and pigs to demonstrate that verotoxin (VT) producing *E. coli* was present in these animals. The VT producing *E. coli* is pathogenic to humans and can cause hemolytic uremic syndrome (HUS) (Beutin et al. 1993, Wells et al. 1991). VT producing *E. coli* also appears to be more regularly associated with ruminants than non-ruminants (Beutin et al. 1993, Wells et al. 1991).

Birds

E. coli can easily be isolated from bird guano due to the fact that it is naturally found in the intestine of birds (Whittam, 1989). During a study on *E. coli* associated with birds, Whittam isolated *E. coli* from the gastro-intestinal tract of birds as well as the secondary environment associated with these birds. The secondary environment included soil, water and the litter of the birds (Whittam, 1989). Among the more than 400 isolates obtained from birds and their environment, the authors found 113 clones. These clones further appeared to be separated into two distinct subpopulations corresponding to their environmental and gastro-intestinal origins (Whittam 1989).

The *E. coli* associated with birds can also be pathogenic to birds (Vidotto et al. 1990). For example, verotoxin-producing *E. coli* has been isolated from wild birds (Wallace et al 1997). Strains possessing the Col V plasmid and that cause colisepticemia have also been reported from birds (Dho-Moulin and Fairbrother, 1999). These strains are known as APEC (Avian pathogenic *E.* coli) (Dho-Moulin and Fairbrother, 1999). The Col V plasmid encodes products needed for the production of colicin V that is involved in the inhibition of growth of related or the same bacterial species, as well as products needed for invasion and pathogenicity (Dho-Moulin and Fairbrother, 1999, Vidotto et al. 1990).

Reptiles

E. coli is not only associated with warm-blooded animals and birds, but also with reptiles (Gordon and Cowling, 2003). The bacterium has been isolated from turtles, crocodiles, snakes and some lizard species (Gordon and Cowling, 2003). In a study conducted by Waturangi et al. (2003) 28 *E. coli* strains were isolated from the faeces of monitor lizards. These *E. coli* isolates were screened for resistance to antimicrobial agents and the authors found that they were similar to those from the human host in terms of their resistance to multiple antimicrobials agents (Waturangi et al. 2003). For example, many of the isolates from the lizards were resistant to tetracycline and kanamycin and

some isolates were even resistant to multiple antimicrobial agents (Waturangi et al. 2003).

1.1.5 Occurrence of *Escherichia coli* in secondary environments

The gastro-intestinal tract of humans and warm-blooded animals is widely regarded as the primary habitat of *E. coli*. Based on this assumption it is thought that the faeces of the primary host is the only source of so-called environmental *E. coli* and that the bacterium cannot multiply or survive outside the host (Byappanahalli and Fujioka, 2004). The reason for this is believed to be the distinctness of the secondary environments (Savageau, 1983) in terms of temperature (secondary environment colder than the primary) and nutrients (availability of carbohydrates and amino acids) (Savageau, 1983, Whittam, 1989). Whittam originally estimated that the half-life of human associated *E. coli* in the secondary environment is only a few days (Whittam, 1989). Subsequent studies have however, shown that certain strains of *E. coli* can persist and multiply in the environment (see below).

Soil and sand

Initial studies of the secondary environment mainly focused on tropical soils where *E. coli* was detected in the absence of any faecal contamination (Byappanahalli and Fujioka, 2004). Later, Ishii et al. (2006) reported that there are naturalized *E. coli* in northern temperate soils of three Lake Superior Watersheds. They demonstrated that these *E. coli* could survive for a long time in the soil and could grow to high densities under laboratory conditions when incubated at 30 or 37°C (Ishii et al. 2006). When incubated at temperatures equal or lower than 25°C these strains survived in the soil for up to a month, proving that these *E. coli* strains grow and persist in the soil (Ishii et al. 2006). These studies clearly demonstrated that naturalized *E. coli* strains are not only found in tropical soils as originally thought but also in temperate soils.

A similar study conducted by Whitman and Nevers (2003) tested for the presence of *E. coli* in the foreshore sand of a Lake Michigan beach. They determined that the sand acted as a suitable habitat for environmental *E. coli*. They also showed that the presence of *E. coli* in the sand resulted in an increase of the faecal indicator bacteria in the lake water. Byappanahalli et al. (2006a) expanded the work to study backshore sand to compare it with the findings related to the foreshore sand. They found that this part of the beach was also a habitat for *E. coli* and that it may even contain pathogens that could pose a health risk (Byappanahalli et al. 2006a). Other researchers found similar results showing that *E. coli* from the sand and soil increase the *E. coli* numbers in water (Desmarais et al. 2002, Alm et al. 2003, Anderson et al. 2005, Ishii et al. 2007).

A study to determine the genetic diversity and distribution of the environmental *E. coli* as well as to assess their ability to persist in soil environments was conducted by Byappanahalli et al. (2006b). They concluded that *E. coli* could persist in the soil for a long time even though they were only found at low concentrations (Byappanahalli et al. 2006b) They also found that although a diversity of *E. coli* isolates was present in the soil, they still formed a cohesive phylogenetic cluster when compared with faecal isolates (Byappanahalli et al. 2006b).

Algae

The occurrence of bacteria on algae is well known (Fisher et al. 1998). Whitman et al. (2003) showed that *E. coli* can be associated with and persist on *Cladophora*, which was the first report of the association between faecal indicator organisms and *Cladophora.* They found that the attached algae supported higher levels of *E. coli* than the unattached algae. However, the amount of sunlight, the water temperature and the thickness of the algal mats also affected the *E. coli* levels (the higher these factors, the higher the *E. coli* level). When the dried *Cladophora* mats were rehydrated, the *E. coli*

concentration increased by about 4 logs, even 96 hours after rehydration (Whitman et al. 2003). This study thus demonstrated that *E. coli* can multiply and persist in association with *Cladophora* (Whitman et al. 2003).

Byappanahalli et al. (2003) examined the genetic relatedness of the *E. coli* associated with different *Cladophora* mats. They found that the isolates were genetically related but highly diverse and could be divided into two clusters, although the isolates did not group according to a specific *Cladophora* mat. The authors interpreted these patterns as evidence for dynamism in the *E. coli* population and that different *E. coli* strains from water can colonize the *Cladophora* mats (Byappanahalli et al. 2003).

Plants

The plant surface is an environment that experience radical changes in moisture, temperature and UV rays, yet plants are still highly colonized by many bacteria and fungi (Lindow and Leveau, 2002, Lindow and Brandl, 2003). The common microorganisms are typical epiphytes (Lindow and Brandl, 2003), although *E. coli* and other pathogenic *Enterobacteriaceae* have been found to occur on or in agricultural crops (e.g., *Pectobacterium*, *Brenneria* and *Pantoea*), oranges, radish sprouts, as well as salads and fruit juice (Brandl, 2006, Cooley et al. 2006, Holden et al. 2008, Caponigro et al. 2010). The presence of these bacteria on plants is usually attributed to cross-contamination by water or contaminated meat (Brandl, 2006, Holden et al. 2008). Contamination has also been linked to the hands of the crop handlers and the equipment used during packaging (Holden et al. 2008).

Solomon et al. (2002) showed that *E. coli* O157:H7 cannot only be isolated from crops such as lettuce, but that the bacterium can also grow on this plant substrate. In the same study by Solomon et al. they showed that this *E. coli* strain could also move through the plant roots if the soil or water is contaminated with faecal matter (Solomon et al. 2002). These findings indicated

that pathogenic *E. coli* can attach to plant material (Solomon et al. 2002, Cooley et al. 2006) and suggested that the bacterium might gain entry into the plant under certain conditions. In fact, recent studies have demonstrated that several *E. coli* strains not only attach to plant surfaces but could also colonize and gain entry to these plants (Holden et al. 2008, Méric et al. 2013). The interior of the plant is a better environment for bacterial survival as it provides both protection and nutrients (Méric et al. 2013). In this process *E. coli* cells use their filamentous structures to attach to plant surfaces (van Elsas et al. 2011) and to enter the plant.

1.1.6 Survival and growth in the secondary environment

There are a number of factors affecting the survival and growth of *E. coli* in the secondary environment. Here, survival of *E. coli* is dependent on its resistance to starvation (van Elsas et al. 2011), because the availability of suitable carbon sources could be limited. Other factors that impact on survival and growth of *E. coli* in the secondary environment include temperature, pH, the availability of water and the microbial community already present in the secondary environment, and obviously the relative importance of these factors will differ, depending on the type of environment.

Soil

Byappanahalli and Fujioka (2004) investigated the effect of other microorganisms present in the soil and the moisture of the soil on the growth of *E. coli*. The hypothesis was that the indigenous soil microorganisms would use the nutrients and limit the growth of the faecal indicator bacteria and that low soil moisture has a similar effect (Byappanahalli and Fujioka, 2004). This study confirmed that the *E. coli* would grow faster and better if more nutrients were available but that its growth under natural conditions may be sporadic and restricted by the presence of other more suitably adapted bacteria

(Byappanahalli and Fujioka, 2004). Brennan et al. (2010) however, found that even in oligotrophic and low temperature conditions some *E. coli* strains have the ability to grow well. These findings thus supported some earlier work reporting that *E. coli* could grow in the soil and could subsequently increase the levels of the bacterium in the water that comes into contact with such soils (Solo-Gabriele et al. 2000).

Water

Survival of *E. coli* in the water is determined by, amongst other environmental factors, the temperature of the water (Rhodes and Kator, 1988, Sampson et al. 2006). *E. coli* blooms, when the bacterium occurs in numbers exceeding 10 000 cells per 100ml, were reported for two Australian lakes (Power et al. 2005). These blooms tend to occur from mid-summer to early autumn when the water temperature is higher than 18°C. The dominant *E. coli* found in these waters were encapsulated and could not be linked to faecal contamination (Power et al. 2005).

The persistence of *E. coli* in water is influenced by the presence of clay particles and sand for the bacterium to adhere to (Sampson et al. 2006). The salt concentration of the water and the level of light that can penetrate the water body will also determine the ability of E. coli to persist (Bordalo et al. 2002). The presence of other microorganisms likely also influence its persistence, as the reduction in *E. coli* levels has been linked to competition with other microorganisms in the water (Rhodes and Kator, 1988).

Algae

In their study Byappanahalli et al. (2003) investigated whether *Cladophora* supported the growth of faecal indicator organisms like *E. coli.* They showed that the growth of *E. coli* in an algal leachate growth medium was best at 35°C but still significant at 25°C (Byappanahalli et al. 2003). The 25°C temperature

corresponded with environmental temperatures, indicating that these *E. coli* strains can exist and multiply at typical environmental temperatures (Byappanahalli et al. 2003). In addition to the organic substances provided by the algae, it is suggested that they can also provide a protected environment for the growth of bacteria (Byappanahalli et al. 2003).

Plants

Méric and colleagues examined 106 plant isolates, which they compared phenotypically to isolates of faecal origin. They were interested in studying the potential traits that differentiate the plant associated strains and that allow for plant colonization (Méric et al. 2013). Their found an increased ability for biofilm formation and production of extracellular matrix (EM) components amongst the plant associated isolates (Méric et al. 2013). They argued that these factors are potentially important for plant colonization. Biofilm formation ability was also enhanced at lower temperatures, which reinforced the notion that increased biofilm formation is most likely an environmental adaptation (Méric et al. 2013). Méric et al. (2013) also showed that the plant associated isolates were different in terms of carbon utilization. The plant associated isolates appeared to be able to utilise fewer carbon sources than the faecal isolates (Méric et al. 2013).

1.1.7 Genetics underpinning persistence in the secondary environment

The survival of *E. coli* in the secondary environment is well recorded, but the mechanisms underpinning the shift to the secondary environment is not well understood. The two environments are very different and there are two hypotheses for how *E. coli* persists in the secondary environment. The first hypothesis is that *E. coli* deals with the transition from the primary to the secondary environment and vice versa by dual regulation (Savageau et al. 1983). If persistence of *E. coli* is due to a dual regulation system, certain genes

which encode products that are necessary for survival will be up-regulated, while the production of other proteins not required in this environment will be down-regulated (Savageau et al. 1983, Whittam, 1989, Gordon et al. 2002). In this scenario the genomes of environmental and clinical strains will not differ greatly as they would need the whole range of genes to be able to survive both types of environments.

The second hypothesis states that specific genotypes are better adapted to the host environment while others are better adapted to the secondary environment (Whittam, 1989, Gordon et al. 2002).These genotypes will always be present and dominate in the specific environment they have been adapted to (Whittam, 1989). This type of adaptation or naturalization will mainly be driven by the dynamic genome structure of *E. coli*, allowing genes that will be required in a specific environment to form part of the accessory genome. This is consistent with the results of a comparative genomic study (Oh et al. 2102), which demonstrated clear differences between isolates associated with humans and those associated with non-humans (Oh et al. 2012). Unique genes were present within the genomes of the environmental isolates that could be used to distinguish environmental from faecal isolates. However, the environmental strains examined by Oh et al. (2102) all belonged to the unique clades or cryptic species of *Escherichia*. The *E. coli sensu stricto* isolates associated with the secondary environment belong to phylogroups B1 and A (Orsi et al. 2007). These phylogroups seem to have adapted for survival in the secondary environment (see below).

1.2 POPULATION STRUCTURE AND GENETIC DIVERSITY OF *ESCHERICHIA COLI*

1.2.1 Typing schemes

Molecular typing methods are important tools for studying pathogenic strains and are useful in understanding and controlling the spread of disease (Tenover

et al. 1995, Maiden et al. 1998). Typing methods detect nucleotide and phenotypic variation within strains (Selander et al. 1986, Smith et al. 2000). It can be used for epidemiological studies and can determine how related an isolate is to other isolates (Maiden et al. 1998). However, the level of discrimination is an important consideration when selecting a typing method; the method must distinguish between different species, but it must also be able to distinguish between isolates of the same species (van Belkum et al. 2001).

Serotyping

Serotyping is used to characterize *E. coli* based on cell surface elements that elicit an immune response (Ballmer et al. 2007). There are three elements used for serotyping. These are the O-antigen (core of lipopolysaccharide layer), K-antigen (capsule of *E. coli*) and H-antigen (Flagellar protein) (Ballmer et al. 2007). Serotyping is performed by agglutination; the bacteria with the unknown serotype are exposed to the antibody. Different antigens are used separately and if agglutination occurs, it is recorded as a positive result (Ballmer et al. 2007). Certain serogroups are consistently associated with the same clinical symptoms and diseases (Ballmer et al. 2007) and serotyping has been widely used to study the epidemiology of *E. coli* (Ballmer et al. 2007). However, this method is very costly, and alternative approaches such as the use of PCR-RFLP (PCR-restriction fragment length polymorphism) have been developed (Ballmer et al. 2007).

Multilocus enzyme electrophoresis (MLEE)

Multilocus enzyme electrophoresis (MLEE) utilizes the electrophoretic mobility of a number of water-soluble housekeeping cellular enzymes to differentiate between strains (Tenaillon et al. 2010). The difference in movement of the enzyme relates to the different alleles that are present at the targeted locus (Tenaillon et al. 2010). This method has been one of the first used to provide insight into the structure of *E. coli* populations (Selander et al. 1986).

Multilocus sequence typing (MLST)

Multilocus sequence typing (MLST) involves analysis of the DNA sequence of a number of housekeeping genes, usually in the order of 7 genes (Tenaillon et al. 2010). MLST can be used on pathogens as well as non-pathogenic strains and can in some cases be used to distinguish between them (Maiden et al. 1998, Smith et al. 2000). An MLST scheme can be extended to include more genes, in a study conducted by Walk et al (2009) 22 *Escherichia* genes were included. The sequences generated are used to either draw phylogenetic trees to determine the evolutionary relationship between isolates or to compare the sequences (alleles) at the different loci with the previously sequenced alleles (Aanensen and Spratt 2005).

Pulsed-field gel electrophoresis (PFGE)

In the past, Pulsed-field gel electrophoresis was widely used for the typing of *E. coli* isolates associated with outbreaks (Tenover et al. 1995). PFGE has also been used to study the origin of chromosome replication, detecting chromosome breaks and mapping (Tenover et al. 1995). As this technique has the ability to separate DNA molecules of up to 12 Mb in size (Maule, 1998) large fragments of DNA obtained after digestion with an infrequent cutting endonuclease are separated. Identical banding patterns are interpreted to indicate that strains are genetically related (Tenover et al. 1995).

1.2.2 *Escherichia coli* phylogroups

Initial studies to investigate the population structure of *E. coli* used 38 enzymes as part of a MLEE study to group the strains into a number of phylogroups. The four main groups detected were A, B1, B2 and D, with C and E as two additional minor phylogroups (Selander et al. 1986, Wirth et al. 2006, Tenaillon et al. 2010). The grouping was based on unique allelic combinations that were found multiple times (Tenaillon et al. 2010). More than a 1000 *E. coli* isolates,

which included human isolates from ECOR (an *E. coli* reference collection that contains genetically variable *E. coli* isolates; Ochman and Selander, 1984a) were used for developing the phylogrouping system.

The ancestral phylogroup was thought to be D but recently it has been suggested that B2 and D diverged simultaneously (Leimbach et al. 2013). Phylogroups A and B1 are very similar and are thus thought to have split at a later stage (Lecointre et al. 1998). Wirth et al. (2006) used a Bayesian approach to show that many of the ECOR isolates fell into the four main phylogroups, but that about a third represented hybrid phylogroups. Two of the hybrids found by Wirth and his colleagues were an A and B1 hybrid (AxB1) and an A, B, and D hybrid (ABD) which is the result of recombination between these phylogroups (Wirth et al. 2006). In fact, they detected recombination within all the phylogroups, but the rate of recombination varied. They also showed that mutation plays a role in shaping some of the phylogroups (Wirth et al. 2006).

These phylogroups can typically be associated with different environments and abilities. Phylogroup A consists mainly of commensal *E. coli,* Phylogroup B1 is associated with both pathogenic and commensal *E. coli,* B2 consists mainly of extraintestinal pathogenic *E. coli* (ExPEC) and group D consists of Uropathogenic *E. coli* (UPEC), enteroaggregative *E. coli* (EAEC), ExPEC and environmental strains (Leimbach et al. 2013). *S*tudies also indicated that the phylogroups found in the secondary environment seem to typically belong to B1 and A (Orsi et al. 2007, Walk et al. 2007, Gordon et al. 2008, Whittam, 1989). Méric and colleagues found that phylogroup A and B2 were linked to the primary host (Méric et al. 2013), while strains isolated from the aerial parts of the plants (Méric et al. 2013) mostly belonged to phylogroup B1. This could indicate that phylogroup B1 *E. coli* isolates are better adapted to plant colonization (Méric et al. 2013).

E. coli strains can be assigned to the four main groups by a triplex PCR. The presence or absence of the *chuA* and *yjaA* genes and the gene fragment TSPE4.C2 are used in this triplex PCR (Clermont et al. 2000). The *chuA* is involved in heme transport, *yjaA* encodes an unknown protein and the gene fragment TSPE4.C2 is part of a putative lipase esterase gene (Clermont et al. 2000). In the original article by Clermont et al. (2000) strains that did not contain any of the gene fragments were grouped with phylogroup A. Gordon et al. (2008), however, stated that assigning phylogroup A to these strains could be erroneous and showed that only 18% of strains that were negative for all three genes belonged to phylogroup A (Gordon et al. 2008). The triplex PCR has subsequently been modified to include an additional marker (i.e. the *arpA* gene) (Clermont et al. 2013). The resulting quadruplex PCR can be used to determine the phylogroups A, B1, B2, C, D, E and F. In other words this system now includes the known major and minor phylogroups as well as phylogroup F (Clermont et al. 2013, Jaureguy et al. 2008). Phylogroup F was denoted as phylogroup D by the original multiplex PCR, but MLST proved that these isolates were not part of phylogroup D (Jaureguy et al. 2008). The new quadruplex PCR has the ability to accurately distinguish between phylogroup D and F.

1.2.3 Clonal composition of *Escherichia coli*

Whittam (1996) defined a clonal population as a group of individuals that are not necessarily phenotypically identical but that share genomic similarities due to a common ancestor. Accordingly, several authors report that *E. coli* has a predominantly clonal population structure (Ochman and Selander, 1984b, Whittam, 1989, Leimbach et al, 2013) but the associated strains are not phenotypically identical (Ishii and Sadowsky, 2008).

One of the reasons for this phenotypic variation amongst the members of a clone involves adaptation to different habitats (Cooper and Lenski, 2000). As

the *E. coli* move to different niches, the habitat can shape the phenotype and the *E. coli* may lose or gain the ability to survive in an environment (Cooper and Lenski, 2000, Ishii and Sadowsky, 2008). These changes could be due to horizontal gene transfer and the acquisition of plasmids, as well as inversions, duplications and mutations occurring during DNA replication (Whittam, 1996, Ishii and Sadowsky, 2008). If one of these changes allows for a competitive advantage in a new environment, this clone will replace existing clones in that environment (i.e., periodic selection) (Whittam, 1996). However, a change that allows for an advantage in one environment may be a disadvantage in another and implies that environmental dynamics also impacts upon population composition (Whittam, 1996, Cooper and Lenski, 2000). In fact, the influence of the environment on a strain could be so substantial that the same strain may not be able to survive in two different environments (Ishii and Sadowsky, 2008).

1.2.4 Cryptic clades of *Escherichia*

A study by Walk et al in 2009 identified five clades or cryptic species within *E. coli sensu lato*. These clades were denoted as CI, CII, CIII, CIV and CV (Walk et al. 2009). These clades were identified by collecting isolates from different hosts and habitats (Walk et al. 2009). An extended MLST analysis using 22 *E. coli* housekeeping genes was performed on all the collected isolates, as well as *E. albertii, E. fergusonii* and four *Shigella* spp. Phylogenetic analysis of these datasets revealed eight clades. Three of these clades respectively presented *E. coli* including *Shigella, E. fergusonii* and *E. albertii* (Walk et al. 2009). The remaining five clades did not group with any of the known species and was suggested to represent separate species distinct from *E. coli* and the other known species of the genus (Walk et al. 2009).

Walk and colleagues also investigated the relatedness among the *Escherichia* species. They found that three (*rpos, fumC* and *lysP*) of the 22 genes analyses always grouped monophyletically for all the clades (Walk et al. 2009). They also

found that *E. coli sensu stricto* and CI, though rarely monophyletic, clustered together. They interpreted this as an indication of the close phylogenetic relationship between the two clades. Within the 22 gene dataset CIII, CIV and CV seldom associated with *E. coli sensu stricto* and these clades also contained most of the environmental strains isolated from freshwater beaches (Walk et al. 2009). The isolates classified as environmental in the study by Oh et al. are part of the clades (Oh et al. 2012).

According to Walk et al. (2009), the genus *Escherichia* shared a common ancestor 48-75 million years ago (mya). As time passed the clades split from the common ancestor accounting for the differences seen in the genus (Walk et al. 2009). *E. fergusonii, E. albertii,* CII and CV diverged from the common ancestor between 38-75 mya and *E. coli,* CI, CIII and CIV split from the common ancestor 19-31 mya. The latter group split into *E. coli*-CI and CIII-CIV more recently (Walk et al. 2009).

Despite fact that some of these clades split from *E. coli* a long period ago, they are still phenotypically indistinguishable from *E. coli* (Walk et al, 2009). Clade III proved to be phenotypically more different from conventional *E. coli,* as it has lost genes for sucrose utilization. This characteristic made Clade III the only clade that is biochemically distinct from the other clades and *E. coli sensu stricto* (Walk et al. 2009).

The clades were compared with isolates from human and animal origin and it was found that these environmental isolates did undergo specialization for adaptation to a new niche (Oh et al. 2012). They gained genes that allowed for increased survival in their new habitat and they lost genes that were no longer necessary (Oh et al. 2012). The study showed that the human-associated isolates have two sets of genes that are conserved and seem to be important for survival in humans; these are genes that allow defence against acids and drugs as well as genes that allow attachment to human tissues (Oh et al. 2012). The environmental isolates lost a number of genes that also seem to allow

improved colonization of the human gut. These genes are not needed for survival in the environment outside the host and code for functions such as nutrient uptake and adherence (Oh et al. 2012). The fact that these genes were lost in the environmental strains could mean that these strains have lost their ability to adhere to the human gut (Oh et al. 2012). Limited gene flow between the clades (with the exception of Clade I) have been observed which could be indicative of the existence of ecological barriers and habitat specific adaptations (Lou et al. 2011, Oh et al. 2012).

In conclusion, it was evident that *E. coli* is a very diverse microorganism that can be associated with the primary host as well as the environment outside the primary host. *E. coli* was primarily associated with the gut of humans and warm-blooded animals; however it has been proven that *E. coli* can multiply and persist in soils, sands, plants, algae and water. *E. coli* strains could naturally occur in the secondary environment, adapt to the secondary environment or undergo differential gene expression depending on the environment. Phylogroups B1 and A are often associated with environmental *E. coli*. Novel *Escherichia* clades (CIII, CIV and CV) also have been shown to be associated with the secondary environment. These clades have additional genes which are specific for survival in the environment and have lost the genes needed for survival in the primary host. These clades and faecal isolates may not exchange genes or have limited ability to exchange genes. This suggests that the genomes of strains found in the secondary environment would differ from the genomes of strains found in the primary host.

1.3 REFERENCES

Aanensen, D.M. and Spratt, B.G. (2005) The multilocus sequence typing network: mlst.net. *Nucleic Acid Research* Volume 33 pp. 728-733.

Alm, E.W., Burke, J. and Spain, A. (2003) Fecal indicator bacteria are abundant in wet sand at freshwater beaches. *Water research* Volume 37 pp. 3978-3982.

Alm, E.W., Walk, S.T. and Gordon, D.M. (2011) 'The Niche of *Escherichia coli*', in Walk, S. T. and Feng, P.C.H. (ed.), Population genetics of bacteria. ASM Press: Washington DC USA pp. 70-84.

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

Ballmer, K., Korczak, B.M., Kuhnert, P., Slickers, P., Ehricht, R. and Hächler, H. (2007) Fast DNA serotyping of *Escherichia coli* by use of an Oligonucleotide microarray. *Journal of Clinical Microbiology* Volume 45 Number 2 pp. 370-379.

Beutin, L., Geier, D., Steinruck, H., Zimmermann, S. and Scheutz, F. (1993) Prevelance and some properties of verotoxin (shiga-like toxin) producing *Escherichia coli* in seven different species of healthy domestic animals. *Journal of Clinical Microbiology* Volume 31 Number 9 pp. 2483-2488.

Bordalo, A.A., Onrassami, R. and Dechsakulwatana, C. (2002) Survival of faecal indicator bacteria in tropical estuarine waters (Bangpakong River, Thailand). *Journal of Applied Microbiology* Volume 93 pp. 864-871.

Brandl, M.T. (2006) Fitness of human enteric pathogens on plants and implications for food safety. *Annual review phytopathology Volume 44 pp. 367-392.*

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

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

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

Byappanahalli, M.N., Whitman, R.L., Shively, D.A., Ting, W.T.E., Tseng, C.C. and Nevers, M.B. (2006a) Seasonal persistence and population characteristics of *Escherichia coli* and enterococci in deep backshore sand of two freshwater beaches. *Journal of Water and Health* Volume 04.3 pp. 313-320.

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

Bywater, R., Deluyker, H., Deroover, E., de Jong, A., Marion, H., McConville, M., Rowan, T., Shryock, T., Shuster, D., Thomas, V., Vallé, M. and Walters, J. (2004) A European survey of antimicrobial susceptibility among zoonotic and commensal bacteria isolated from food-producing animals. *Journal of Antimicrobial Chemotherapy* Volume 54 pp. 744-754.

Caponigro, V., Ventura, M., Chiancone, I., Amato, L., Parente, E. and Piroa, F. (2010) Variation of microbial load and visual quality of ready-to-eat salads by vegetable type, season, processor and retailer. *Food Microbiology* Volume 27 pp. 1071-1077.

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

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* Volume 5 Number 1 pp. 58-65.

Cooley, M.B., Chao, D. and Mandrell, R.E. (2006) *Escherichia coli* O157:H7 survival and growth on lettuce is altered by the presence of epiphytic bacteria. *Journal of Food Protection* Volume 69 Number 10 pp. 2329-2335.

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

Desmarais, T.R., Solo-Gabriele, H.M. and Palmer, C.J. (2002) Influence of soil on fecal indicator organisms in a tidally influenced subtropical environment. *Applied and Environmental Microbiology* Volume 68 Number 3 pp. 1165-1172.

Dho-Moulin, M. and Fairbrother, J.M. (1999) Avian pathogenic *Escherichia coli* (APEC). *Veterinary Research* Volume 30 pp. 299-316.

Fisher, M.M., Wilcox, L.W. and Graham, L.E. (1998) Molecular characterization of epiphytic bacterial communities on Charophycean green algae. *Applied and Environmental Microbiology* Volume 64 Number 11. pp. 4384-4389.

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

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

Gordon, D.M. and Cowling, A. (2003) The distribution and genetic structure of *Escherichia coli* in Australian vertebrates: host and geographic effects. *Microbiology* Volume 149 pp. 3575-3586.

Gyllenberg, H.G., Gyllenberg, M., Koski, T., Lund, T., Schindler, J. and Verlaan, M. (1997) Classification of *Enterobacteriaceae* by minimization of stochastic complexity. *Microbiology* Volume 143 pp. 721-732.

Holden, N., Pritchard, L. and Toth, I. (2008) Colonization outwith the colon: plants as an alternative environmental reservoir for human pathogenic enterobacteria. *Federation of European Microbiological Societies Review* Volume 33 pp. 689-703.

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 Science and Technology* Volume 41 pp. 2203-2209.

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

Ishii, S. and Sadowski, M.J. (2008) *Escherichia coli* in the environment: Implications for water quality and human health. *Microbes and Environments* Volume 23 Number 2 pp. 101-108.

Jaureguy, F., Landraud, L., Passet, V., Diancourt, L., Frapy, E., Guigon, G., Carbonnelle, E., Lortholary, O., Clermont, O., Denamur, E., Picard, B., Nassif, X. and Brisse, S. (2008) Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains. *BioMedCentral Genomics* Volume 9 pp. 560-573.

Khachatryan, A.R., Hancock, D.D., Besser, T.E. and Call, D.R. (2004) Role of calf-adapted *Escherichia coli* in maintenance of antimicrobial drug resistance in dairy calves. *Applied and Environmental Microbiology* Volume 70 Number 2 pp. 752-757.

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* Volume 8 Issue 2 pp. 217-226.

Lecointre, G., Rachdi, L., Darlu, P. and Denamur, E. (1998) *Escherichia coli* molecular phylogeny using the incongruence length difference test. *Molecular Biology and Evolution* Volume 15 Number 12 pp.1685-1695.

Leimbach, A., Hacker, J. and Dobrindt, U. (2013) *E. coli* as an all-rounder: The thin line between commensalism and pathogenicity. *Current Topics in Microbiology and Immunology* Volume 358 pp. 3-32.

Lindow, S.E. and Brandl, M.T. (2003) Microbiology of the phyllosphere. *Applied and Environmental Microbiology* Number 69 Volume 4 pp. 1875-1883.

Lindow, S.E. and Leveau, J.H.J. (2002) Phyllosphere microbiology. *Current Opinion in Biotechnology* Volume 13 pp. 238-243.

Luo, C., Walk, S.T., Gordon, D.M., Feldgarden, M. and Tiedje, J.M. (2011) Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species.

The Proceedings of the Natural Academy of Sciences USA Volume 108 Number17 pp. 7200-7205.

Maiden, M.C.J., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M. and Spratt, B.G. (1998) Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *The Proceedings of the Natural Academy of Sciences USA* Volume 95 pp. 3140-3145.

Maule, J. (1998) Pulsed-field gel electrophoresis. *Molecular Biotechnology* Volume 9 pp. 107-126.

Méric, G., Kemsley, E.K., Falush, D., Saggersm E.J. and Lucchini, S. (2013) Phylogenetic distribution of traits associated with plant colonization in *Escherichia coli. Environmental Microbiology* Volume 15 Number 2 pp. 487-501.

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

Ochman, H. and Selander, R.K. (1984b) Evidence for clonal population structure in *Escherichia coli. The Proceedings of the Natural Academy of Sciences USA* Volume 81 pp. 198-201.

Oh, S., Buddenborg, S., Yoder-Himes, D.R., Tiedje, J.M. and Konstantinidis, K.T. (2012) Genome diversity of *Escherichia* isolates from diverse habitats. *Public Library of Science* Volume 7 Issue 10 pp. 1-9.

Orsi, R.H., Stoppe, N.C., Sato, M.I.Z. and Ottoboni, L.M.M. (2007) Identification of *Escherichia coli* from groups A, B1, B2 and D in drinking water in Brazil. *Journal of Water and Health* Volume 5 pp. 323-327.

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

Priest, F.G. and Barker, M. (2010) Gram-negative bacteria associated with brewery yeasts: reclassification of *Obesumbacterium proteus* biogroup 2 as *Shimwellia pseudoproteus* gen. nov., sp. nov., and transfer of *Escherichia blattae* to *Shimwellia blattae* comb. nov. *International Journal of Systematic and Evolutionary Microbiology* Volume 60 pp. 828-833.

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* Volume 190 Number 20 pp. 6881-6893.

Rhodes, M.W. and Kator, H. (1988) Survival of *Escherichia coli* and *Salmonella* spp. in estuarine environments. *Applied and Environmental Microbiology* Volume 54 Number 12 pp. 2902-2907.

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 a northern lake water microcosm. *Journal of Water and Health* Volume 4 Number 3 pp. 389-393.

Savageau, M.A. (1983) *Escherichia coli* habitats, cell types and molecular mechanisms of gene control. *The American Society of Naturalists* Volume 122 Number 6 pp. 732-744.

Selander, R.K., Caugant, D.A., Ochman, H., Musser, J.M., Gilmour, M.N. and Whittam, T.S. (1986) Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. *Applied and Environmental Microbiology* Volume 51 Number 5 pp. 873-884.

Smith, J.M., Feil, E.J. and Smith, N.H. (2000) Population structure and evolutionary dynamics of pathogenic bacteria. *BioEssays* Volume 22 Issue 12 pp. 1115-1122.

Solo-Gabriele, H.M., Wolfert, M.A., Desmarais, T.R. and Palmer, C.J. (2000) Sources of *Escherichia coli* in a coastal subtropical environment. *Applied and Environmental Microbiology* Volume 66 Number 1 pp. 230-237.

Solomon, E.B., Yaron, S. and Matthews, K.R. (2002) Transmission of *Escherichia coli* O157:H7 for contaminated manure and irrigation water to lettuce plant tissue and its subsequent internalization. *Applied and Environmental Microbiology* Volume 68 Number 1 pp. 397-400.

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

Tenover, F.C., Arbeit, R.D., Goering, R.V., Mickelsen, P.A., Murray, B.E., Persing, D. H. and Swaminathan, B. (1995) Interpreting Chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: Criteria for bacterial strain typing. *Journal of Clinical Microbiology* Volume 33 Number 9 pp. 2233-2239.

van Belkum, A., Struelens, M., de Visser, A., Verbrugh, H. and Tibayrenc, M. (2001) Role of genomic typing in taxonomy, evolutionary genetics and microbial epidemiology. *Clinical Microbiology Reviews* Volume 14 Number 3 pp. 547-560.

van Elsas, J.D., Semenov, A.V., Costa, R. and Trevors, J.T. (2011) Survival of *Escherichia coli* in the environment: fundamental and public health aspects. *The International Society for Microbial Ecology Journal* Volume 5 pp. 173-183.

Vidotto, M.C., Müller, E.E., de Freitas, J.C., Alfieri, A.A., Guimarães, I.G. and Santos, D.S. (1990) Virulence factors of avian *Escherichia coli. Avian Diseases* Volume 34 pp. 531-538.

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

Walk, S.T., Alm, E.W., Gordon, D.M., Ram, J.R., Toranzos, G. A., Tiedje, J.M. and Whittam, T. S. (2009) Cryptic Lineages of the Genus *Escherichia. Applied and Environmental Microbiology* Volume 75 Number 20 pp. 6534-6544.

Wallace, J.S., Cheasty, T. and Jones, K. (1997) Isolation of vero cytotoxin-producing *Escherichia coli* O157 from wild birds. *Journal of Applied Microbiology* Volume 82 pp. 399-404.

Waturangi, D.E., Suwanto, A., Schwarz, S. and Erdelen, W. (2003) Identification of class 1 integrons- associated gene cassettes in *Escherichia coli* isolated from *Varanus* spp. in Indonesia. *Journal of Antimicrobial Chemotherapy* Volume 51 pp. 175-177.

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

Welch, R. A., V. Burland, G. Plunkett III, P. Redford, P. Roesch, D. Rasko, E. L. Buckles, S. R. Liou, A. Boutin, J. Hackett, D. Stroud, G. F. Mayhew, D. J. Rose, S. Zhou, D. C. Schwartz, N. T. Perna, H. L. Mobley, M. S. Donnenberg, and Blattner F.R. (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *The Proceedings of the Natural Academy of Sciences USA* Volume 99 pp. 17020-17024.

Wells, J.G., Shipman, L.D. Greene, K.D., Sowers, E.G. Cameron, D.N., Downes, F.P., Martin, M.L. Griffin, P.M., Ostroff, S.M., Potter, M.E. Tauxe, R.V. and Wachsmuth, I.K. (1991) Isolation of *Escherichia coli* serotype 0157:H7 and other Shiga-like toxin producing *E. coli* from dairy cattle*. Journal of Clinical Microbiology* Volume 29 Number 5 pp. 985-989.

Whitman, R.L. and Nevers, M.B. (2003) Foreshore sand as a source of *Escherichia coli* in nearshore water of a Lake Michigan Beach. *Applied and Environmental Microbiology* Volume 69 Number 9 pp. 5555-5562.

Whitman, R.L., Shively, D.A., Pawlik, H., Nevers, M.B. and Byappanahalli, M.N. (2003) Occurrence of *Escherichia coli* and Enterococci in *Cladophora* (Chlorophyta) in nearshore water and beach sand of Lake Michigan. *Applied and Environmental Microbiology* Volume 69 Number 8 pp. 4714-4719.

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

Whittam, T.S. (1996) 'Genetic variation and evolutionary processes in natural populations of *Escherichia coli',* in Neidhardt, F.C. (ed.), *Escherichia coli* and *Salmonella*: cellular and molecular biology. *American Society for Microbiology,* Washington, DC p. 2708-2720.

Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., Wieler, L.H., Karch, H., Reeves, P.R., Maiden, M.C.J., Ochman, H. and Achtman, M. (2006) Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Molecular Microbiology* Volume 60 Number 5 pp. 1136-1151.

CHAPTER 2

ESCHERICHIA COLI DIVERSITY IN SUBTROPICAL CATCHMENTS

ESCHERICHIA COLI DIVERSITY IN SUBTROPICAL CATCHMENTS

2.1 INTRODUCTION

E. coli is a diverse bacterial species which consists of both commensal and pathogenic bacteria (Lavigne and Blanc-Potard, 2008). Although the gastro-intestinal tract of warm-blooded animals is regarded as the primary habitat of *E. coli*, this species may also occur in so-called secondary environments that include soil, water, plants and algae (Byappanahalli and Fujioka, 2004). Due to the marked differences between the primary and secondary environments (Alm et al. 2003, Whitman et al. 2003, Byappanahalli and Fujioka, 2004, Anderson et al. 2005, Bergholz et al. 2011), it is generally thought that the population numbers of the bacterium will rapidly decrease when entering the secondary environment (Oh et al. 2012). For this reason *E. coli* is widely used to indicate faecal contamination of water sources (World Health Organization, 2010).

Various studies have shown that *E. coli* can persist and multiply in the environment outside of the host (Savageau, 1983, Byappanahalli et al. 2003, Ishii et al, 2007, van Elsas et al. 2011). In such situations it has been shown and that the presence of specific isolates can often not be linked to faecal contamination (Byappanahalli et al. 2006). At the phenotypic level these environmental persistent strains can usually not be differentiated from the faecal isolates (Oh et al. 2012). Although initially reported for the warmer tropical and sub-tropical environments, it was also observed for different temperate environments (Byappanahalli and Fujioka, 2004, Ishii et al. 2006). To explain the existence of these *E. coli* populations, various authors hypothesize that some of the environmental isolates represent specific self-sustaining

genotypes that are well adapted to these environments (Whittam 1989, Power et al. 2005, Ishii et al. 2007).

The impact of the secondary environment on the diversity, population biology and evolution of *E. coli* is not well understood. *E. coli* is currently divided into phylogroups based on the use of *chuA* (heme transport), *yjaA* (unknown function) marker genes and TSPE4.C2, a DNA fragment (putative lipase esterase) (Clermont et al. 2000, Gordon et al. 2008, Clermont et al. 2013). Accordingly, there are four main phylogroups, groups A, B1, B2 and D, and the minor phylogroups, groups C, E and F, which do not consistently form separate clusters (Selander et al. 1986, Wirth et al. 2006, Jaureguy et al. 2008, Tenaillon et al. 2010, Leimbach et al. 2013). Various researchers have reported that there may be some link between the phylogroups and the environments they are associated with. For example, strains from rivers and soils primarily belong to phylogroup B1 and to a lesser extend to group A (Walk et al, 2007, Ratajczak et al. 2010). Phylogroup B1has also been described as generalists with minimal distinct adaptations to link them with a specific host (White et al. 2011).

The evolution of *E. coli* has also been studied using a Multilocus sequence typing (MLST) scheme based on 22 housekeeping genes (Walk et al. 2009). Analysis of these genes separated isolates that were collected from a wide range of environments and that were phenotypically identified as *E. coli* into six groups. Only one of these groups represent *E. coli sensu stricto*, while the other five represent novel clades (CI-CV) or cryptic species within the *Escherichia* genus (Walk et al. 2009). The most ancestral clade seems to be CV followed by CII, CIII and CIV and then finally CI which is the closest to *E. coli* (Walk et al. 2009, Luo et al 2011). Although some strains representing these clades were isolated from humans and animals, Clade III, IV and V were dominated by strains isolated from environmental samples (Walk et al. 2009).

Genome sequence information provides further support for the observed differentiation among isolates associated with the so-called primary and secondary environments of *E. coli.* When the genomes of strains representing the different clades were compared to those of strains belonging to *E. coli sensu stricto*, a large number of genes were found to be specific to the environmental isolates and to the isolates of enteric origin (Luo et al. 2011). Therefore, different gene sets are likely important for survival in specific habitats (Luo et al. 2011) and the different groups have likely lost the genes that are not needed for survival in a particular environment (Oh et al. 2012). Consistent with the close relationship between *E. coli sensu stricto* and Clade I, the latter was also the only clade that appeared to regularly exchange genes with enteric *E. coli*. Ecological barriers apparently limit gene exchange between the other clades and *E. coli sensu stricto* (Luo et al. 2011).

Little is known about the diversity of *E. coli* in subtropical aquatic environments. All previous studies investigating the diversity of *E. coli* found in secondary environment focused on soil and sand from tropical and temperate environments. Furthermore, the results of these previous studies suggested that the diversity of *E. coli* is likely to be uniquely distributed among different locations due to the significant effects of environmental factors (e.g., temperature and nutrient availability). The overall objective of this study was therefore to determine the diversity of *E. coli* within a subtropical aquatic environment. For this purpose, *E. coli* isolates associated with a range of niches within a fresh water dam in South Africa were investigated. To determine whether specific forms of *E. coli* are associated with aquatic plants, isolates from an additional seven fresh water dams were examined. The *rpoS* gene region was sequenced to determine if any of the isolates belonged to the cryptic environmental clades. This region and part of the *uidA* gene was also used to identify genetically unique clusters that potentially represent environmental isolates.

2.2 MATERIALS AND METHODS

2.2.1 Sample collection and *E. coli* isolation

Rietvlei Dam

The main environment sampled during this study was the Rietvlei dam (Pretoria, Gauteng, South Africa). This dam supplies about 10% of the drinking water that Pretoria requires through the Rietvlei Dam Water Treatment Works operated by the Tshwane Metropolitan Municipality (Bodenstein et al. 2006). The dam is also an important recreational facility as it hosts yacht and canoe clubs. The dam was ideally suited for this study as urban areas drain into this dam and it has limited inflow from one sewage treatment works. Sediment, water and plant debris samples were collected from different locations on the Rietvlei dam (Figure 2.1). The samples were diluted to 10^{-3} in sterile quarter strength Ringer's solution. From each dilution, 100 µl was plated onto MLGA (Membrane Lactose Glucuronide Agar) (Oxoid). This medium contained 40 g/l Peptone, 6 g/l Yeast extract, 30 g/l Lactose, 0.2 g/l Phenol red, 1 g/l Sodium lauryl sulphate, 0.5 g/l sodium pyruvate, 10 g/l Agar and 0.2 g/l X-glucuronide. The inoculated MLGA plates were incubated overnight at 37°C. The samples that did not yield any isolates when diluted were concentrated by filtration of 1 ml of the sample through 0.45 µm Whatman® filters (Merck). The filters were then placed on MLGA and incubated overnight at 37°C. Green colonies were assumed to be *E. coli* as MLGA is a selective media that differentiates *E. coli* from other coliforms (Fricker et al. 2008). The green colonies were re-streaked to obtain pure colonies for the subsequent analyses.

Plant samples

Aquatic plants were sampled from 8 dams in the Gauteng Province of South Africa including the Rietvlei Dam. The other dams included Leeukraal, Hartebeespoort, Klipvoor, Roodekoppies, Bon Accord, Roodeplaat and

Buffelspoort dams. During collection, plant samples were placed in plastic buckets and processed within 24 hours of receiving the samples in the laboratory. For analysis, a section of the plant was cut and placed into 50 ml of sterile Ringer's solution. After vigorous shaking, 10 ml was filtered through 0.45 µm Whatman® filters (Merck) and plated out on MLGA. The plates were incubated at 37°C overnight. The green colonies were re-streaked to obtain pure colonies as they were assumed to be *E. coli.*

Other isolates included

Additional isolates from a parallel study investigating the *E. coli* populations in similar aquatic environments in the Gauteng Province of South Africa were also included. The samples included 13 isolates associated with water hyacinth from Roodeplaat dam, six isolates from the Zeekoegat sewage treatment works and another from Baviaanspoort sewage treatment works. Lastly, 4 isolates from rivers in the vicinity of the Kusile power station construction site was included.

In an effort to obtain isolates representing *E. coli* circulating in the human population, samples were taken from the Hartebeesfontein sewage works (Kempton Park, Gauteng, South Africa). The final effluent from this plant drains into the Rietvlei dam. Samples were taken of both the raw sewage and the final effluent. The raw samples were diluted to 10^{-6}, final samples were diluted 10^{-4} and 100µl of each dilution was plated on MLGA and incubated overnight at 37°C. The green colonies were assumed to be *E. coli* and were re-streaked to obtain pure colonies.

Compost tea samples received from Kareebosch during the course of this study were also included. Compost tea is a water mixture used to fertilize vegetables. It is made by mixing water, woodchips, compost, cattle manure and goat manure. The mixture is left for a two weeks, after which the liquid phase is removed and used as organic fertilizer. Samples of compost tea (5), the water used to produce compost tea (1) and samples of the individual components of compost tea (4) were received. The compost tea, water and component

samples were diluted to 10^{-6} and 100 µl of each of these dilutions were plated onto MLGA. The samples were also filtered through Whatman® filters (Merck) if necessary and the filters were placed on MLGA and incubated overnight at 37°C. The green colonies were assumed to be *E. coli* and were re-streaked to obtain pure colonies.

During this study *E. coli* isolates from drinking water distribution networks were received and included. These *E. coli* strains were isolated from a large distribution network. The strains could thus not be linked to any specific faecal contamination event.

2.2.2 Verification of the *E. coli* isolates

The identity of the *E. coli* isolates was confirmed using the Colilert ®-18 (Dehteq) test. One colilert capsule was added to 100 ml distilled H_2O and then 5 ml of the Colilert medium was aliquoted into test tubes. All the selected green colonies were inoculated in this medium and incubated at 37°C for 18 hours. The test tubes were thereafter viewed under ultraviolet light for fluorescence to confirm the presence of *E. coli.* All confirmed *E. coli* isolates were stored at -70°C according to the manufacturer's instruction using Microbank™ (Pro-lab Diagnostics) beads.

2.2.3 PCR and Sequencing

For extracting DNA, the isolates confirmed as *E. coli* was streaked for pure culture on MLGA and incubated overnight at 37°C. DNA was extracted using the ZR Genomic DNA II Kit ™ (Quick- gDNA™ miniprep) (Zymo research). Genomic DNA was extracted according to the manufacturer's instructions.

Polymerase chain reaction (PCR) was used to amplify the genes encoding sigma factor S (*rpoS*) and β-D-glucuronidase (*uidA*) as described previously

(Michigan State University MLST database; www.shigatox.net). For these amplifications, each 25 ul reaction mixture contained 2 mM $MgCl_2$ (Separation Scientific), 250 µM of each dNTP (Fermentas),5 pmol of the forward and of the reverse primer (Table 2.1), 2-4 ng/ul of template DNA, 0.006 U/µl Super-therm *Taq* polymerase and 1X reaction buffer (Separation Scientific). Amplification was done using a Veriti Thermal Cycler (Applied Biosystems). The amplification cycle was initiated at 94°C for 10 minutes, followed by 30 cycles of denaturing at 92°C for 1 minute, annealing at 60°C for 1 minute and extension at 72°C for 1 minute. The final extension was done at 72°C for 5 minutes. A negative control containing no DNA was included for each set of PCR reactions to ensure that no cross-contamination occurred.

The amplified products were subjected to 1% agarose gel electrophoresis (Sambrook and Russell, 2001) and visualized using GelRed (Biotium) according to the manufacturer's specifications. Products in the expected size range i.e., 618 base pairs for *rpoS* and 658bp for *uidA* were purified enzymatically using Exonuclease I (Fermentas) Fast AP Alkaline Phosphatase (Fermentas) following the manufacturer's protocol. The purified products were then sequenced with the same forward primers as before, and the BigDye® Terminator v3.1 Cycle Sequencing Kit (Applied BioSystems) using the Veriti Thermal Cycler (Applied Biosystems) and the standard sequencing protocol.

All raw sequences were viewed and trimmed using BioEdit Sequence Alignment Editor v. 7.0.9.0 (Hall, 1999). All sequences were also compared to those in the nucleotide database of the National Centre for Biotechnology Information (NCBI; http://blast.ncbi.nlm.nih.gov/Blast) BLAST (Altschul et al. 1990) to confirm that the sequences represented those of the *rpoS* and *uidA* genes, respectively.

2.2.4 Phylogenetic comparison

Apart from the sequences for the isolates obtained in this study, the sequences for a number of other isolates for which whole genome sequence are available, were included. The *rpoS* dataset included 32 *E. coli* strains with known phylogroup assignment (Leimbach et al. 2013), which represented commensal, ETEC, EHEC, EAEC, EPEC, ExPEC and AIEC *E. coli* isolates (Leimbach et al. 2013). This dataset also included 37 strains representing the cryptic clades of *Escherichia* (Walk et al. 2009) and the type strains of the four *Shigella* species (Leimbach et al. 2013), as well as *E. fergusonii* for outgroup purposes. The *uidA* dataset also included the sequences for the 32 *E. coli* strains representing known phylogroup assignment (Leimbach et al. 2013), but only eight strains representing the cryptic clades (Luo et al, 2010) were included. The GenBank accession numbers (Benson et al. 2011, http://www.ncbi.nlm.nih.gov/genbank/) of the genomes of these isolates are represented in Table 2.2.

The sequences were aligned using both ClustalW multiple alignment (Thompson et al. 1994) and MAFFT (Katoh et al. 2002). Maximum Likelihood (ML) phylogenetic analyses were conducted using PhyML 3.0 (Guindon et al. 2010) and the best-fit evolutionary model parameters as indicated by jModelTest software v. 0. 1. 1 (Posada, 2008) based on the Akaike Information Criterion (Akaike, 1974, Posada and Buckley, 2004). The *rpoS* dataset utilized the TIM transitional model (Posada, 2003) with gamma correction (G) to account for among site rate variation. The *uidA* dataset utilized the TVM transversional model (Posada, 2003) with G and a proportion of invariable sites (I). Branch support was evaluated using 1000 bootstrap replicates and the same ML parameters as before (Felsenstein, 1981). Trees were viewed and edited using MEGA5 (Tamura et al. 2011) and Inkscape v. 0.48.4 (http://inkscape.org/).

2.3 RESULTS

2.3.1 Isolation of *E. coli*

A total 281 *E. coli* isolates were obtained from the various samples collected during this study (Tables 2.3 and 2.4). From the sampling at the Rietvlei dam, 81 *E. coli* isolates were obtained and 15 isolates came from the Hartebeesfontein sewage works. The samples associated with compost tea production provided 82 *E. coli* isolates. From the aquatic plant samples collected in 8 different dams, 68 *E. coli* isolates were obtained (Table 2.4). A further 35 other isolates were included, these isolates were from a drinking water distribution network, Roodeplaat dam, Zeekoegat and Baviaanspoort sewage treatment works.

2.3.2 Phylogenetic analysis

The *rpoS* dataset consisted of 364 sequences that included those of the 281 isolates obtained in this study, as well as 37 sequences representing the novel clades of *Escherichia*, 36 sequences representing the *E. coli* phylogroups and four *Shigella* type strains and finally nine *rpoS* sequences of known *E. coli* isolates and the outgroup (*E. fergusonii*). The *rpoS* dataset contained 446 aligned nucleotides and include 6 alignment gaps.

The *uidA* dataset consisted of 330 sequences that in addition to those for the 281 isolates from this study also included 32 sequences representing the *E. coli* phylogroups, four sequences for the *Shigella* type strains and five sequences representing known *uidA E. coli* isolates. This dataset also included eight sequences representing the novel *Escherichia* clades where two sequences represent each of clades I, III, IV and V. The uidA dataset contained 510 aligned nucleotides and included 11 alignment gaps.

The ML phylogenies inferred from both datasets revealed that the isolates examined were highly diverse, although the *uidA* tree was generally better resolved (Figures 2.2 and 2.3). The isolates from the sewage treatment works that represent the primary human habitat was also diverse and grouped throughout both phylogenetic trees. Many of the environmental isolates could not be separated from those representing strains with a human origin. However, none of the sequences for the isolates collected during this study grouped with those representing the novel cryptic clades of *Escherichia*, indicating that they all belong to *E. coli sensu stricto.*

For both *rpoS* and *uidA*, the phylogroups were spread across the trees, with phylogroups A and B1 not separated from each other in either of the trees. In the *rpoS* tree single isolates belonging to phylogroups B2 and D2 (re-classified as F, Jaureguy et al. 2008) did not group with the rest of the isolates from that specific phylogroup. The only phylogroup that formed a well-defined cluster, which was well separated from the environmental strains from this study, was phylogroup D2 (F) in the *uidA* tree. The *E. coli* isolates obtained during this study grouped with or in the vicinity of three of the phylogroups. Most of the isolates grouped with phylogroups A and B1. None of the isolates from this study grouped with isolates representing phylogroup E and only a single isolate grouped with phylogroup D2 (F) in the *rpoS* phylogenetic tree.

The results of this study also revealed the existence of groups that are genetically unique and that potentially represent environmental groups. Three such groups (Clusters 1-3) were supported by both the *rpoS* and *uidA* trees (Figures 2.2 and 2.3), where Cluster 2 was associated with phylogroup D1, and contained strains isolated from plant debris from the Rietvlei dam. The plant debris isolates grouped closely with sediment isolates from the Rietvlei dam in both phylogenetic trees. Cluster 3 was associated with phylogroup B2, and consisted of the water hyacinth isolates from the Roodeplaat dam. Cluster 1 did not group closely to any of the phylogroups, and consisted of isolates from wood chips which is one of the components of compost tea. An additional four groups (Clusters 4R-7R) potentially of environmental origin were supported by

the *rpoS* data and another ten (Clusters 4U-13U) were supported by the *uidA* data (Figures 2.2 and 2.3).

2.4 DISCUSSION

The *rpoS* and *uidA* phylogenetic trees showed that diversity amongst the *E. coli* isolates obtained from these aquatic and related environments were high. Most of the isolates did not group according to the type of sample or the geographic origin of the sample. The majority of environmental samples could also not be separated from those isolates that represented the primary human habitat.

Although grouping with the rest of the clades in the *rpoS* tree, Clade I grouped within the middle of the unrooted *uidA* tree. This is not unexpected for the unrooted tree as Clade I has previously been shown to be the clade that is most closely related to *E. coli* and is considered by Clermont and colleagues to be an eighth phylogroup of *E. coli* (Walk et al. 2009, Lou et al. 2011, Clermont et al. 2013). In both trees the strains representing the other clades grouped separately, which is consistent with the notion that, despite being phenotypically indistinguishable from *E. coli sensu stricto*, these clades were genetically unique (Walk et al. 2009). This supports the idea that the clades (Clade III-V) are potential cryptic species within the genus *Escherichia* (Walk et al. 2009).

Both phylogenetic trees clearly showed that the *E. coli* isolated for this study are conspecific with *E. coli* sensu stricto. However, the fact that none of the isolates collected during this study were related to any of the cryptic clades of *Escherichia* was rather unexpected. At least three of the cryptic clades are primarily associated with isolates of environmental origin (Walk et al. 2009). Studies have shown that these three clades have evolved for survival in the environment and have lost many genes associated with growth in the human intestines (Luo et al. 2011, Oh et al. 2012). However none of the *E. coli* isolates isolated during this study were part of the *Escherichia* clades. This could mean

that the clades are not as wide spread as thought or that the clades are restricted to specific secondary environments.

The phylogroups are widely distributed across the *rpoS* and *uidA* phylogenetic trees. Most of the *E. coli* isolates seemed to group with or to be associated with phylogroups A and B1. This was not unexpected as these phylogroups are widely reported to be associated with isolates obtained from the environment (Gordon et al. 2008, Tenaillon et al. 2010, Méric et al. 2013). The majority of isolates isolated from the environment (mineral water and wells) were identified as being part of phylogroup A and B1 (Orsi et al. 2007, Méric et al. 2013). A study by White and colleagues in 2011 found that phylogroup B1 is associated with non-human environments. Furthermore a study by Méric et al. in 2013 indicated that isolates with the ability to grow on plants are part of phylogroup B1. This implies that different phylogroups have different environmental capabilities (Méric et al. 2013). This could mean that the ability of *E. coli* to survive in the environment depend on the presence or absence of genes (*chuA, yjaA*) and DNA fragment TSPE4.C2 which are used to differentiate between the phylogroups (White et al. 2011).

Several isolates also grouped with phylogroup B2 and D1. Isolates that are part of phylogroups B2 and D are expected to cause extra-intestinal infections. Virulence genes are found more regularly in phylogroups B2 and D, than phylogroups A and B1 (Picard et al. 1999, White et al. 2011). This makes sense as the isolates generally associated with phylogroup B2 are ExPEC (extraintestinal pathogenic *E. coli*) strains and with D1 is UPEC (uropathogenic *E. coli*) and EAEC (enteroaggregative *E. coli*) (Leimbach et al. 2013). Phylogroups B2 and D are less likely to be isolated from the environment and are usually associated with humans (Gordon et al. 2008). In the current study, two of the unique environmental groups (Clusters 2 and 3) were associated with phylogroups D1 and B respectively. This could mean that environmental isolates are not limited to phylogroups A and B1.

None of the isolates examined in this study could clearly be linked with phylogroup E and D2 (F). The isolates which form part of phylogroup F were

previously misclassified as phylogroup D (Jaureguy et al. 2008, Tenaillon et al. 2010, Clermont et al. 2013). Phylogroup E was only recently identified as a known phylogroup, which is due to the fact that phylogroup E is a minor group and does not group consistently (Tenaillon et al. 2010, Clermont et al. 2013). Phylogroup E is associated with EHEC (enterohaemorrhagic *E. coli*) and EPEC (enteropathogenic *E. coli*), thus it would not be unexpected that environmental isolates did not group with phylogroup E (Leimbach et al. 2013). Phylogroup F consists of ExPEC (extraintestinal pathogenic *E. coli*) and a single isolate identified as environmental (SMS-3-5) (Leimbach et al. 2013). This isolate has multiple antibiotic resistance and was obtained from an industrial, toxic metal-contaminated coastal environment (J. Craig Venter Institute, http://gsc.jcvi.org/projects/msc/e_coli_and_shigella/escherichia_coli_secec_sms-3-5/index.shtm). Given the environments with which these phylogroups are associated, it is not surprising that the isolates examined in this study were not related to them.

The *rpoS* and *uidA* phylogenetic trees both yielded possible unique environmental clusters. These clusters could represent unique environmental isolates as no human associated isolates grouped within these clusters. A study by Byappanahalli et al. (2006) found that *E. coli* isolated from the secondary environment (soil) were diverse, yet formed groupings that were distinct from faecal isolates. In another study Whittam (1989) showed that the *E. coli* isolated from the environment surrounding birds and bird faeces were genetically distinct and formed separate clusters. This indicated that faecal/sewage and environmental *E. coli* can be distinguished based on genetic differentiation, which means that the clusters in this study that do not contain any sewage isolates could possibly be environmental *E. coli* clusters. The clusters most likely to be environmental clusters are Clusters 1-3.

2.5 CONCLUSIONS

A diverse range of *E. coli* has been isolated from the aquatic and associated habitats sampled during this study. The *rpoS* and *uidA* genes of these isolates were sequenced and it was established that the *E. coli* isolates are not part of the cryptic clades but belonged to *E. coli sensu stricto.* These isolates could therefore be included in a population genetic study to investigate gene flow amongst *E. coli* strains isolated form aquatic and related environments.

The environmental *E. coli* isolates are very diverse, yet most could not be separated from the human associated isolates. There was however three unique clusters observed in both the *rpoS* and *uidA* trees that could represent unique environmental strains. These three clusters associated with phylogroups B2 and D that are more likely to cause extra-intestinal infection (Clermont et al. 2013) and did not group with phylogroup A or B1, which are typically associated with environmental isolates.

It would be important to determine whether these three unique environmental groups remained constant when additional genes are analyzed. To get better resolution the genes need to be carefully selected to ensure that they are more variable than the *rpoS* and *uidA* genes used in the current study. It is also hoped that they would be able to differentiate better between sewage isolates and some of the other potential environmental clusters observed during this study.

It is clear from the above that the diverse *E. coli* isolates from this study all belonged to the same species. It is important to next determine if this population of diverse strains has a specific structure and whether this structure (sub-populations) can be linked to the specific niches from which they were isolated. Without this information if would be difficult to determine how these environmental strains have adapted or evolved and what impact they might have on human health.

2.6 REFERENCES

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* Volume 19 Number 6 pp. 716–723.

Alm, E.W., Burke, J. and Spain, A. (2003) Fecal indicator bacteria are abundant in wet sand at freshwater beaches. *Water research* Volume 37 pp. 3978-3982.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology* Volume 215 Number 3 pp. 403-410.

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

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Sayers, E.W. (2011) GenBank. *Nucleic Acids Research* Volume 39 pp. 32-37.

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* Volume 77 Number 1 pp. 211-219.

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.

Byappanahalli, M., Fowler, M., Shively, D. and Whitman R. (2003) Ubiquity and Persistence of *Escherichia coli* in a Midwestern Coastal Stream. *Applied and Environmental Microbiology* Volume 69 Number 8 pp. 4549-4555.

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* Volume 50 Number 1 pp. 27–32.

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

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

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* Volume 5 Number 1 pp. 58-65.

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

Fricker, C.R., DeSarno, M., Warden, P.S. and Eldred, B.J. (2008) False-negative β-D-glucuronidase reactions in membrane lactose glucuronide agar medium used for the simultaneous detection of coliforms and *Escherichia coli* from water. *Letters in Applied Microbiology* Volume 47 pp. 539-542.

Gordon, D.M., Clermont, O., Tolley, H. and Denamur, E. (2008) Assigning *Escherichia coli* strains to phylogenetic groups: multilocus sequence typing

versus the PCR triplex method. *Environmental Microbiology* Volume 10 Number 10 pp. 2484-2496.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W. and Gascuel, O. (2010) New algorithms and methods to estimate Maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systemic Biology* Volume 59 Number 3 pp. 307-321.

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

Inkscape (2011) Inkscape: Draw freely, viewed 2013 < http://inkscape.org/>

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 Science and Technology* Volume 41 pp. 2203-2209.

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

J. Craig Venter Institute (2008) *Escherichia coli* SECEC SMS-3-5. <http://gsc.jcvi.org/projects/msc/e_coli_and_shigella/escherichia_coli_secec_sms-3-5/index.shtm>

Jaureguy, F., Landraud, L., Passet, V., Diancourt, L., Frapy, E., Guigon, G., Carbonnelle, E., Lortholary, O., Clermont, O., Denamur, E., Picard, B., Nassif, X. and Brisse, S. (2008) Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains. *BioMedCentral Genomics* Volume 9 pp. 560-573.

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

Lavigne, J. and Blanc-Potard, A. (2008) Molecular evolution of *Salmonella enterica* serovar *Typhimurium* and pathogenic *Escherichia coli*: From pathogenesis to therapeutics. *Infection, Genetics and Evolution* Volume 8 Issue 2 pp. 217-226.

Leimbach, A., Hacker, J. and Dobrindt, U. (2013) *E. coli* as an all-rounder: The thin line between commensalism and pathogenicity. *Current Topics in Microbiology and Immunology* Volume 358 pp. 3-32.

Luo, C., Walk, S.T., Gordon, D.M., Feldgarden, M. and Tiedje, J.M. (2011) Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *The Proceedings of the Natural Academy of Sciences USA* Volume 108 Number17 pp. 7200-7205.

Méric, G., Kemsley, E.K., Falush, D., Saggersm E.J. and Lucchini, S. (2013) Phylogenetic distribution of traits associated with plant colonization in *Escherichia coli. Environmental Microbiology* Volume 15 Number 2 pp. 487-501.

Michigan State University (2006) Department of Microbiology and Molecular genetics. <www.shigatox.net>

Oh, S., Buddenborg, S., Yoder-Himes, D.R., Tiedje, J.M. and Konstantinidis, K.T. (2012) Genome diversity of *Escherichia* isolates from diverse habitats. *Public Library of Science* Volume 7 Issue 10 pp. e47005-e47005.

Orsi, R.H., Stoppe, N.C., Sato, M.I.Z. and Ottoboni, L.M.M. (2007) Identification of Escherichia coli from groups A, B1, B2 and D in drinking water in Brazil. *Journal of Water and Health* Volume 05.2 pp. 323-327.

Picard, B., Garcia, J.S., Gouriou, S., Dureiz, P., Brahimi, N., Bingen, E., Elion, J. and Denamur, E. (1999) The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection. *Infection and Immunity* Volume 67 Number 2 pp. 546-553.

Posada, D. (2008) jModelTest: Phylogenetic Model Averaging, *Molecular Biology and Evolution,* Volume 25 Number 7 pp. 1253-1256.

Posada, D. (2003) Selecting a model of nucleotide substitution. *Current Protocols in Bioinformatics* pp. 6.5.1-6.5.14. John Wiley & Sons, Inc., New York.

Posada, D. and Buckley, T.R. (2004) Model selection and model averaging in phylogenetics: Advantages of Akaike Information Criterion and Bayesian approaches over likelihood ratio tests. *Society of Systematic Biologists* Volume 53 Number 5 pp. 793-808.

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

Ratajczak, M., Laroche, E., Berthe, T., Clermont, O., Pawlak, B., Denamur, E. and Pitet, F. (2010) Influence of hydrological conditions on the *Escherichia coli* population structure in the water of a creek on a rural watershed. *BMC Microbiology* Volume 10 pp. 222-231.

Sambrook, J. and Russell, D.W. (2001) Molecular cloning: A laboratory manual. 3rd edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Savageau, M.A. (1983) *Escherichia coli* habitats, cell types and molecular mechanisms of gene control. *The American Society of Naturalists* Volume 122 Number 6 pp. 732-744.

Selander, R.K., Caugant, D.A., Ochman, H., Musser, J.M., Gilmour, M.N. and Whittam, T.S. (1986) Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. *Applied and Environmental Microbiology* Volume 51 Number 5 pp. 873-884.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, and Kumar S (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution* Volume 28 Number 10 pp. 2731-2739.

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

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

van Elsas, J.D., Semenov, A.V., Costa, R. and Trevors, J.T. (2011) Survival of *Escherichia coli* in the environment: fundamental and public health aspects. *The International Society for Microbial Ecology Journal* Volume 5 pp. 173-183.

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

Walk, S.T., Alm, E.W., Gordon, D.M., Ram, J.R., Toranzos, G. A., Tiedje, J.M. and Whittam, T. S. (2009) Cryptic Lineages of the Genus *Escherichia. Applied and Environmental Microbiology* Volume 75 Number 20 pp. 6534–6544.

White, A.P., Sibley, K.A., Sibley, C.D., Wasmuth, J.D., Schaefer, R., Surette, M.G., Edge, T.A. and Neumann, N.F. (2011) Intergenic sequence comparison of *Escherichia coli* isolates reveals lifestyle adaptations but not host specificity. *Applied and Environmental Microbiology* Volume 77 Number 21 pp. 7620-7632.

Whitman, R.L., Shively, D.A., Pawlik, H., Nevers, M.B. and Byappanahalli, M.N. (2003) Occurrence of *Escherichia coli* and Enterococci in *Cladophora* (Chlorophyta) in nearshore water and beach sand of Lake Michigan. *Applied and Environmental Microbiology* Volume 69 Number 8 pp. 4714-4719.

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

Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., Wieler, L.H., Karch, H., Reeves, P.R., Maiden, M.C.J., Ochman, H. and Achtman, M. (2006) Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Molecular Microbiology* Volume 60 Number 5 pp. 1136-1151.

World Health Organization (2010). Guidelines for Drinking-water Quality. Volume 1: Recommendations 4th edition. Geneva, World Health Organization.

Table 2.1: List of primers used for amplification and sequencing of selected genes

Primer name	Sequence (5'-3')
*rpoS	CGC CGG ATG ATC GAG AGT AA
rpoS	GAG GCC AAT TTC ACG ACC TA
*uidA	CAT TAC GGC AAA AGT GTG GGT CAAT
uidA	TCA GCG TAA GGG TAA TGC GAG GTA

* Indicates forward primer

Table 2.2: The accession numbers of the isolates used

Isolate name	Accession number	Isolate details
AF002204 *E. coli*	AF002204.1	*Escherichia coli* O55:H7 *rpoS* gene
AF002207 *E. coli*	AF002207 .1	*Escherichia coli* O157:H7 *rpoS* gene
AF002205 *E. coli*	AF002205.1	*Escherichia coli* O157:NM *rpoS* gene
AF002206 *E. coli*	AF002206 .1	*Escherichia coli* O78:K80 *rpoS* gene
AF002208 *E. coli*	AF002208.1	*Escherichia coli* O157:H7 *rpoS* gene
AF002209 *E. coli*	AF002209.1	*Escherichia coli* O157:H7 *rpoS* gene
AY698845 *E. coli*	AY698845.1	*Escherichia coli* strain 5898-71 *rpoS gene*
AY698870 *E. coli*	AY698870.1	*Escherichia coli* strain 1758-70 *rpoS* gene
AY723475 *E. coli*	AY723475.1	*Escherichia coli* strain DEC 2a *rpoS* gene
CU928158.2 *E. fergusonii*	CU928158.2	*Escherichia fergusonii ATCC 35469 rpoS gene*
AB454547.1 *E. coli*	AB454547.1	*Escherichia coli* CK28 *uidA* gene
AB334714 *E. coli*	AB334714.1	*Escherichia coli* 12646 *uidA* gene
AY698445 *E. coli*	AY698445.1	*Escherichia coli* LT-82 *uidA* gene
HM221006 *E. coli*	HM221006.1	*Escherichia coli* RFA16 *uidA* gene
HM221054 *E. coli*	HM221054.1	*Escherichia coli* RFA19 *uidA* gene
B827I		*Escherichia* Clade I strain B827
E1492I		*Escherichia* Clade I strain E1492
E807I		*Escherichia* Clade I strain E807
H442I		*Escherichia* Clade I strain H442
M863I		*Escherichia* Clade I strain M863
TW10509I		*Escherichia* Clade I strain TW10509
TW11930I		*Escherichia* Clade I strain TW11930
TW11966I		*Escherichia* Clade I strain TW11966
TW15838I		*Escherichia* Clade I strain TW15838
B1147II		*Escherichia* Clade II strain B1147
TW09231III		*Escherichia* Clade III strain TW09231
TW09276III		*Escherichia* Clade III strain TW09276
TW09254III		*Escherichia* Clade III strain TW09254
TW09266III		*Escherichia* Clade III strainTW09266
TA04III		*Escherichia* Clade III strain TA04
B685III		*Escherichia* Clade III strain B685
TW14182IV		*Escherichia* Clade IV strain TW14182
TW11588IV		*Escherichia* Clade IV strain TW11588
H605IV		*Escherichia* Clade IV strain H605
B49IV		*Escherichia* Clade IV strain B49
TW09308V		*Escherichia* Clade V strain TW09308
B1225V		*Escherichia* Clade V strain B1225
B646V		*Escherichia* Clade V strain B646
E1118V		*Escherichia* Clade V strain E1118
E1195V		*Escherichia* Clade V strain E1195
E1196V		*Escherichia* Clade V strain E1196
E471V		*Escherichia* Clade V strain E471
E472V		*Escherichia* Clade V strain E472
E620V		*Escherichia* Clade V strain E620
M1108V		*Escherichia* Clade V strain M1108
TA290V		*Escherichia* Clade V strain TA290

Table2.2: continued

TW14263V		*Escherichia* Clade V strain TW14263
TW14264V		*Escherichia* Clade V strain TW14264
TW14265V		*Escherichia* Clade V strain TW14265
TW14266V		*Escherichia* Clade V strain TW14266
TW14267V		*Escherichia* Clade V strain TW14267
RL325/96V		*Escherichia* Clade V strain RL325/96
Z205V		*Escherichia* Clade V strain Z205
ATCC8739A	CP000946.1	Commensal *Escherichia coli* ATCC 8739 Phylogroup A
HS A	NC_009800.1	Commensal *Escherichia coli* HS Phylogroup A
K12A	U00096.2	Commensal *Escherichia coli* K-12 MG1655 Phylogroup A
H10407A	NC_017633.1	ETEC *Escherichia coli* H10407 Phylogroup A
O26H11B1	NC_013361.1	EHEC *Escherichia coli* O26:H11 11368 Phylogroup B1
O111HB1	NC_013364.1	EHEC *Escherichia coli* O111:H- 11128 Phylogroup B1
O103H2B1	NC_013353.1	EHEC *Escherichia coli* O103:H2 12009 Phylogroup B1
55989B1	NC_011748.1	EAEC *Escherichia coli* 55989 Phylogroup B1
E24377AB1	NC_009801.1	ETEC *Escherichia coli* E24377A Phylogroup B1
SE11B1	NC_011415.1	Commensal *Escherichia coli* SE11 Phylogroup B1
IAI1B1	NC_011741.1	Commensal *Escherichia coli* IAI1 Phylogroup B1
CB9615E	NC_013941.1	EPEC *Escherichia coli* CB9615 Phylogroup E
EDL933E	NC_002655.2	EHEC *Escherichia coli* O157:H7 EDL 933 Phylogroup E
SakaiE	NC_002695.1	EHEC *Escherichia coli* O157:H7 Sakai Phylogroup E
042D1	NC_017626.1	EAEC *Escherichia coli* 042 Phylogroup D1
UMN026D1	NC_011751.1	ExPEC *Escherichia coli* UMN026 Phylogroup D1
SMS35D2	CP000970.1	ExPEC Escherichia coli SECEC SMS-3-5 Phylogroup D2 (F)
IAI39D2	NC_011750.1	ExPEC *Escherichia coli* IAI39 Phylogroup D2 (F)
CE10D2	NC_017646.1	ExPEC *Escherichia coli* CE10 Phylogroup D2 (F)
ABU83972B2	NC_017631.1	ExPEC *Escherichia coli* ABU 83972 Phylogroup B2
CFT073B2	NC_004431.1	ExPEC *Escherichia coli* CFT073 Phylogroup B2
LF82B2	NC_011993.1	AIEC *Escherichia coli* LF82 Phylogroup B2
NRG857CB2	CP001855.1	AIEC *Escherichia coli* NRG 857C Phylogroup B2
ED1aB2	NC_011745.1	Commensal *Escherichia coli* ED1a Phylogroup B2
APEC01B2	NC_008563.1	ExPEC *Escherichia coli* APEC 01 Phylogroup B2
S88B2	NC_011742.1	ExPEC *Escherichia coli* S88 Phylogroup B2
IHE3034B2	NC_017628.1	ExPEC *Escherichia coli* IHE3034 Phylogroup B2
UTI89B2	NC_007946.1	ExPEC *Escherichia coli* UTI89 Phylogroup B2
536B2	NC_008253.1	ExPEC *Escherichia coli* 536 Phylogroup B2
SE15B2	NC_013654.1	Commensal *Escherichia coli* SE15 Phylogroup B2
NA114B2	NC_017644.1	ExPEC *Escherichia coli* NA114 Phylogroup B2
E2348/69B2	NC_011601.1	EPEC *Escherichia coli* E2348/69 Phylogroup B2
S. sonnei	NC_007384.1	*Shigella sonnei* Ss046
S. boydii	NC_007613.1	*Shigella boydii* Sb227
S. flexneri	NC_004337.2	*Shigella flexneri* 2a str. 301
S. dysenteriae	NC_007606.1	*Shigella dysenteriae* Sd197

Table 2.3: Indicates the isolate names and the sample types and points

Isolate name	Sample type	Sample point
TS1B, TS2A, TS2B, TS3A, TS3B,TS4B, TS5A, TS5B, TS6A, TS6B, TS7A, TS7B, TS8B, TS9A, TS10A, TS11A, TS12B, TS13A, TS14B, TS15A, TS15B, TS16A, TS17A, TS17B, TS18A	Dam water	Rietvlei Dam
S21.1G, S21.2G, S21.3G, S21.4G, S21.5G, S21.6G, S21.7G, S21.8G, S21.9G, S21.10G, S21.11G, S21.13G, S21.14G, S226Y, S2F1.1G, S2F1.2G, S2F1.3G, S2F1.4G, S2F1.5G, S2F1.6G, S2F1.7G, S2F1.8G, S2F1.11G, S2F1.12G, S2F1.13G, S2F1.14G, S2F1.15G, S2F2.1G, S2F2.3G, S3F2.1G, S3F2.2G, S3F2.3G	Sediment	Rietvlei Dam
DWWF1.2G, DWWF1.3G, DWWF1.4G, DWWF2.1G, DWWF2.2G, DWWF2.3G, DWWF2.4G, DWWF2.5G, DWWF2.6G, DWWF2.7G, DWWF2.8G, DWWF2.9G, DWWF2.10G, DWWF2.11Y/G, DWPD5.1, DWPD5.2, DWPD5.3, DWPD5.4, DWPD5.5, DWPD5.6, DWPD5.7, DWPD5.8, DWPD5.9, DWPD5.10	Plant debris	Rietvlei Dam
Raw4.5G, Raw4.6G, Raw4.8G2, Raw4.9G2, FINAL2.1G, FINAL2.2G, FINAL2.3G, FINAL2.5G, FINAL2.6G, FINAL2.7G, FINAL2.8G, FINAL2.8G2, FINAL2.9G, FINAL2.10G, FINAL2.15Y	Sewage	Hartebeesfontein sewage works
Ep141, Ep153, Ep154, Ep168, SP004, SP1000, NP720, NP707, NP705	Drinking water	Drinking water distribution networks
TomB1G, TomB3G, TomB7G, TomB8G, TomB9G, TomB10G, TomB11G, TomB12G, TomB13G, TomB14G, TomB15G, TomBF1.1G, TomBF1.2G, TomBF2.4G, TomCF1.1G, TomCF1.2G, TomCF2.3G, KBCT1G, KBCT2G, KBCTF1.1G, KBCTF1.2G, NCTR1, NCTR2, NCTR3, NCTR4, NCTR5, NCTR6, NCTR7, 1NKBKT, 2NKBKT, NKBKT2, NKBKT3, NKBKT4, NKBKT5, NKBKT6	Compost tea	
CTWF1.1, CTWF1.2, CTWF1.3, CTWF1.4, CTWF1.5, CTWF2.6	Compost tea water	

Table 2.3 continued:

Isolate name	Sample type	Sample point
WC1.1, WC1.2, WC1.3, WC1.4, WC1.5, WC1.6, WC1.7, WC2.8, WC2.9, WCF1.10, WCF2.11, BeM1.1, BeMF1.9, BeMF2.2, BeMF2.3, BeMF2.4, BeMF2.5, BeMF2.6, BeMF2.7, BeMF2.8, C2d3.9, C2d3.10, C2d3.11, C2d3.12, C2d3.13, C2d3.14, C2d3.15, C2d2.16, C2d2.17, C2d2.18, C2dF1.1, C2dF2.2, C2dF2.3, C2dF2.4, C2dF2.5, C2d5.6, C2d5.7, C2d5.8, BoM1.1, BoM1.2, BoMF1.5	Components of compost tea	
LKD3.1, L1F1.2, L1F1.3, L1F2.1, L1F2.3	Plant	Leeukraal Dam
RVD1.1, RVD2.2, RVD3.1, RVD4.1, RVD4.2, RVD5.2, RV1F1.1, RV1F1.2, RV1F1.3, RV1F2.1, RV1F2.2, RV1F2.3	Plant	Rietvlei Dam
HBPD1.1, HBPD1.2, HBPD2.1, HBPD5.2, H2F1.2, H2F2.2, H2F2.3, H3F1.1, H3F2.1	Plant	Hartebeespoort Dam
KVD1.1, KVD1.2, KVD3.1, KVD5.1, KVD5.2, K1F1.2, K1F1.3	Plant	Klipvoor Dam
BAD1.1, BAD1.2, BAD2.1, BAD2.2, BAD3.1, BAD3.2, BAD4.1, BAD4.2, BAD5.1, BAD5.2, BA1F1.2, BA1F2.1, BA1F2.2, BA1F2.3	Plant	Bon Accord Dam
RKD1.1, RKD1.3, RKD2.2, RKD3.1	Plant	Roodekopjies Dam
R1F1.1, R1F1.2, R1F1.3, R1F1.5, R1F2.1, R1F2.5, R1S1.1, R1S1.2, R1S1.3, R2F1.2, R2F1.3, R2F1.4	Plant	Roodeplaat Dam
B1F1.1, B1F1.2, B1F1.3, B1F2.2, B1F2.3	Plant	Buffelspoort Dam
Q098, Q021	Dam water	Roodeplaat Dam
ZA2.4, ZA1.8, ZA2.7, ZB2.9, ZB2.7, ZB1.2	Sewage	Zeekoegat sewage treatment works
B1.2	Sewage	Baviaanspoort sewage treatment works
Q02H1, Q02H2, Q02H3, Q02H4, Q02H5, Q02H6, Q02H8, Q02H9, Q02H10, Q02H11, Q02H12, Q02H13, Q02H15	Water hyacinth	Roodeplaat Dam
SW7FE8, SW9FE12, SW11FE16, Spring2FE4	Surface water	Kusile power station construction site.

Table 2.4: List of the plant species sampled from the 8 dams

Plant sample	Dam	Common name	Scientific name
RKD 3	Roodekopjies	Water hornwort	*Ceratophyllum demersum*
RKD 2	Roodekopjies	Parrot's feather	*Myriophyllum aqauticum*
RKD 1	Roodekopjies	Parrot's feather	*Myriophyllum aqauticum*
BAD 5	Bon Accord	Parrot's feather	*Myriophyllum aqauticum*
BAD 4	Bon Accord	Parrot's feather	*Myriophyllum aqauticum*
BAD 3	Bon Accord	Parrot's feather	*Myriophyllum aqauticum*
BAD 2	Bon Accord	Parrot's feather	*Myriophyllum aqauticum*
BAD 1	Bon Accord	Parrot's feather	*Myriophyllum aqauticum*
KVD 5	Klipvoor	Parrot's feather	*Myriophyllum aqauticum*
KVD 3	Klipvoor	Parrot's feather	*Myriophyllum aqauticum*
KVD 2	Klipvoor	Water hyacinth	*Eichhornia crassipes*
KVD 1	Klipvoor	Parrot's feather	*Myriophyllum aqauticum*
HBPD 5	Hartebeespoort	Water weed	*Isolepis fluitans*
HBPD 2	Hartebeespoort	Water hornwort	*Ceratophyllum demersum*
HBPD 1	Hartebeespoort	Water hornwort	*Ceratophyllum demersum*
RVD 5	Rietvlei	Parrot's feather	*Myriophyllum aqauticum*
RVD 4	Rietvlei	Water weed	*Isolepis fluitans*
RVD 3	Rietvlei	Parrot's feather	*Myriophyllum aqauticum*
RVD 2	Rietvlei	Parrot's feather	*Myriophyllum aqauticum*
RVD 1	Rietvlei	Water weed	*Isolepis fluitans*
LKD 3	Leeukraal	Water weed	*Isolepis fluitans*
H2	Hartebeespoort	Water hornwort	*Ceratophyllum demersum*
R1	Roodeplaat	Water weed	*Isolepis fluitans*
RV1	Rietvlei	Water weed	*Isolepis fluitans*
H3	Hartebeespoort	Water weed	*Isolepis fluitans*
BA1	Bon Accord	Water hornwort	*Ceratophyllum demersum*
R2	Roodeplaat	Parrot's feather	*Myriophyllum aqauticum*
K1	Klipvoor	Parrot's feather	*Myriophyllum aqauticum*
L1	Leeukraal	Water weed	*Isolepis fluitans*
B1	Buffelspoort	Parrot's feather	*Myriophyllum aqauticum*

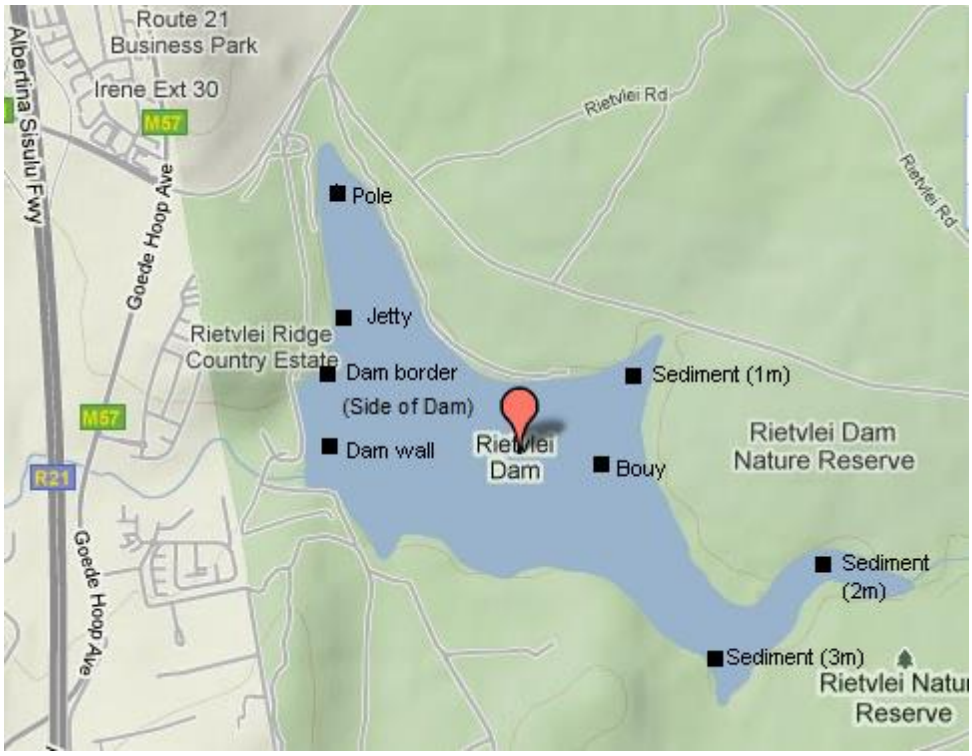

Figure 2.1: Map of Rietvlei dam indicating where the samples were taken. (©Google Maps)

Figure 2.2: The maximum likelihood tree of the *rpoS* gene of the 281 isolates, the phylogroups and the clades. Bootstrap values are represented as a percentage of 1000 replicates, all bootstrap values below 50% were not included. The brackets indicate the clusters that do not contain any sewage isolates. The key indicates the colour associated with a specific sample type as well as how the phylogroups are indicated in the tree.

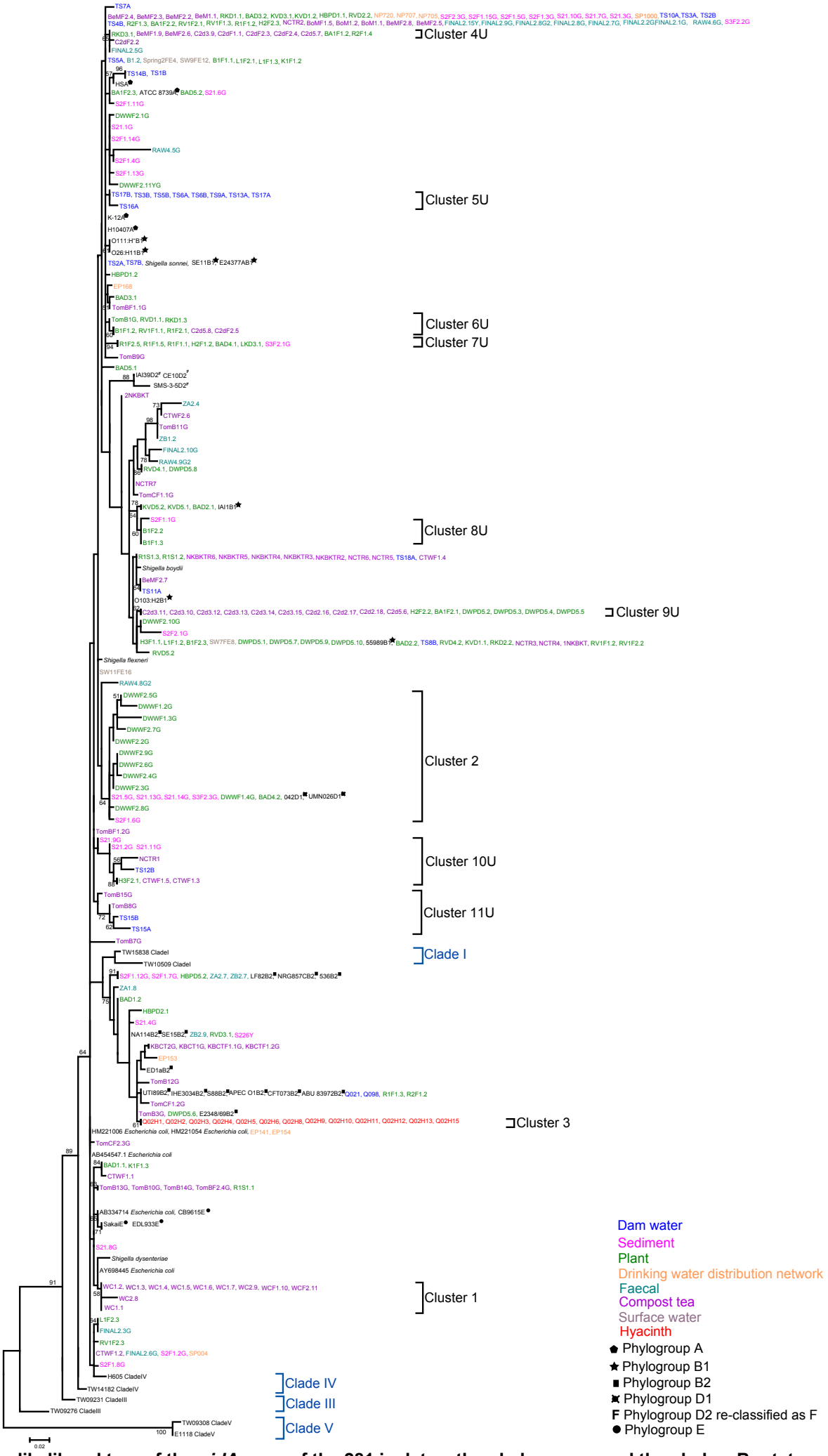

Figure 2.3: The maximum likelihood tree of the *uidA* gene of the 281 isolates, the phylogroups and the clades. Bootstrap values are represented as a percentage of 1000 replicates, all bootstrap values below 50% were not included. The brackets indicate the clusters that do not contain any sewage isolates. The key indicates the colour associated with a specific sample type as well as how the phylogroups are indicated in the tree.

CHAPTER 3

POPULATION DYNAMICS OF *ESCHERICHIA COLI*

POPULATION DYNAMICS OF *ESCHERICHIA COLI*

3.1 INTRODUCTION

Escherichia coli (*E. coli*) is primarily found in the gastro-intestinal tract of humans and warm-blooded animals (Gordon and Cowling, 2003). Many survival studies have indicated that *E. coli* does not live in non-host environments and that its detection in these environments is the result of continuous input from the primary habitat (Winfield and Groisman, 2004). There are however, numerous reports that *E. coli* can also be associated with environments outside the primary host (Savageau, 1983, Byappanahalli and Fujioka, 2004). These secondary habitats include soil and water and associations with plants and algae (Alm et al. 2003, Whitman and Nevers 2003, Byappanahalli and Fujioka, 2004, Anderson et al. 2005, Bergholz et al. 2011). The primary and secondary environments vary markedly in terms of prevailing biotic and abiotic factors and conditions (Savageau, 1983, Whittam, 1989, Walk et al. 2007). Knowledge of how *E. coli* has evolved and adapted to survive in these different environments is still fragmented.

Currently the main hypothesis is that certain phylogenetic lineages (phylogroups) within *E. coli* have adapted to persist and grow in the secondary environment although they still have the ability to circulate through the primary host. This hypothesis is supported by the evidence that the most dominant *E. coli* strains present in the environment belong to phylogroup A and B1 (Walk et al. 2007; White et al. 2011), although strains belonging to these two phylogroups are also commonly isolated from the human gut (Duriez et al. 2001). Savageau (1983) suggested that these *E. coli* have a dual control mechanism to regulate certain metabolic functions. The outcome of such a system will be the existence of two cell types that could deal with the different

nutritional demands of the primary and secondary habitats. This regulation system will allow the bacterium to survive longer in the secondary environment and thereby increase its ability to again colonize the primary host. This environmental adaptation could also be linked to the evolution of other properties. For example, a study by Méric and co-workers (2013) showed that isolates belonging to phylogroup B1 were more likely to possess phenotypic characteristics important for survival on plants. White and co-workers (White et al. 2011) also reported that phenotypes associated with environmental survival were significantly higher amongst the B1 isolates than amongst isolates belonging to the B2 phylogroup.

A variation on this main hypothesis to explain the growth and survival of *E. coli* in the secondary environment is that certain phylogenetic lineages of the species have evolved as free-living populations only present in the secondary environment. The idea that independent environmental populations occur in nature was initially based on the fact that *E. coli* could be isolated from the environment without any indication or link to faecal contamination having taken place (Walk et al. 2007). Ishii et al. (2006) also demonstrated that some "naturalized" *E. coli* strains formed part of the overall soil microbial community. Subsequently Walk et al. (2009) reported the existence of a group of genetically distinct environmental and animal isolates of *E. coli*. Although phenotypically indistinguishable from true members of the species (i.e. *E. coli sensu stricto*), phylogenetic analysis showed that they belonged to separate and, as yet, undescribed species in the genus *Escherichia*. Genomic studies (Luo et al. 2011, Oh et al. 2012) subsequently demonstrated that these cryptic species of *E. coli sensu lato* typically lack more than 90 genes considered to be important for interactions with human epithelial cells and the immune response. Under this hypothesis and sub-hypothesis, the ability of *E. coli* to proliferate in the environment has gradually evolved within specific phylogenetic lineages. As these lineages become more distinct recombination and gene flow between them will be limited resulting in a clear separation of the different populations.

An alternative to the main hypothesis (and its variant) for explaining why certain *E. coli* persist in the environment is that the genetic elements responsible for this ability are subject to horizontal gene transfer. In other words, the genes required for environmental persistence may be located on mobile elements and form part of the accessory genome of this species. This implies that the phenotypic property in question would have a patchy distribution among the various lineages of *E. coli*. If this hypothesis is true, environmental isolates will be randomly associated with the different phylogenetic lineages and will not be part of one or more well-defined populations.

In order to develop a better understanding of the emergence of environmental *E. coli*, insight into the population dynamics and related genetic processes associated with environmental populations are required. Therefore, the overall objective of this study was to determine the population dynamics amongst various freshwater and human *E. coli* populations. For this purpose, the DNA sequence information for four core genome genes (*rpoS*, *uidA*, *mutS* and *fadD*) were utilized. To infer how populations of the bacterium have evolved and are structured in this environment, coalescence and parsimony based methodologies were used (Huson, 1998, Posada and Crandall, 2001). This information is important to further our understanding of the emergence of environmental *E. coli* and could provide important clues as to how these populations potentially interact.

3.2 MATERIALS AND METHODS

3.2.1 *E. coli* isolates

A number of diverse *E. coli* strains representing various aquatic environments were included. Several of these isolates were obtained from aquatic plants sampled at different localities, whereas others were obtained from water and

sediments sampled at the Rietvlei dam. The collection also included strains isolated from drinking water distribution networks, sewage and compost tea used as a liquid fertilizer. The *E. coli* isolates therefore represented seemingly distinct niches within the environment. The isolation and verification procedure used was fully described in Chapter 2. A total of 281 *E. coli* isolates were isolated from distinct niches associated with the specific aquatic environments targeted. The codes of the *E. coli* isolates and sites these *E. coli* were isolated from are provided in Table 3.1. For the plant associated *E. coli* isolates information on the plant species and the dam from which these plant were sampled are shown in Table 3.2.

3.2.2 Selection of additional housekeeping genes

Of the four genes utilized in this study, the sequences for two (*uidA* and *rpoS*) were available from a previous study (Chapter 2 of this Dissertation). Two additional and more variable genes were identified by evaluating the 22 housekeeping gene sequences included in an extended MLST study for *E. coli* (Walk et al. 2009). The sequence data for the 22 genes were downloaded from GenBank (Benson et al. 2011, http://www.ncbi.nlm.nih.gov/genbank/). The 22 gene sequences were aligned using ClustalW multiple alignment (Thompson et al. 1994) and MAFFT (Katoh et al. 2002). The alignments were then trimmed and analysed for percentage variability using MEGA 5 (Tamura et al. 2011). For the selection of the two additional genes, the percentage variability within the various datasets was compared to those obtained for the *uidA* and *rpoS* datasets (Chapter 2 of this Dissertation).

3.2.3 Gene sequencing

Based on the analysis described above the methyl-directed mismatch repair (*mutS*) and fatty acyl-CoA synthetase (*fadD*) were selected as they were the most variable. For the present study the genes were sequenced using the PCR

protocol, PCR cycles and primers (Table 3.3) obtained from the Michigan State University (MSU) database (www.shigatox.net). The PCR was modified and performed as discussed in Chapter 2. The PCR clean-up, sequencing reaction and precipitation was performed as discussed in Chapter 2. The purified products were sequenced using an ABI Prism DNA Automated Sequencer (Perkin-Elmer). The identities of the sequences were confirmed using the National Centre for Biotechnology Information (NCBI) database blast tool (Altschul et al. 1990, http://blast.ncbi.nlm.nih.gov/Blast).

3.2.4 PHI test for recombination

The analysis was first performed on the *E. coli* strains isolated during this study (281 sequences) and then for the same strains as well as the *E. coli* representing the known phylogroups (Leimbach et al. 2013) (317 sequences). The analysis was run for the two collections of isolates, first using the sequences for each gene individually and then using the concatenated dataset (in the order *rpoS, uidA, mutS* and *fadD*).

To detect evidence of recombination the Φ_w or PHI (pairwise homoplasy index) test was performed using SplitsTree v. 4.12.8 (Huson and Bryant, 2006). This test involves calculation of the mean refined incompatibility scores obtained for nearby nucleotide sites along the sequences (Bruen et al. 2006), by making use of informative positions only (i.e., a site that has at least two different character states and these character states are found in at least two different sequences) (Bruen et al. 2006). The null hypothesis (Ho) of the PHI test assumes no recombination. This hypothesis is tested using simulations based on a natural coalescent model with recombination (Kingman 1982, Hudson 1983) and a window size of 100 bp (Bruen et al. 2006).

3.2.5 Determining the relationships between the isolates

Maximum Likelihood phylogenetic analyses

Datasets containing the *mutS* and *fadD* gene sequences of the strains was generated and combined with the sequences of the *E. coli* strains representing the phylogroups (Leimbach et al. 2013), the cryptic Clades I-V (Walk et al. 2009), and other *E. coli* and *E. fergusonii* sequences downloaded from the GenBank database (Benson et al. 2011, http://www.ncbi.nlm.nih.gov/genbank/). This dataset included 357 sequences, of which 281 were from the isolates examined in this study. The remaining sequences include 32 representatives of the various *E. coli* phylogroups, the four *Shigella* species (Leimbach et al. 2013), 37 *E. coli* sequences representing the Clades (Walk et al. 2009) and two known *E. coli* isolates (Walk et al. 2009).The accession numbers of the downloaded sequences are listed in Table 3.4.

Geneious v 4.8.5 (Drummond et al. 2010) was used to compile a concatenated dataset for *E. coli sensu stricto*. This was done by using the aligned *mutS* and *fadD* gene sequences determined here, as well as the *rpoS* and *uidA* data from Chapter 2. The concatenated dataset included only those taxa (i.e., 317 isolates) for which the sequence data for all four genes were available.

Maximum Likelihood (ML) phylogenetic analyses of the two individual genes and the four-gene dataset were conducted using PhyML 3.0 (Felsenstein, 1981, Guindon et al. 2010) and the best-fit evolutionary model parameters as indicated by jModelTest software v. 0. 1. 1 (Posada, 2008) based on the Akaike Information Criterion (AIC) (Akaike, 1974, Posada and Buckley, 2004, Posada, 2008). Branch support was evaluated using 1000 bootstrap replicates using the same model parameters and using PhyML 3.0 (Guindon et al. 2010). For the *mutS* and *fadD* trees *E. fergusonii was* used as the outgroup, and for the concatenated tree, the tree was rooted at midpoint with no outgroup as no *E. fergusonii* sequence is available for *uidA.* ML phylogenetic trees were viewed

as well as edited using MEGA5 (Tamura et al. 2011) and Inkscape v. 0.48.4 (http://inkscape.org/).

Estimation of genealogical relationships

The method of Templeton, Crandall and Sing (1992) (TCS) was used to estimate genealogies using statistical parsimony networks (Posada and Crandall, 2001). This procedure uses statistical parsimony to determine an ancestor (Posada and Crandall, 2001), by first collapsing the sequences into unique haplotypes (Table 3.5) and then generating a cladogram using statistical parsimony (Templeton et al. 1992; Clement et al. 2000). The analysis was done using TCS v.1.21 (Clement et al. 2000) with a default 95% connection limit based on the generated concatenated dataset.

Inference of phylogenetic networks

To infer phylogenetic networks, SplitsTree v. 4.12.8 (Huson and Bryant, 2006) was employed. The SplitsTree methodology used Neighbour-Net to determine weighted splits from which networks were drawn (Bryant and Moulton, 2004; Huson, 1998). Because these analyses included only unique sequences, haplotypes within the concatenated datasets were identified with DAMBE v. 5.3.12 (Xia and Xie, 2001). The codes for the isolates with identical sequences (i.e. haplotypes) are shown in Table 3.5. A Phylip 4(.phy) file format of the 258 unique concatenated sequences was used to generate a SplitsTree network.

3.3 RESULTS

3.3.1 Selection of additional housekeeping genes

Among the 22 housekeeping genes utilized in the extended MLST for *E. coli*, *mutS* and *fadD* were selected as additional markers. These genes were the most variable with 14.21% and 10.9% variability respectively (Figure 3.1). The *rpoS* gene was found to be the least variable with 1.06% variability. The *uidA* gene was more variable than *rpoS* but less variable than the two new genes with 6.73% variability.

3.3.2 PHI test for recombination

Tests for recombination within the various gene datasets, the PHI test for recombination were run using SplitsTree4 (Huson and Bryant, 2006). Inspection of the p-values for these tests revealed statistically significant evidence ($p<0.05$) for recombination only in the *mutS, fadD* and concatenated datasets (Table 3.6). In other words, for these three datasets the null hypothesis of no recombination could be rejected at the 95 % confidence level.

3.3.3 Determining the relationships between the isolates

Maximum Likelihood phylogenetic analyses

The *mutS* dataset contained 497 aligned nucleotides and included 2 alignment gaps. The *fadD* dataset contained 489 aligned nucleotides and included 16 alignment gaps and finally the concatenated dataset contained 1991 aligned nucleotides and included 15 alignment gaps. The substitution model selected by jModelTest for the *mutS* phylogenetic tree was TrN with gamma correction (G; determines the range of among site rate variation) and a proportion of invariable sites (I; gives the frequency of sites that do not evolve) (Tamura and

Nei, 1993). The model used for the *fadD* dataset was TPM2uf with G and I (Kimura, 1981). The concatenated dataset utilized the TVM model with G and I (Posada, 2003).

Apart from the majority of strains that did not group according to environmental sources, the *mutS* gene (Figure 3.2) tree had twelve unique clusters that were not associated with any of the sewage isolates. These clusters were denoted as Cluster 4M to 12M as well as Cluster 1, 2 and 3. Clusters 1-3 corresponded with the unique groups previously seen in the *rpoS* and *uidA* phylogenetic trees (Chapter 2 of this Dissertation). In the *mutS* tree, Cluster 1 was not clearly associated with a phylogroup whereas Cluster 2 was associated with phylogroup D1 and Cluster 3 with phylogroup B2.

The *fadD* phylogenetic tree (Figure 3.3) had seven unique clusters that did not contain any sewage isolates. These were denoted as Cluster 2 and Cluster 4F to Cluster 9F. The original Cluster 1 was split in the *fadD* tree in Cluster 6F and wood chip isolates grouped close to Cluster 6F. Cluster 2 is still present and well defined in *fadD*. Cluster 3 was not found in *fadD,* but was split into Cluster 9F and another cluster of Hyacinth isolates that grouped close to Cluster 9F.

When looking at the unique environmental clusters that formed, the *mutS* phylogenetic tree had 12 possible clusters which were not associated with any sewage isolates. The *fadD* had seven unique groups that only contained environmental isolates. The *fadD* phylogenetic tree was the least variable of the four gene trees. Cluster 1, 2 and 3 of the *mutS* phylogenetic tree were similar to the Cluster 1, 2 and 3 seen in the *rpoS* and *uidA* phylogenetic trees. Only Cluster 2 formed a well-defined grouping in the *fadD* phylogenetic tree. Cluster 6F and 9F of the *fadD* tree was not exactly the same but were similar to Cluster 1 and 3.

The concatenated phylogenetic tree (Figure 3.4) was rooted at midpoint (without an outgroup) due to the fact that *E. fergusonii* does not have the *uidA* gene and could not be used as the outgroup.The diversity revealed by this tree was high, where the three unique clusters (Cluster 1, 2 and 3) possibly representing environmental groups, clustered separately and were well supported. The bootstrap values of these groups were better in the concatenated phylogenetic tree than in any of the individual gene trees. Most of the isolates grouped with phylogroups A and B1. Phylogroups E and D2 (F) grouped with the least number of isolates from this study.

Estimation of genealogical relationships

The statistical parsimony analysis was done on the concatenated dataset of 281 isolates from this study as well as the 32 *E. coli* sequences representing the phylogroups and 4 *Shigella* species. The resulting TCS cladogram revealed similar groupings to those that were seen in the concatenated phylogenetic tree (Figure 3.4). Although there were a number of haplotypes that were connected only to a single haplotype or completely lacked connection to other haplotypes, the majority formed two separate parts. In the one, most of the haplotypes appeared to have radiated from a single ancestral node, while the other is a network and showed how these isolates are connected and the number of changes between them.

Inference of phylogenetic networks

The groupings observed in the TCS cladogram (Figure 3.5) corresponded to those observed in the SplitsTree network generated by the coalescent analysis of the *E. coli* haplotypes (Figure 3.6). However, the two TCS cycles did not form two separate groupings in the SplitsTree analysis. Also, the haplotypes representing phylogroups A and B1 are restricted to the one side of the

SplitsTree. There were also groups that seemed to differentiate from other haplotypes, especially those in Cluster 2. The “unconnected” haplotypes and those connected to single haplotypes in the TCS cladogram also grouped separately and further from the centre in the SplitsTree analysis.

3.4 DISCUSSION

The single gene trees as well as the concatenated tree confirmed that all the isolates included in this study belonged to *E. coli sensu stricto.* None of the isolates from this study, grouped with any of the *mutS* sequences representing Clades I-V as determined by Walk et al. (2009). In the *fadD* tree one of the isolates representing Clade I grouped amongst the *E. coli sensu stricto* isolates. This was not totally unexpected as it is still debated whether Clade I should be considered to represent a separate species or if it should only be considered as the 8th *E. coli* phylogroup (Clermont et al. 2013).

The *mutS* phylogenetic tree had the best resolution and provided good separation between isolates (compare Figures 3.1, 3.2 and 3.3). In contrast the *fadD* tree provided limited resolution, even less than what was observed for the *uidA* tree and especially among environmental isolates. This was unexpected as the initial sequence comparison showed the *fadD* gene to be more variable than the *uidA* gene. The *fadD* gene encodes a protein involved in fatty acid synthesis (Hsu et al. 1989) and has been used by Lacher et al. (2007) to study EPEC. They found that of the 483 sites examined, 32 were variable (this is equal to 6.6% a variability), which is comparable to the results of the current study, where the 508 nucleotide *fadD* dataset harboured 40 variable sites (7.87 % variability). The *fadD* variability in the two studies were respectively 6.6% and 7.87%, and only the *cyaA* gene were more variable (8.84%) of the 22 housekeeping genes used in the study by Walk et al. 2009. Because of its variability, the *fadD* gene is also commonly used as marker during MLST and

Single Nucleotide Polymorphism (SNP) analyses and is known to be variable (Tarr et al. 2002, Lacher et al. 2007, Sheludchenko et al. 2010, Bergholz et al. 2011). Why the *fadD* gene should be less variable amongst environmental strains therefore remains unclear.

Based on the PHI test only *fadD* and *mutS* showed clear signs of recombination. It could be that frequent recombination in these genes led to the increased variability predicted for the *mutS* and *fadD* genes. It is, however, unclear how this data should be interpreted in light of the limited variation observed for *fadD* amongst the environmental isolates. Contreras et al. in 2011 found recombination within the *fadD* gene, although they only detected 20 variable sites within the 483 site dataset examined (Contreras et al. 2011). This study also found recombination even though the variability was low within the *fadD* gene (Contreras et al. 2011). There seems to be a recombinational "hot spot" within the *mutS* gene thus it is not surprising that recombination was detected within *mutS* (Prunier and Leclercq, 2005).

The phylogenetic, TCS and SplitsTree analyses revealed the same overall structure for the *E. coli* populations in the aquatic environment as indicated by the colour coding used. These overall groupings that formed were not associated with specific environments, but corresponded well with the phylogroups. All phylogroups apart from A and B1 were clearly separated. Phylogroup A and B1 are considered to be sisters groups and have previously also been shown to group close together (Walk et al. 2007).

Most of the isolates from the environment grouped with phylogroup A and B1. This was seen for the individual gene trees based on the *mutS* and *fadD* sequence data as well as the phylogenetic, TCS and SplitsTree analyses based on the concatenated dataset. The fact that the majority of the isolates belonged to these two phylogroups supported the current hypothesis that environmental adaptation has taken place within specific lineages of *E. coli* (Whittam, 1989,

Gordon et al. 2002). This is because phylogroups A and B1, are often reported as the phylogroups most regularly associated with isolates from the environment (White et al, 2011, Tenaillon et al. 2010, Gordon et al. 2008). Phylogroup B1 specifically has been shown to be associated with *E. coli* isolated from plants (Méric et al. 2013) as well.

Three coherent groups of isolates were observed in most of the trees and most likely represent unique environmental isolates. Some environmental strains have previously also been observed to form coherent groupings. Both Whittam (1989) and Byappanahalli et al. in (2006) observed the separate clustering of isolates obtained from the secondary environment. Whittam (1989) also demonstrated that *E. coli* isolated from environments associated with birds and those from bird faeces formed two separate clusters.

The three unique environmental groups, which were consistently observed in all phylogenetic analyses, did not cluster with phylogroup A and B1 as would be expected for environmental isolates. Cluster 2 containing Dam wall weed (DWW) isolates grouped consistently with isolates known to belong to phylogroup D1. Isolates from Cluster 3 isolated from water hyacinths, on the other hand, consistently grouped with phylogroup B2 representatives. The association of these unique clusters with phylogroups B2 and D needs to be further investigated to determine if they are unique isolates only present in the aquatic environment or whether they could still be associated with humans and have an impact on their health. These two phylogroups are known to include extraintestinal pathogenic *E. coli* (ExPEC) isolates (Escobar-Páramo et al. 2004a, Escobar-Páramo et al. 2004b). The ExPEC isolates belonging to phylogroup B2 has been associated with several virulence factors (Escobar-Páramo et al. 2004a, Escobar-Páramo et al. 2004b). However, phylogroups B2 and D are not known to contain any *E. coli* isolates that produce toxins or that are enteroinvasive (Escobar-Páramo et al. 2004a, Escobar-Páramo et al. 2004b).

Some of the possible environmental clusters were not clearly associated with strains representing the specific phylogroups. This included the unique, environmental Cluster 1. For this group the closest representative strains belonged to *Shigella*. It would be important to perform the newly described phylotyping quadraplex PCR on these isolates to determine which phylogroup they form part of (Clermont et al. 2013).

3.5 CONCLUSIONS

The overall structure observed for these aquatic *E. coli* populations clearly showed support for the main hypothesis that specific lineages within *E. coli* (Phylogroup A and B1) are better adapted for environmental survival. Most of the isolates from the aquatic environment grouped within this broader grouping of isolates well known for their ability to survive in the environment (White et al, 2011, Tenaillon et al. 2010, Gordon et al. 2008). The current literature suggests that although these phylogroups A and B1 are associated with the environment, they can still circulate within the human population (White et al. 2011). The phylogroups are therefore known as host generalists (White et al. 2011).

The observed structuring of the populations in the aquatic environment did not exclude the possibility that the alternative hypothesis for the origin of environmentally adapted *E. coli* could also be valid. This hypothesis states that the ability to survive in the environment may not be limited to a specific lineage but that the ability could be acquired through horizontal gene transfer. In this scenario, the genes for the survival in the secondary environment would then be carried on the accessory genome and the ability to grow in the environment will not be associated with specific phylogroups. This hypothesis is supported by the fact that the apparently unique environmental groups identified during this study are part of phylogroups B2 and D not A and B1 as expected and the

ability to survive in the secondary environment does not seem to be limited to members of phylogroups A and B1 alone.

In future the full genomes of isolates representing the unique environmental *E. coli* clusters will be sequenced to compare these genomes to those of faecal *E. coli* genomes previously sequences and available. This will be done to determine if the environmental and faecal *E. coli* contain genes that are specific to their specific environment. The function of these unique genes will then be determined to conclude whether they could confer any niche specific functions to these isolates and allow for an adaptive advantage.

3.6 REFERENCES

Akaike, H. (1974) A new look at the Statistical Model Identification. *IEEE Transactions on Automatic Control* Volume 19 Number 6 pp. 716-723.

Alm, E. W., Burke, J. and Spain, A. (2003) Fecal indicator bacteria are abundant in wet sand at freshwater beaches. *Water research* Volume 37 pp. 3978-3982.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology* Volume 215 Number 3 pp. 403-410.

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

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Sayers, E.W. (2011) GenBank. *Nucleic Acids Research* Volume 39 pp. 32-37.

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* Volume 77 Number 1 pp. 211-219.

Bruen, T.C., Philippe, H. and Bryant, D. (2006) A simple and robust statistical test for detecting the presence of recombination. *Genetics* Volume 172 pp. 2665-2681.

Bryant, D. and Moulton, V. (2004) Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. *Molecular Biology and Evolution* Volume 21 Number 2 pp. 255-265.

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* Volume 50 Number 1 pp. 27–32.

Byappanahalli, M.N., Whitman, R.L., Shively, D.A., Ting, W.T.E., Tseng, C.C. and Nevers, M.B. (2006) Seasonal persistence and population characteristics of *Escherichia coli* and enterococci in deep backshore sand of two freshwater beaches. *Journal of Water and Health* Volume 04.3 pp. 313-320.

Clement, M., Posada, D., Crandall, K.A. (2000) TCS: a computer program to estimate gene genealogies. *Molecular Ecology* Volume 9 Number 10 pp. 1657-1660.

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* Volume 5 Number 1 pp. 58-65.

Contreras, C.A., Ochoa, T.J., Ruiz, J., Lacher, D.W., Rivera, F.P., Saenz, Y., Chea-Woo, E., Zavaleta, N., Gill, A.I., Lanata, C.F., Huicho, L., Maves,

R.C., Torres, C., DebRoy, C. and Cleary, TG. (2011) Phylogenetic relationships of Shiga toxin-producing *Escherichia coli* isolated from Peruvian children. *Journal of Medical Microbiology* 60 pp. 639–646.

Drummond, A.J., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Duran, C., Field, M., Heled, J., Kearse, M., Markowitz, S., Moir, R., Stones-Havas, S., Sturrock, S., Thierer, T., Wilson, A. (2010) Geneious, v. 4.8.5 Geneious, Auckland, New Zealand.

Duriez, P., Clermont, O., Bonacorsi, S., Bingen, E., Chaventre, A. and Elion, J. (2001) Commensal *Escherichia coli* isolates are phylogenetically distributed among geographically distinct human populations. *Microbiology* Volume 147 pp. 1671-1676.

Escobar-Páramo, P., Clermont, O., Blanc-Potard, A., Bui, H., Le Bouguénec, C. and Denamur, E. (2004a) A specific genetic background is required for acquisition and expression of virulence factors in *Escherichia coli. Molecular Biology and Evolution* Volume 21 Number 6 pp. 1085-1094.

Escobar-Páramo, P., Grenet, K., Le Menac'h, A., Rode, L., Salgado, E., Amorin, C., Gouriou, S., Picard, B., Rahimy, M.C., Andremont, A., Denamur, E. and Ruimy, R. (2004b) Large-scale population structure of human commensal *Escherichia coli* isolates. *Applied and Environmental Microbiology* Volume 70 Number 9 pp. 5698-5700.

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

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

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

Gordon, D.M. and Cowling, A. (2003) The distribution and genetic structure of *Escherichia coli* in Australian vertebrates: host and geographic effects. *Microbiology* Volume 149 pp. 3575-3586.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W. and Gascuel, O. (2010) New algorithms and methods to estimate Maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systemic Biology* Volume 59 Number 3 pp. 307-321.

Hsu, L., Jackowski, S. and Rock, C.O. (1989) Uptake and acylation of 2-Acyl-lysophospholipids by *Escherichia coli. Journal of Bacteriology* Volume 171 Number 2 pp. 1203-1205.

Hudson, R. (1983) Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* Volume 23 pp. 183–201.

Huson, D.H. (1998) SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* Volume 14 Number 1 pp. 68-73.

Huson, D.H. and Bryant, D. (2006) Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* Volume 23 Number 2 pp. 254-267.

Inkscape (2011) Inkscape: Draw freely, viewed 2013 < http://inkscape.org/>

Ishii, S., Ksoll, W.B., Hicks, R.E. and Sadowsky, M.J. (2006) Presence and growth of naturalized *Escherichia coli* in temperate soils from Lake Superior Watersheds. *Applied and Environmental Microbiology* Volume 72 Number 1 pp. 612–621.

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

Kimura, M. (1981) Estimation of evolutionary distances between homologous nucleotide sequences. *The Proceedings of the Natural Academy of Sciences USA* Volume 78 Number pp. 454-458.

Kingman, J. (1982) The coalescent. *Stochastic Processes and their applications.* Volume 13 pp. 235-248.

Lacher, D.W., Steinsland, H., Blank, T.E., Donnenberg, M.S. and Whittam T.S. (2007) Molecular Evolution of Typical Enteropathogenic *Escherichia coli*: Clonal Analysis by Multilocus Sequence Typing and Virulence Gene Allelic Profiling. *Journal of Bacteriology* Volume 189 Number 2 pp. 342-350.

Leimbach, A., Hacker, J. and Dobrindt, U. (2013) *E. coli* as an all-rounder: The thin line between commensalism and pathogenicity. *Current Topics in Microbiology and Immunology* Volume 358 pp. 3-32.

Luo, C., Walk, S.T., Gordon, D.M., Feldgarden, M. and Tiedje, J.M. (2011) Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *The Proceedings of the Natural Academy of Sciences USA* Volume 108 Number17 pp. 7200-7205.

Méric, G., Kemsley, E.K., Falush, D., Saggersm E.J. and Lucchini, S. (2013) Phylogenetic distribution of traits associated with plant colonization in *Escherichia coli. Environmental Microbiology* Volume 15 Number 2 pp. 487-501.

Michigan State University (2006) Department of Microbiology and Molecular genetics. <www.shigatox.net>

Oh, S., Buddenborg, S., Yoder-Himes, D.R., Tiedje, J.M. and Konstantinidis, K.T. (2012) Genome diversity of *Escherichia* isolates from diverse habitats. *Public Library of Science* Volume 7 Issue 10 pp. e47005-e47005.

Posada, D. (2008) jModelTest: Phylogenetic Model Averaging. *Molecular Biology and Evolution* Volume 25 Number 7 pp. 1253-1256.

Posada, D. (2003) Selecting a model of nucleotide substitution. *Current Protocols in Bioinformatics* pp. 6.5.1-6.5.14. John Wiley & Sons, Inc., New York.

Posada, D. and Buckley, T.R. (2004) Model selection and model averaging in phylogenetics: Advantages of Akaike Information Criterion and Bayesian approaches over likelihood ratio tests. *Society of Systematic Biologists* Volume 53 Number 5 pp. 793-808.

Posada, D. and Crandall, K.A. (2001) Intraspecific gene genealogies: trees grafting into networks. *TRENDS in Ecology and Evolution* Volume 16 Number 1 pp. 37-45.

Prunier, A-L. and Leclercq, R. (2005) Role of *mutS* and *mutL* genes in hypermutability and recombination in *Staphylococcus aureus. Journal of Bacteriology* Volume 187 Number 10 pp. 3455-3464.

Savageau, M.A. (1983) *Escherichia coli* habitats, cell types and molecular mechanisms of gene control. *The American Society of Naturalists* Volume 122 Number 6 pp. 732-744.

Sheludchenko, M.S. Huygens, F., Hargreaves, M.H. (2010) Highly discriminatory single-nucleotide polymorphism interrogation of *Escherichia coli* by use of allele specific Real-Time PCR and eBurst analysis. *Applied and Environmental Microbiology* Volume 76 Number 13 pp. 4337-4345.

Tamura, K. and Nei, M. (1993) Estimation of the number of nucleotide substitutions in the control region of Mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution* Volume 10 Number 3 pp. 512-526.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution* Volume 28 Number 10 pp. 2731-2739.

Tarr, C.L., Large, T.M., Moeller, C.L., Lacher, D.W., Tarr, P.I., Acheson, D.W. and Whittam, T.S. (2002) Molecular characterization of a serotype O121:H19 clone, a distinct Shiga toxin-producing clone of pathogenic *Escherichia coli. Infection and Immunity* Volume 70 Number 12 pp. 6853-6859.

Templeton, A.R., Crandall, K.A. and Sing, C.F. (1992) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and dna sequence data. III. Cladogram estimation. *Genetics* Volume 132 pp. 619-633.

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

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

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

Walk, S.T., Alm, E.W., Gordon, D.M., Ram, J.R., Toranzos, G. A., Tiedje, J.M. and Whittam, T. S. (2009) Cryptic Lineages of the Genus *Escherichia. Applied and Environmental Microbiology* Volume 75 Number 20 pp. 6534–6544.

White, A.P., Sibley, K.A., Sibley, C.D., Wasmuth, J.D., Schaefer, R., Surette, M.G., Edge, T.A. and Neumann, N.F. (2011) Intergenic sequence comparison of *Escherichia coli* isolates reveals lifestyle adaptations but not host specificity. *Applied and Environmental Microbiology* Volume 77 Number 21 pp. 7620-7632.

Whitman, R.L. and Nevers, M.B. (2003) Foreshore sand as a source of *Escherichia coli* in nearshore water of a Lake Michigan Beach. *Applied and Environmental Microbiology* Volume 69 Number 9 pp. 5555–5562.

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

Winfield, M.D. and Groisman, E.A. (2004) Phenotypic differences between Salmonella and *Escherichia coli* resulting from the disparate regulation of homologous genes. *The Proceedings of the Natural Academy of Sciences* Volume 101 Number 49 pp. 17162-17167.

Xia, X. and Xie, Z. (2001) DAMBE: Software package for data analysis in molecular biology and evolution. *The Journal of Heredity* Volume 92 Number 4 pp. 371-373.

Table 3.1: Indicates the isolate names and the sample types and points

Isolate name	Sample type	Sample point
TS1B, TS2A, TS2B, TS3A, TS3B,TS4B, TS5A, TS5B, TS6A, TS6B, TS7A, TS7B, TS8B, TS9A, TS10A, TS11A, TS12B, TS13A, TS14B, TS15A, TS15B, TS16A, TS17A, TS17B, TS18A	Dam water	Rietvlei Dam
S21.1G, S21.2G, S21.3G, S21.4G, S21.5G, S21.6G, S21.7G, S21.8G, S21.9G, S21.10G, S21.11G, S21.13G, S21.14G, S226Y, S2F1.1G, S2F1.2G, S2F1.3G, S2F1.4G, S2F1.5G, S2F1.6G, S2F1.7G, S2F1.8G, S2F1.11G, S2F1.11G, S2F1.12G, S2F1.13G, S2F1.14G, S2F1.15G, S2F2.1G, S2F2.3G, S3F2.1G, S3F2.2G, S3F2.3G	Sediment	Rietvlei Dam
DWWF1.2G, DWWF1.3G, DWWF1.4G, DWWF2.1G, DWWF2.2G, DWWF2.3G, DWWF2.4G, DWWF2.5G, DWWF2.6G, DWWF2.7G, DWWF2.8G, DWWF2.9G, DWWF2.10G, DWWF2.11Y/G, DWPD5.1, DWPD5.2, DWPD5.3, DWPD5.4, DWPD5.5, DWPD5.6, DWPD5.7, DWPD5.8, DWPD5.9, DWPD5.10	Plant debris	Rietvlei Dam
Raw4.5G, Raw4.6G, Raw4.8G2, Raw4.9G2, FINAL2.1G, FINAL2.2G, FINAL2.3G, FINAL2.5G, FINAL2.6G, FINAL2.7G, FINAL2.8G, FINAL2.8G2, FINAL2.9G, FINAL2.10G, FINAL2.15Y	Sewage	Hartebeesfontein sewerage works
Ep141, Ep153, Ep154, Ep168, Sp004, Sp1000, NP720, NP707, NP705	Drinking water	Drinking water distribution networks
TomB1G, TomB3G, TomB7G, TomB8G, TomB9G, TomB10G, TomB11G, TomB12G, TomB13G, TomB14G, TomB15G, TomBF1.1G, TomBF1.2G, TomBF2.4G, TomCF1.1G, TomCF1.2G, TomCF2.3G, KBCT1.1G, KBCT1.2G, KBCTF1.1G, KBCTF1.2G, NCTR1, NCTR2, NCTR3, NCTR4, NCTR5, NCTR6, NCTR7, 1NKBKT, 2NKBKT, NKBKT2, NKBKT3, NKBKT4, NKBKT5, NKBKT6	Compost tea	
CTWF1.1, CTWF1.2, CTWF1.3, CTWF1.4, CTWF1.5, CTWF2.6	Compost tea water	

Table 3.1 continued:

Isolate name	Sample type	Sample point
WC1.1, WC1.2, WC1.3, WC1.4, WC1.5, WC1.6, WC1.7, WC2.8, WC2.9, WCF1.10, WCF2.11, BeM1.1, BeMF1.9, BeMF2.2, BeMF2.3, BeMF2.4, BeMF2.5, BeMF2.6, BeMF2.7, BeMF2.8, C2d3.9, C2d3.10, C2d3.11, C2d3.12, C2d3.13, C2d3.14, C2d3.15, C2d2.16, C2d2.17, C2d2.18, C2dF1.1, C2dF2.2, C2dF2.3, C2dF2.4, C2dF2.5, C2d5.6, C2d5.7, C2d5.8, BoM1.1, BoM1.2, BoMF1.5	Components of compost tea	
LKD3.1, L1F1.2, L1F1.3, L1F2.1, L1F2.3	Plant	Leeukraal Dam
RVD1.1, RVD2.2, RVD3.1, RVD4.1, RVD4.2, RVD5.2, RV1F1.1, RV1F1.2, RV1F1.3, RV1F2.1, RV1F2.2, RV1F2.3	Plant	Rietvlei Dam
HBPD1.1, HBPD1.2, HBPD2.1, HBPD5.2, H2F1.2, H2F2.2, H2F2.3, H3F1.1, H3F2.1	Plant	Hartebeespoort Dam
KVD1.1, KVD1.2, KVD3.1, KVD5.1, KVD5.2, K1F1.2, K1F1.3	Plant	Klipvoor Dam
BAD1.1, BAD1.2, BAD2.1, BAD2.2, BAD3.1, BAD3.2, BAD4.1, BAD4.2, BAD5.1, BAD5.2, BA1F1.2, BA1F2.1, BA1F2.2, BA1F2.3	Plant	Bon Accord Dam
RKD1.1, RKD1.3, RKD2.2, RKD3.1	Plant	Roodekopjies Dam
R1F1.1, R1F1.2, R1F1.3, R1F1.5, R1F2.1, R1F2.5, R1S1.1, R1S1.2, R1S1.3, R2F1.2, R2F1.3, R2F1.4	Plant	Roodeplaat Dam
B1F1.1, B1F1.2, B1F1.3, B1F2.2, B1F2.3	Plant	Buffelspoort Dam
Q098, Q021	Dam water	Roodeplaat Dam
ZA2.4, ZA1.8, ZA2.7, ZB2.9, ZB2.7, ZB1.2	Sewage	Zeekoegat sewage treatment works
B1.2	Sewage	Baviaanspoort sewage treatment works
Q02H1, Q02H2, Q02H3, Q02H4, Q02H5, Q02H6, Q02H8, Q02H9, Q02H10, Q02H11, Q02H12, Q02H13, Q02H15	Water hyacinth	Roodeplaat Dam
SW7FE8, SW9FE12, SW11FE16, Spring2FE4	Surface water	Kusile power station construction site.

Table 3.2: List of the plant species sampled from the 8 dams

Plant sample	Dam	Common name	Scientific name
RKD 3	Roodekopjies	Water hornwort	*Ceratophyllum demersum*
RKD 2	Roodekopjies	Parrot's feather	*Myriophyllum aqauticum*
RKD 1	Roodekopjies	Parrot's feather	*Myriophyllum aqauticum*
BAD 5	Bon Accord	Parrot's feather	*Myriophyllum aqauticum*
BAD 4	Bon Accord	Parrot's feather	*Myriophyllum aqauticum*
BAD 3	Bon Accord	Parrot's feather	*Myriophyllum aqauticum*
BAD 2	Bon Accord	Parrot's feather	*Myriophyllum aqauticum*
BAD 1	Bon Accord	Parrot's feather	*Myriophyllum aqauticum*
KVD 5	Klipvoor	Parrot's feather	*Myriophyllum aqauticum*
KVD 3	Klipvoor	Parrot's feather	*Myriophyllum aqauticum*
KVD 2	Klipvoor	Water hyacinth	*Eichhornia crassipes*
KVD 1	Klipvoor	Parrot's feather	*Myriophyllum aqauticum*
HBPD 5	Hartebeespoort	Water weed	*Isolepis fluitans*
HBPD 2	Hartebeespoort	Water hornwort	*Ceratophyllum demersum*
HBPD 1	Hartebeespoort	Water hornwort	*Ceratophyllum demersum*
RVD 5	Rietvlei	Parrot's feather	*Myriophyllum aqauticum*
RVD 4	Rietvlei	Water weed	*Isolepis fluitans*
RVD 3	Rietvlei	Parrot's feather	*Myriophyllum aqauticum*
RVD 2	Rietvlei	Parrot's feather	*Myriophyllum aqauticum*
RVD 1	Rietvlei	Water weed	*Isolepis fluitans*
LKD 3	Leeukraal	Water weed	*Isolepis fluitans*
H2	Hartebeespoort	Water hornwort	*Ceratophyllum demersum*
R1	Roodeplaat	Water weed	*Isolepis fluitans*
RV1	Rietvlei	Water weed	*Isolepis fluitans*
H3	Hartebeespoort	Water weed	*Isolepis fluitans*
BA1	Bon Accord	Water hornwort	*Ceratophyllum demersum*
R2	Roodeplaat	Parrot's feather	*Myriophyllum aqauticum*
K1	Klipvoor	Parrot's feather	*Myriophyllum aqauticum*
L1	Leeukraal	Water weed	*Isolepis fluitans*
B1	Buffelspoort	Parrot's feather	*Myriophyllum aqauticum*

Table 3.3: List of primers used for amplification and sequencing of variable genes

Primer name	Sequence (5'-3')
*mutS	GGC CTA TAC CCT GAA CTA CA
mutS	GCA TAA AGG CAA TGG TGT C
*fadD	GCT GCC GCT GTA TCA CAT TT
fadD	GCG CAG GAA TCC TTC TTC AT

* Indicates forward primer

Table 3.4: The accession numbers associated with the isolates used:

Isolate name	Accession number	Isolate details
AF002204 *E. coli*	AF002204.1	*Escherichia coli* O55:H7 *rpoS* gene
AF002207 *E. coli*	AF002207 .1	*Escherichia coli* O157:H7 *rpoS* gene
AF002205 *E. coli*	AF002205.1	*Escherichia coli* O157:NM *rpoS* gene
AF002206 *E. coli*	AF002206 .1	*Escherichia coli* O78:K80 *rpoS* gene
AF002208 *E. coli*	AF002208.1	*Escherichia coli* O157:H7 *rpoS* gene
AF002209 *E. coli*	AF002209.1	*Escherichia coli* O157:H7 *rpoS* gene
AY698845 *E. coli*	AY698845.1	*Escherichia coli* strain 5898-71 *rpoS gene*
AY698870 *E. coli*	AY698870.1	*Escherichia coli* strain 1758-70 *rpoS* gene
AY723475 *E. coli*	AY723475.1	*Escherichia coli* strain DEC 2a *rpoS* gene
CU928158.2 *E. fergusonii*	CU928158.2	*Escherichia fergusonii* ATCC 35469 *rpoS* gene
AB454547.1 *E. coli*	AB454547.1	*Escherichia coli* CK28 *uidA* gene
AB334714 *E. coli*	AB334714.1	*Escherichia coli* 12646 *uidA* gene
AY698445 *E. coli*	AY698445.1	*Escherichia coli* LT-82 *uidA* gene
HM221006 *E. coli*	HM221006.1	*Escherichia coli* RFA16 *uidA* gene
HM221054 *E. coli*	HM221054.1	*Escherichia coli* RFA19 *uidA* gene
B692		*Escherichia coli mutS* gene Walk et al. 2009
E677		*Escherichia coli mutS* gene Walk et al. 2009
GU969899 *E. fergusonii*	GU969899.1	*Escherichia fergusonii* B691 *mutS* gene
B692		*Escherichia coli fadD* gene Walk et al. 2009
E677		*Escherichia coli fadD* gene Walk et al. 2009
GU952664 *E. fergusonii*	GU952664.1	*Escherichia fergusonii* B372 *fadD* gene
B827I		*Escherichia* Clade I strain B827
E1492I		*Escherichia* Clade I strain E1492
E807I		*Escherichia* Clade I strain E807
H442I		*Escherichia* Clade I strain H442
M863I		*Escherichia* Clade I strain M863
TW10509I		*Escherichia* Clade I strain TW10509
TW11930I		*Escherichia* Clade I strain TW11930
TW11966I		*Escherichia* Clade I strain TW11966
TW15838I		*Escherichia* Clade I strain TW15838
B1147II		*Escherichia* Clade II strain B1147
TW09231III		*Escherichia* Clade III strain TW09231
TW09276III		*Escherichia* Clade III strain TW09276
TW09254III		*Escherichia* Clade III strain TW09254
TW09266III		*Escherichia* Clade III strainTW09266
TA04III		*Escherichia* Clade III strain TA04
B685III		*Escherichia* Clade III strain B685
TW14182IV		*Escherichia* Clade IV strain TW14182
TW11588IV		*Escherichia* Clade IV strain TW11588
H605IV		*Escherichia* Clade IV strain H605
B49IV		*Escherichia* Clade IV strain B49
TW09308V		*Escherichia* Clade V strain TW09308
B1225V		*Escherichia* Clade V strain B1225
B646V		*Escherichia* Clade V strain B646
E1118V		*Escherichia* Clade V strain E1118
E1195V		*Escherichia* Clade V strain E1195

Table 3.4: Continued

E1196V		*Escherichia* Clade V strain E1196
E471V		*Escherichia* Clade V strain E471
E472V		*Escherichia* Clade V strain E472
E620V		*Escherichia* Clade V strain E620
M1108V		*Escherichia* Clade V strain M1108
TA290V		*Escherichia* Clade V strain TA290
TW14263V		*Escherichia* Clade V strain TW14263
TW14264V		*Escherichia* Clade V strain TW14264
TW14265V		*Escherichia* Clade V strain TW14265
TW14266V		*Escherichia* Clade V strain TW14266
TW14267V		*Escherichia* Clade V strain TW14267
RL325/96V		*Escherichia* Clade V strain RL325/96
Z205V		*Escherichia* Clade V strain Z205
ATCC8739A	CP000946.1	Commensal *Escherichia coli* ATCC 8739 Phylogroup A
HS A	NC_009800.1	Commensal *Escherichia coli* HS Phylogroup A
K12A	U00096.2	Commensal *Escherichia coli* K-12 MG1655 Phylogroup A
H10407A	NC_017633.1	ETEC *Escherichia coli* H10407 Phylogroup A
O26H11B1	NC_013361.1	EHEC *Escherichia coli* O26:H11 11368 Phylogroup B1
O111HB1	NC_013364.1	EHEC *Escherichia coli* O111:H- 11128 Phylogroup B1
O103H2B1	NC_013353.1	EHEC *Escherichia coli* O103:H2 12009 Phylogroup B1
55989B1	NC_011748.1	EAEC *Escherichia coli* 55989 Phylogroup B1
E24377AB1	NC_009801.1	ETEC *Escherichia coli* E24377A Phylogroup B1
SE11B1	NC_011415.1	Commensal *Escherichia coli* SE11 Phylogroup B1
IAI1B1	NC_011741.1	Commensal *Escherichia coli* IAI1 Phylogroup B1
CB9615E	NC_013941.1	EPEC *Escherichia coli* CB9615 Phylogroup E
EDL933E	NC_002655.2	EHEC *Escherichia coli* O157:H7 EDL 933 Phylogroup E
SakaiE	NC_002695.1	EHEC *Escherichia coli* O157:H7 Sakai Phylogroup E
042D1	NC_017626.1	EAEC *Escherichia coli* 042 Phylogroup D1
UMN026D1	NC_011751.1	ExPEC *Escherichia coli* UMN026 Phylogroup D1
SMS35D2	CP000970.1	ExPEC Escherichia coli SECEC SMS-3-5 Phylogroup D2 (F)
IAI39D2	NC_011750.1	ExPEC *Escherichia coli* IAI39 Phylogroup D2 (F)
CE10D2	NC_017646.1	ExPEC *Escherichia coli* CE10 Phylogroup D2 (F)
ABU83972B2	NC_017631.1	ExPEC *Escherichia coli* ABU 83972 Phylogroup B2
CFT073B2	NC_004431.1	ExPEC *Escherichia coli* CFT073 Phylogroup B2
LF82B2	NC_011993.1	AIEC *Escherichia coli* LF82 Phylogroup B2
NRG857CB2	CP001855.1	AIEC *Escherichia coli* NRG 857C Phylogroup B2
ED1aB2	NC_011745.1	Commensal *Escherichia coli* ED1a Phylogroup B2
APEC01B2	NC_008563.1	ExPEC *Escherichia coli* APEC 01 Phylogroup B2
S88B2	NC_011742.1	ExPEC *Escherichia coli* S88 Phylogroup B2
IHE3034B2	NC_017628.1	ExPEC *Escherichia coli* IHE3034 Phylogroup B2
UTI89B2	NC_007946.1	ExPEC *Escherichia coli* UTI89 Phylogroup B2
536B2	NC_008253.1	ExPEC *Escherichia coli* 536 Phylogroup B2
SE15B2	NC_013654.1	Commensal *Escherichia coli* SE15 Phylogroup B2
NA114B2	NC_017644.1	ExPEC *Escherichia coli* NA114 Phylogroup B2
E2348/69B2	NC_011601.1	EPEC *Escherichia coli* E2348/69 Phylogroup B2
S. sonnei	NC_007384.1	*Shigella sonnei* Ss046
S. boydii	NC_007613.1	*Shigella boydii* Sb227
S. flexneri	NC_004337.2	*Shigella flexneri* 2a str. 301
S. dysenteriae	NC_007606.1	*Shigella dysenteriae* Sd197

Table 3.5: The unique isolates used in the TCS and SplitsTree analysis and the isolates they represent.

Isolate name	Represents the following isolates
S21.5G	S21.5G, S21.14G
FINAL2.1G	FINAL2.1G, FINAL2.9G
KBCT2G	KBCT2G, KBCTF1.1G, KBCTF1.2G
WC1.1	WC1.1, WC1.4, WC1.5, WC1.6, WC1.7, WC2.9, WCF2.11
C2dF2.4	C2dF2.4, C2d5.7
HBPD1.1	HBPD1.1, BoM1.1
NKBKTR2	NKBKTR2, NKBKTR3, NKBKTR4, NKBKTR5, NKBKTR6
Q02H1	Q02H1, Q02H2, Q02H3, Q02H4, Q02H5, Q02H6, Q02H8, Q02H9, Q02H10, Q02H11
Q02H12	Q02H12, Q02H13
C2d3.10	C2d3.10, C2d3.11, C2d3.12, C2d3.13, C2d3.14, C2d3.15, C2d2.16, C2d2.17, C2d2.18, C2d5.6, H2F2.2
BeMF2.3	BeMF2.3, BeMF2.4, R1F1.2
R1F1.5	R1F1.5, R1F2.5
C2d3.9	C2d3.9, C2dF2.3, R2F1.4
BeMF2.2	BeMF2.2, BeMF2.8, L1F2.1
R1F2.1	R1F2.1, B1F1.2
DWPD5.2	DWPD5.2, DWPD5.3, DWPD5.4
DWPD5.9	DWPD5.9, DWPD5.10
FINAL2.8G	FINAL2.8G, FINAL2.8G2, K12A
NCTR2	NCTR2, RV1F2.1, E24377AB1
BoM1.2	BoM1.2, SE11B1
LF82B2	LF82B2, NRG857CB2
R2F1.2	R2F1.2, APEC01B2, S88B2, IHE3034B2, UTI89B2
ZA2.7	ZA2.7, 536B2
ZB2.9	ZB2.9, SE15B2
EDL933E	EDL933E, SakaiE

Table 3.6: The PHI test results for each gene and the concatenated dataset

Gene (Number of sequences)	Number of informative sites [a]	Mean [b]	Variance [c]	observed [d]	p-value [e]	k [f]	Reject or fail to reject null hypothesis of no recombination [g]
rpoS (281)	32	0.11290323	3.01E-04	0.1122449	0.4848635	7	Fail to reject
rpoS (317)	35	0.11092437	2.76E-04	0.11059908	0.4921824	7	Fail to reject
uidA(281)	53	0.24165457	2.88E-04	0.24564797	0.5929892	11	Fail to reject
uidA(317)	58	0.21415608	2.24E-04	0.20064725	0.1832954	12	Fail to reject
mutS (281)	115	0.53897788	1.26E-04	0.47952723	5.71E-08	23	Reject
mutS (317)	116	0.55217391	1.27E-04	0.49707358	5.20E-07	23	Reject
fadD (281)	62	0.19936542	1.91E-04	0.17267267	0.026861	12	Reject
fadD (317)	86	0.14774282	8.54E-05	0.12605042	0.0094522	17	Reject
Concatenated (281)	262	0.41715656	5.89E-05	0.29713424	0.0	13	Reject
Concatenated (317)	295	0.37149775	4.39E-05	0.27154472	0.0	15	Reject

a: site with at least two different character states and are found in at least two different sequences; **b**: mean of a normal probability distribution used to calculate p-value

c: variability of a normal probability distribution used to calculate p-value; **d**: value of Φ_w for the original alignment; **e**: p-value is the probability of obtaining a test result at least as extreme as the one observed

f: k=wq, w being the window size and q the proportion of invariable sites; **g**: based on the p-value the null hypothesis will be rejected or not rejected

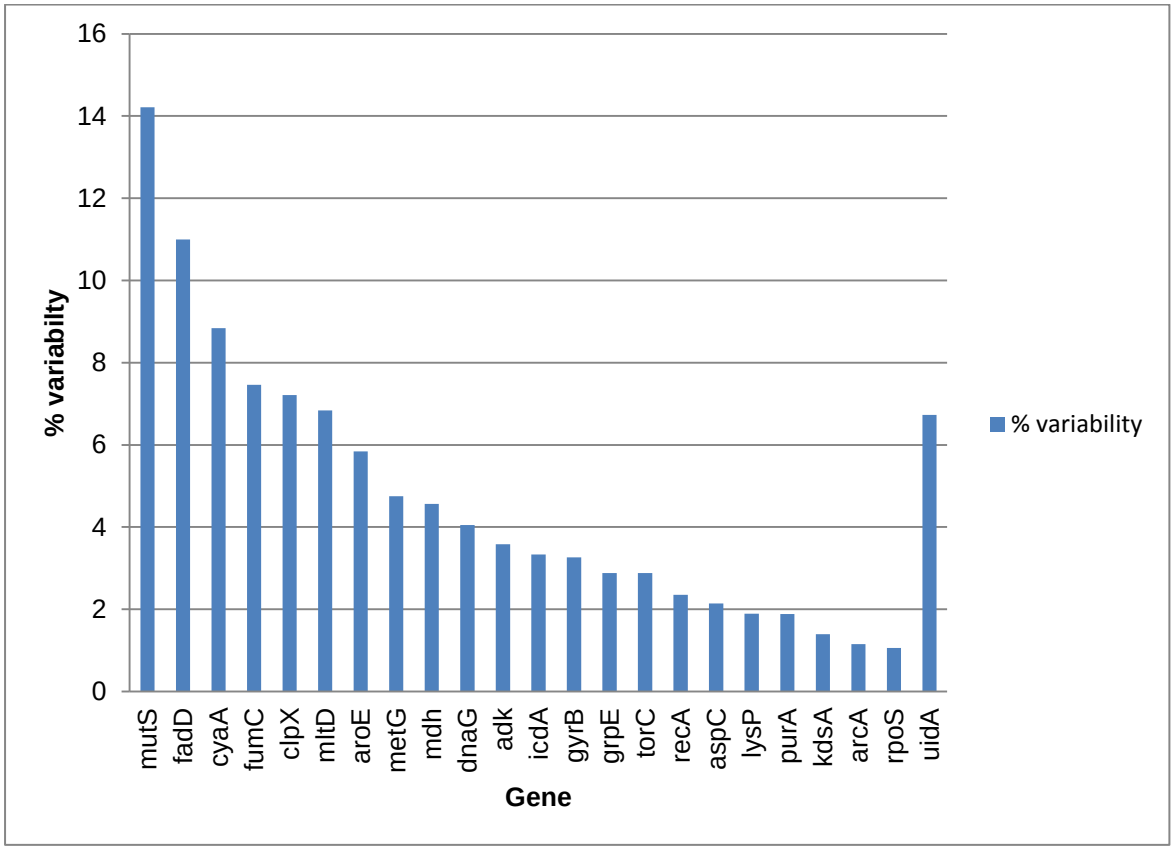

Figure 3.1: A graph depicting the percentage variability within the 22 housekeeping genes used in the Walk et al. 2009 study. The variability was calculated by number of variable sites/total number of sites*100.

Figure 3.2: The maximum likelihood tree of the *mutS* gene of the 261 isolates, the phylogroups and the clades. Bootstrap values are represented as a percentage of 1000 replicates, all bootstrap values below 50% were not included. The brackets indicate the clusters that do not contain any sewage isolates. The key indicates the colour associated with a specific sample type as well as how the phylogroups are indicated in the tree.

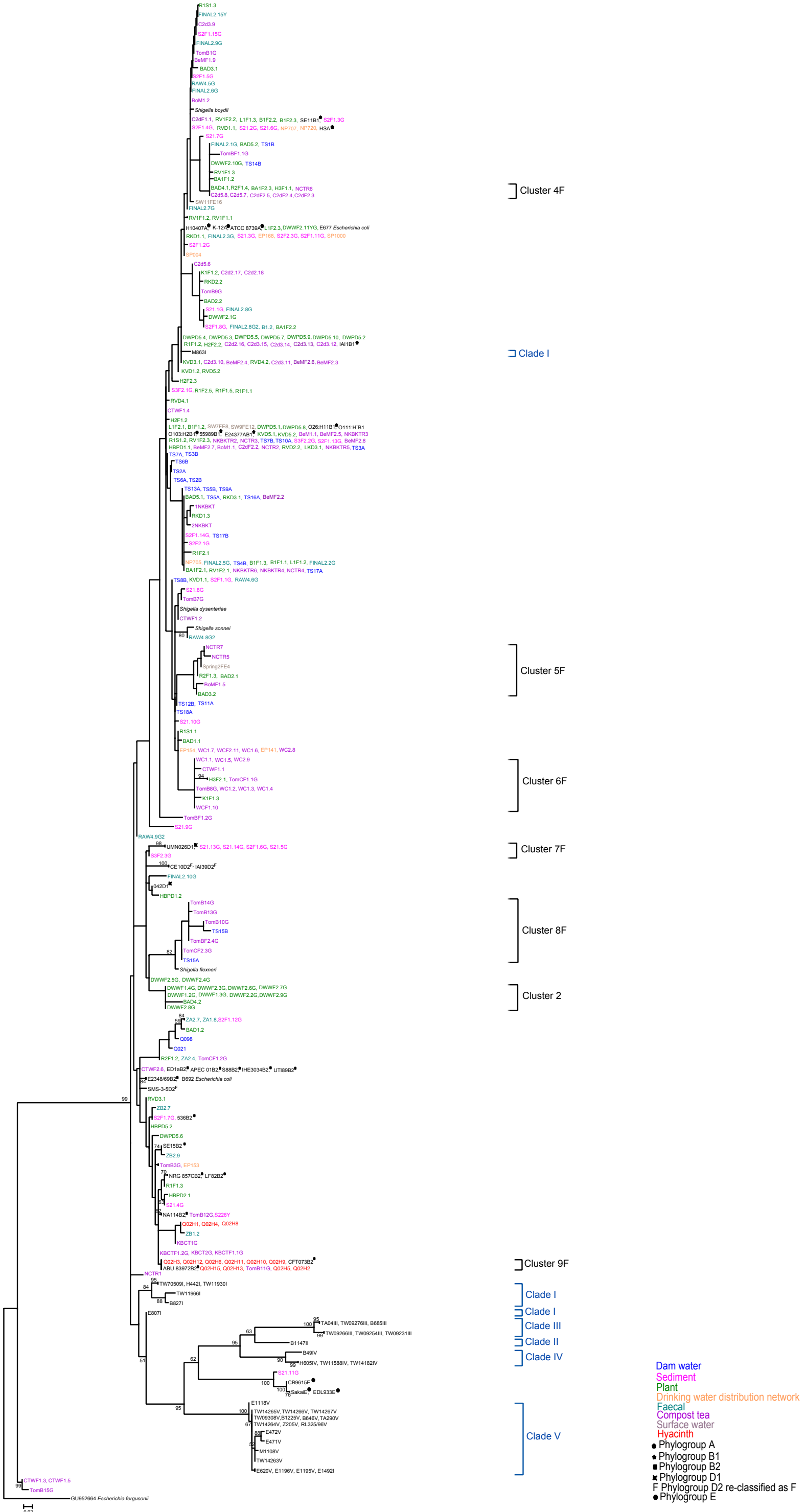

Figure 3.3: The maximum likelihood tree of the *fadD* gene of the 281 isolates, the phylogroups and the clades. Bootstrap values are represented as a percentage of 1000 replicates, all bootstrap values below 50% were not included. The brackets indicate the clusters that do not contain any sewage isolates. The key indicates the colour associated with a specific sample type as well as how the phylogroups are indicated in the tree.

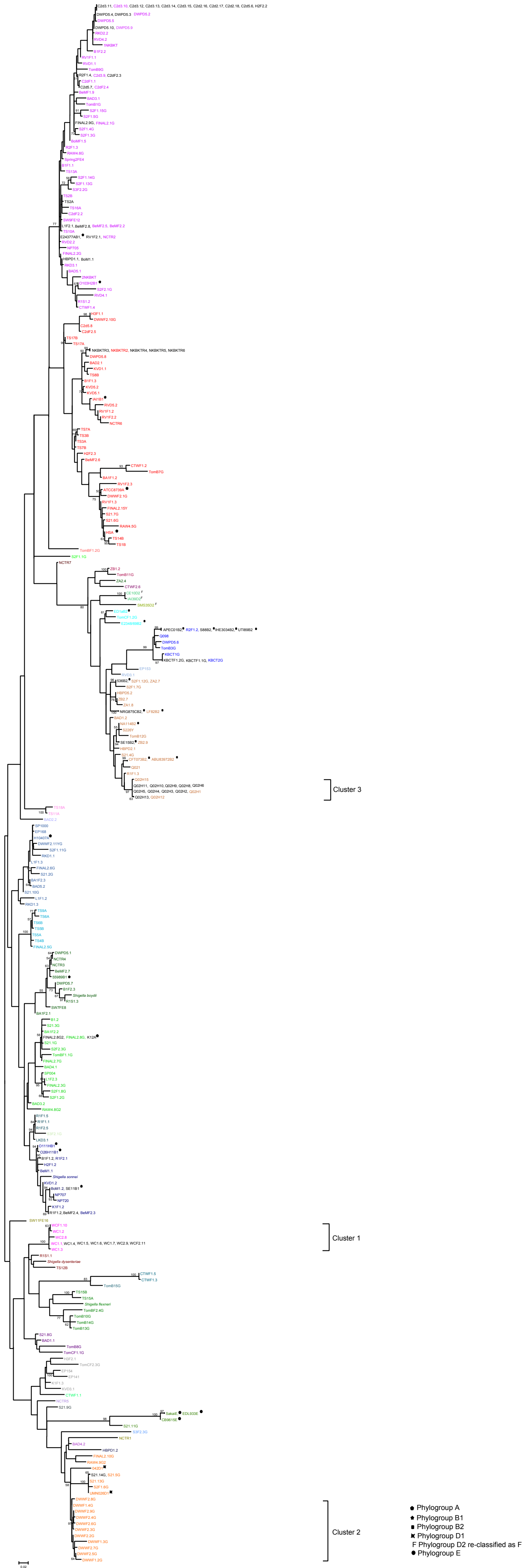

Figure 3.4: The maximum likelihood tree of the concatenated dataset of the 221 isolates from this study and the phylogroups. Bootstrap values are represented as a percentage of 1000 replicates, all bootstrap values below 50% were not included. The brackets indicate the unique clusters. The isolates were coloured according to consistent grouping within this, TCS and SplitsTree analyses. The key indicates how the phylogroups are indicated in the tree.

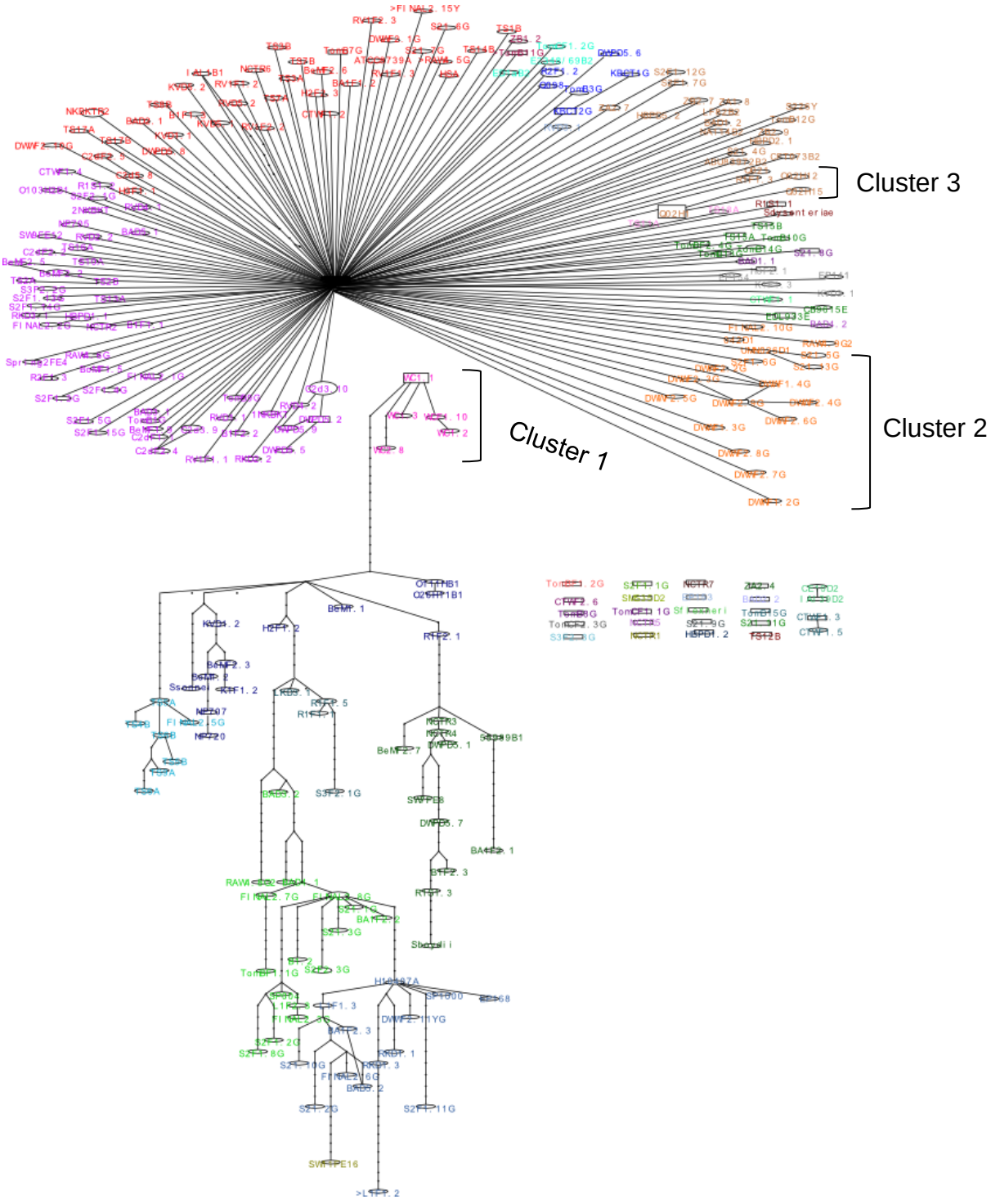

Figure 3.5: A TCS analysis using the concatenated dataset. Sequences were collapsed into haplotypes and only the first name is kept, this is indicated in Table 3.5. The brackets indicate the unique clusters. The isolates were coloured according to consistent groupings within this, phylogenetic and SplitsTree analyses.

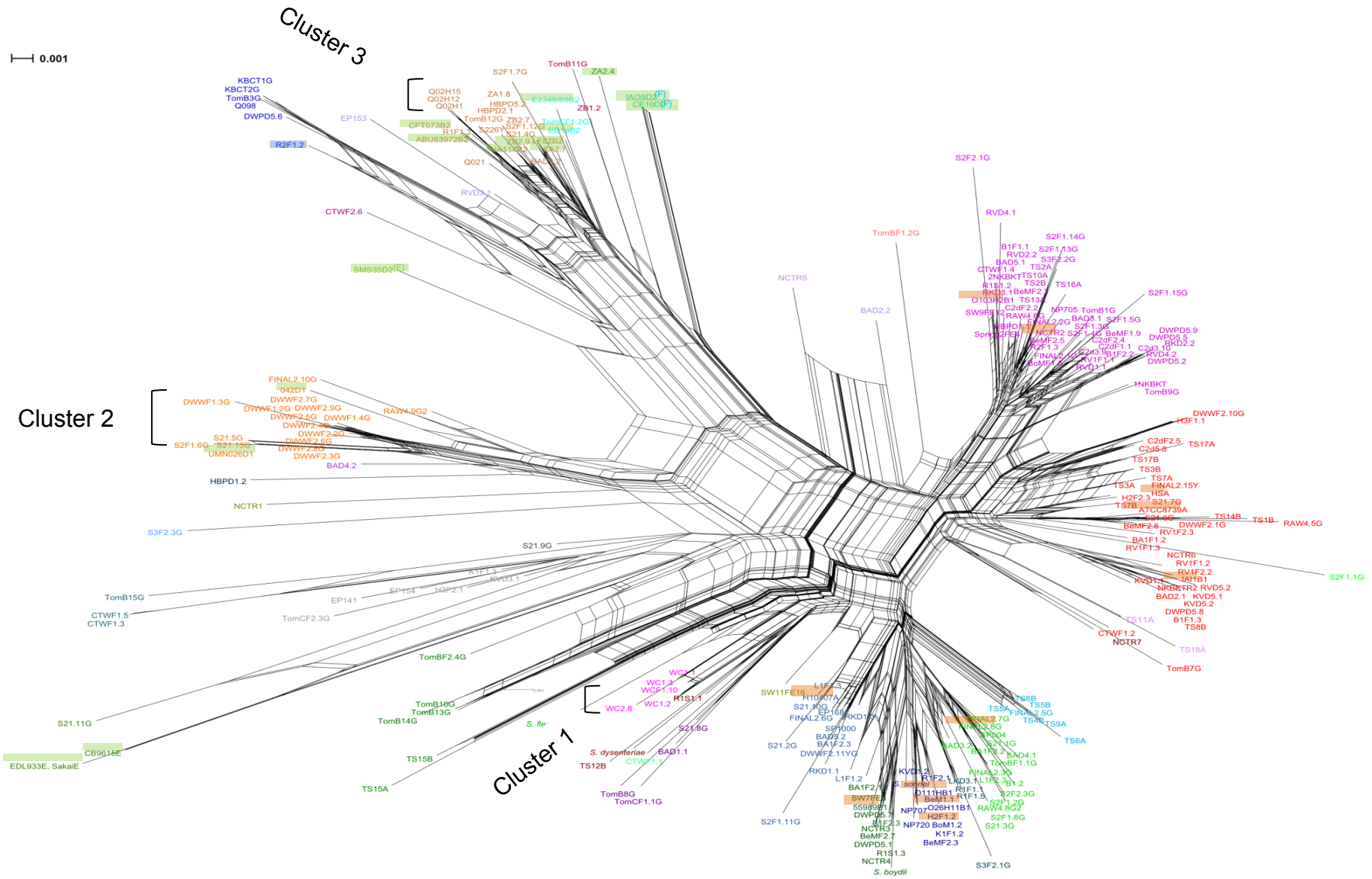

Figure 3.6: A neighbour-net SplitsTree analysis using the concatenated dataset. The brackets indicate the unique clusters. The isolates with the background colour orange are phylogroups A and B1, those coloured by a green background are part of the other phylogroups. The isolates were coloured according to consistent groupings within this, phylogenetic and TCS analyses.

SUMMARY

Escherichia coli (*E. coli*) is an indicator of faecal contamination as it is assumed that faecal contamination is the main source of these bacteria in the environment. Recent studies have, however, shown that *E. coli* can be found in the environment without any apparent link to faecal contamination. These environmental *E. coli* isolates multiply and survive in niches including soil, sand, sediment and water. Environmental *E. coli* are usually associated with phylogroups A and B1, two of the 7 phylogroups (A, B1, B2, C, D, E and F) typically used to group *E. coli* isolates. Some environmental isolates have also been linked to Clades III-V, which are novel undescribed *Escherichia* species. In this study *E. coli* was isolated from different niches within various freshwater dams. To represent the *E. coli* circulating in the human population, *E. coli* was isolated from sewage samples. To determine the diversity within the *Escherichia coli* population and to identify possible environmental *E. coli*, the sigma factor S (*rpoS*), β-D-glucuronidase (*uidA*), methyl-directed mismatch repair (*mutS*) and fatty acyl-CoA synthetase (*fadD*) genes were sequenced and Maximum Likelihood trees were drawn using the individual and concatenated datasets. The phylogenetic trees were also used to determine which phylogroup the isolates are associated with and if any of the isolates belonged to the undescribed species or clades. The population dynamics was determined using TCS and SplitsTree analysis. The phylogenetic trees showed that the diversity amongst these isolates was high. None of the isolates were part of the clades and most of the isolates group with phylogroups A and B1 as expected. Three possible unique clusters of environmental isolates were observed which respectively formed part of phylogroups B2, D and no specific phylogroup. Phylogroups B2 and D are usually associated with isolates that cause extra-intestinal infection and was not expected to be represented by environmental isolates. Population structure analyses indicated that these clusters could be part of sub-populations within the larger *E. coli* population and may be genetically separate from the rest of the isolates.