May 2016

The Selective Survival of Escherichia Coli in Freshwater Beach Sand

Natalie Ann Rumball
University of Wisconsin-Milwaukee

THE SELECTIVE SURVIVAL OF *ESCHERICHIA COLI* IN FRESHWATER BEACH SAND

by

Natalie Ann Rumball

A Dissertation Submitted in

Partial Fulfilment of the

Requirements for the Degree of

Doctor of Philosophy

in Freshwater Sciences

at

The University of Wisconsin-Milwaukee

May 2021

ABSTRACT

THE SELECTIVE SURVIVAL OF *ESCHERICHIA COLI* IN FRESHWATER BEACH SAND

by

Natalie Ann Rumball

The University of Wisconsin-Milwaukee, 2021
Under the Supervision of Professor Sandra McLellan

The quantification of *Escherichia coli* or *E. coli* is the most common method used to detect recent fecal pollution in recreational water, as this species is known for its high abundance in fecal matter and assumed host-associated nature. However, it has been determined that some strains are capable of long-term survival and potential propagation in non-host environments, such as the beach sand. These long-term environmental survivors are host-independent and are not associated with the same health risks as those *E. coli* from recent fecal pollution. However, they have been shown to impact how water quality is perceived as they are reintroduced into the water column by wave action and are counted in monitoring efforts. Current monitoring methods are unable to differentiate long-term surviving populations from populations originating from recent fecal pollution. Despite this known discrepancy *E. coli* enumeration is still relied upon to estimate levels of fecal pollution and used to determine the need for beach closures. The aim of this work was to identify genetic indicators of long-term survival that can be used to develop tools to improve beach monitoring. *E. coli* capable of long-term environmental survival were identified through a series of microcosm experiments, in which populations from sand, sewage, and gull waste (n=198 each) were seeded into sand treatments (unaltered native sand, nutrient-limited baked sand, and nutrient-abundant autoclaved sand) and buried 0.5 m deep in the backshore of Lake Michigan for 6-8 weeks. The populations were monitored over the course of

the study, and those capable of environmental survival increased in frequency by the end of the experiment. Survival-associated genes were identified through a novel population genetics approach in which composite samples from each source and timepoint were shot-gun sequenced and mapped to a scaffold of *E. coli* accessory genes from 21 genomes. Genes that had >25% higher depth of coverage in output populations compared to those from the input were considered enriched in long-term surviving populations. It was determined that *E. coli* from each source tested were capable of long-term survival in beach sand, with the ability to survive varying based on phylotype association and accessory gene ownership. Through Clermont phylotyping it was determined that members of A and B1 increased in frequency by the end of the experiment, suggesting that members of these groups may be better suited for survival in secondary environments. Overall, there were a total of 198 survival-associated functions shared among each sand, sewage, and gull surviving populations, which were largely associated with metabolism enzymes and transport proteins. Several pathway modules were identified in these surviving populations, including the betaine biosynthesis pathway, which allows the production of compatible solutes that prevent dehydration, and the GABA biosynthesis and the GABA shunt modules, which are associated with flexibility in nutrient utilization. Overall, the distribution of these survival related functions were shown to vary, with some being more widely distributed (i.e., among non-clade members), while others were more narrowly distributed among members of select phylogroups (A/B1/cryptic clades), demonstrating that survivability varies based on accessory gene ownership and phylotype association.

To

my parents for instilling a love for learning in me and

my husband for supporting me while I follow my dreams

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBRVIATIONS

CFU	Colony Forming Units
CN	Copy Number
DNA	Deoxyribonucleic acid
E. coli	*Escherichia coli*
FIB	Fecal Indicator Bacteria
MLEE	Multilocus Enzyme Electrophoresis
MLST	Multilocus Sequence Typing
OC	Organic Carbon
ON	Organic Nitrogen
PCR	Polymerase Chain Reaction
qPCR	Quantitative Polymerase Chain Reaction
QMRA	Qualitative Microbial Risk Assessment
TN	Total Nitrogen
TOC	Total Organic Carbon
TP	Total Phosphorus
USEPA	United States Environmental Protection Agency
RWQC	U.S. Recreational Water Quality Criteria

CHAPTER 1
Introduction

The historical use of *E. coli* as a fecal indicator

It has been estimated that swimming in fecal contaminated beach water causes over 120 million cases of gastrointestinal disease and 50 million cases of acute respiratory diseases world-wide yearly (Shuval, 2003). In effort to prevent exposure to conditions that would cause such water related illnesses the US developed guidelines for recreational water quality. These guidelines are laid out in the Beaches Environmental Assessment and Coastal Health Act (BEACH) of 2000 and the U.S. Clean Water Act (CWA) of 1972 (US EPA, 2000), which mandate that beach water is monitored for the presence of pathogens associated with fecal contamination and that the public is notified when the water does not meet these quality standards. Current monitoring efforts enforced by these acts often rely upon the quantification of single organisms in recreational waters to act as indicators of fecal pollution. One of the most common indicators used is the bacterial species *Escherichia coli* or *E. coli*.

E. coli is a gram-negative facultative anaerobic bacterial species that belongs to the Enterobacteriaceae family (Blount, 2015). Its primary habitat is within the biofilms that line the mucus layer of the gut (Chang et al., 2004) and is shed into host-independent secondary habitats, such as water, sediment, and soil, at a density of $10^6 – 10^9$ cells per gram of fecal matter (Savageau, 1983). Exposure to such fecal matter is considered a public health risk, as it is associated with a multitude of pathogens that can result in communicable diseases and illness (Arnold et al., 2016; Graciaa et al., 2018; Korajkic, Brownell, & Harwood, 2011). In effort to prevent exposure to fecal pollution the EPA has recommended the use of *E. coli* as a fecal indicator in recreational waters (US EPA, 1986), with a recommended daily maximum limit of

235 CFU/100 ml for single samples or a geometric mean of 126 CFU/100 ml for samples collected over the past 30 days (US EPA, 2012).

Current monitoring methods rely upon the culture detection of *E. coli* with no additional consideration other than its concentration. This monitoring is conducted using either direct counts or a most probable number estimation. Common methods include direct enumeration of cells through plate counts, the detection of species-specific DNA markers through qPCR, and most probable number techniques that use the prevalence of species-specific metabolic identifiers, such as the IDEXX chromogenic test or multiple tube fermentation tests (Noble et al., 2003). When these tools detect *E. coli* levels higher than the recommended cutoff values that beach is then assumed to be impacted by fecal pollution and is temporarily closed until *E. coli* levels decline. These methods, though efficient at determining total concentrations of *E. coli*, fall short as they do not address the contribution of host-independent populations from these environments, which can result in an over-estimation of recent fecal pollution.

E. coli have a much more complicated set of potential outcomes in open environments than current monitoring methods account for, which rely on the assumption that all *E. coli* decay quickly in open environments. For instance, though *E. coli* is best known for its primary habitat, the gut of warm-blooded organisms, it is also known to establish host independent populations in open environments such as water, sediment, or soil (Savageau, 1983). Here *E. coli* are subjected to a variety of environmental and biotic stresses, where an isolate can either enter a new host, parish, or shift to long-term survival strategies. Evidence suggests that not all strains of *E. coli* react to these selection pressures equally, with certain phylotype associations and the ownership of certain accessory genes suggested to increase the likelihood of environmental survival. The

presence of these populations complicates current monitoring efforts, as they cannot distinguish those *E. coli* that are a part of recent fecal pollution from those that are part of the secondary environment population. This project addresses this short coming and aims to address genetic differences in these populations, adding to the body of research on bacterial survival in secondary environments.

E. coli in freshwater beach sand and the impact on perceived water quality

The shoreline is a highly complex environment. It is at the intersection of both terrestrial and aquatic ecosystems and is subjected to influence from both environments. The berm or swash zone of the shoreline, where the beach is within the range of wave action, is home to a highly heterogeneous microbial population (Strayer & Findlay, 2010) that has been shown to have a higher level of taxonomic richness and alpha diversity than populations in paired water samples (Halliday, McLellan, Amaral-Zettler, Sogin, & Gast, 2014; Mohiuddin, Salama, Schellhorn, & Golding, 2017). These highly diverse microbial populations are a consequence of this ecosystem's dynamic influences including fluctuating water chemistry, UV light exposure, wave intensity, and nutrient availability. Species that can survive here, such as *E. coli*, must be highly adaptable and able to overcome many selection pressures.

E. coli have been shown to overcome these selection pressures and have been demonstrated to be constantly present in freshwater beach sand. This was observed by several research groups across the region, with it being detected throughout the year in each season (Alm, Burke, & Hagan, 2006; Hartz et al., 2008; R. L. Whitman & Nevers, 2003). These populations have been consistently observed to be 10 - 100X higher in sand samples compared to paired water samples (Alm, Burke, & Spain, 2003; Staley, Vogel, Robinson, & Edge, 2015; R. L.

Whitman & Nevers, 2003), suggesting that the sand is forming a reservoir for *E. coli* where they can congregate and potentially propagate, becoming members of the autochthonous, or native microbial community in sand.

In addition to observational studies, *E. coli* has also been demonstrated to be capable of environmental survival through several experiments. One common method used has been microcosms or growth chambers that allow survival to be observed *In vivo* or in the environment. In these experiments *E. coli* was subjected to various treatments, sealed in microcosms and buried at the beach for extended periods of time (weeks to months), with the population dynamics being observed throughout deployment. Using this design, it was determined that *E. coli* is better at survival in beach sand then in water, as observed by slower decay rate in sand (Hartz et al., 2008). This enhanced ability to survive and form populations was observed through an additional microcosm study that showed *E. coli* survival under environmental conditions for over 48 days at Lake Huron beaches in autoclaved sand (Alm et al., 2006), with cellular replication in beach sand observed under laboratory observations (Hartz et al., 2008). Overall, the results of these studies suggest that *E. coli* is capable of being a stable member of the microbial community in beach sand, with both survival and replication observed in this habitat. These populations are subjected to many uncontrollable environmental conditions including sand erosion, which acts as the main transport of these organisms into the surrounding aquatic ecosystem (Vogel, O'Carroll, Edge, & Robinson, 2016).

Wave action and rain are the major modes of bacterial transport form sand to water. These events increase the levels of *E. coli* in adjacent water and effect the accuracy of monitoring efforts that aim to detect recent fecal pollution. The analysis of *E. coli* in paired water

and sand samples collected over the summer of 2000 at the 63rd street beach in Chicago demonstrated wind speed to be the major contributing factor in determining the relationship between *E. coli* concentrations in sand and water (R. L. Whitman & Nevers, 2003). Similar results were found in a 2016 study that demonstrated that the majority of *E. coli* transport was occurring through particle attachment rather than through pore water (Vogel et al., 2016). The force of heavy rain causes a similar event called a sand washout, which is the transport of sand into water through run-off following heavy rain, with *E. coli* levels demonstrated to increase nearly 100-fold after 30 minutes of rain at a beach on Lake Michigan in Milwaukee (Beversdorf, Bornstein-Forst, & McLellan, 2007). Despite evidence demonstrating that these sand-associated populations can impact the perception of local water quality, current beach monitoring techniques do not account for those *E. coli* that have accumulated and propagated in beach sand. The accuracy of these methods can be improved with an enhanced understanding of *E. coli* survival in the beach sand, with more work needing to be conducted on determining the genetic factors that allow for *E. coli* survival in secondary environments.

Potential genetic determinants of survival

E. coli is an organism that is widely distributed across many environments. Its adaptability is due to the large number of genetic variances within the species, including a wide range of phylogenic groups and an extensive accessory genome. Previous work done has shown that certain strains of *E. coli* are associated with the beach sand environment. DNA microarray analysis of *E. coli* populations isolated from beach sand and water demonstrated that isogenic populations of *E. coli* were found in beach sand, but not in water (Kon et al., 2007; Oh, Buddenborg, Yoder-Himes, Tiedje, & Konstantinidis, 2012), suggesting that specific strains of *E. coli* are able to survive and propagate in beach sand. Further investigation into the distribution

of *E. coli* sequence types at beaches using DNA fingerprinting (Ishii, Hansen, Hicks, & Sadowsky, 2007; Whitman & Nevers, 2003) and REP-PCR (Beversdorf et al., 2007) revealed genetically distinct populations of *E. coli* forming in the beach sand, which could be differentiated from host-associated populations. The ability to recover the same genetically distinct populations of *E. coli* from multiple beaches overtime suggests that some strains of *E. coli* may be associated with secondary environments.

The relationship between genetic composition and survival has been further explored through assessing the phylogeny of *E. coli* isolated from various sources, with several phylogroups being associated with secondary environments. Phylogeny is determined by the evolutionary relationships within a species based on shared genetic characteristics (Chaudhuri & Henderson, 2012), with those strains belonging to more closely related phylogenetic groups sharing more similar genetic composition. This results in more closely related phylogroups containing similar characteristics, such as those that allow for the adaption to secondary environments. The phylogeny of *E. coli* was initially determined through multilocus enzyme electrophoresis or MLEE analysis of 35 loci, designating phylogroups A, B1, B2, C, D, and E (Selander, Caugant, & Whittam, 1987), with phylogroups F (Jaureguy et al., 2008) and cryptic clades I-V (Clermont et al., 2011) later being characterized through multilocus sequence typing or MLST.

It has been demonstrated that the distribution of phylogroups among *E. coli* sources is non-random, with populations of *E. coli* isolated from the beach containing a unique pattern in their phylotype distribution when compared to populations isolated from hosts (Gordon, Clermont, Tolley, & Denamur, 2008). Using an early Clermont phylotyping method it was

demonstrated that B1 (56%) was the most common phylotype recovered from six Lake Michigan beaches followed by phylogroups A (23%), D (15%), and B2 (6%) (Walk, Alm, Calhoun, Mladonicky, & Whittam, 2007). This distribution was distinct from those observed in populations from human gut samples, in which A (40.5%) and B2 (25.5%) were the most common phylogroups detected, followed by B1 and D being less common (17% each) (Tenaillon, Skurnik, Picard, & Denamur, 2010). Other work has also demonstrated B1 to be the most common phylotype associated with secondary environments such as estuary water (Berthe, Ratajczak, Clermont, Denamur, & Petit, 2013), manure lagoons (Howard, Martin, Gentry, Feagley, & Karthikeyan, 2017), and surface water and sediments (Tymensen et al., 2015), suggesting that members of this group may have select phylo-associated survival advantages.

In addition to A and B1, there are also several other minor phylogroups associated with secondary environments called the cryptic clades. These cryptic clade members are detected more commonly in secondary environments than in the gastrointestinal tract of hosts, and were originally characterized through a MLST analysis of isolates collected from freshwater beaches (Walk et al., 2009). An investigation into the genetic content of these clade members was conducted through whole genome sequencing, which revealed a set of 84 genes that are unique to these clade members, including the complete pathways for diol utilization and lysozyme production (Luo et al., 2011). Comparative genomic analysis of these clade members to their host-associated counterparts revealed that these groups did not exchange genetic material, indicating that independent populations of *E. coli* are forming in the beach sand.

Another method being used to identify the genetic composition of environmental populations is pangenomics. A pangenome is the complete library of genes found in an entire

species or within a subset of representative samples. It consists of the core genome, which contains basic house-keeping genes that belong to all members of that species, and the accessory genome, which contains genes that are variable in distribution that are generally associated with a particular phylotype or environment (Medini, Donati, Tettelin, Masignani, & Rappuoli, 2005). It is the ownership of these accessory genes that allow for the adaption to specific environmental conditions and can determine the habitat-range of an *E. coli.* Specific ownership of these accessory genes based on geographic location forms what is referred to as an ecotype, which is a subset of a species that is specifically adapted to the local environment. The goal of this work is to expand the limited genetic comparison that has been done using a novel population level bioinformatic workflow to identify the specific ecotype traits associated with freshwater beach sand survival.

The scope of this thesis work

Despite its limited accuracy, the enumeration of *E. coli* is still extensively used as a method to detect fecal pollution in recreational waters. Other members of the research community have also been critical of the use of such methods as have been laid out in the U.S. Recreational Water Quality Criteria (RWQC). In 2013, a group of 19 recreational water experts at the "US Recreation Water Criteria: A Vision for the Future" conference publicly stated that the last RWQC guidelines issued in 2012 fell short as they did not include results from recommended studies, including those that would be used to determine microbiological criteria for sand. The group concluded that the accuracy of recreational water quality criteria cannot be improved until there is a better understanding of sand microbial populations (Fujioka, Solo-Gabriele, Byappanahalli, & Kirs, 2015).

This work aims to fill this research need by using a genomic approach to study the survival of *E. coli* in beach sand. In this project, microcosm experiments, next-generation sequencing, and a novel bioinformatic workflow were used to identify those *E. coli* capable of long-term survival. The specific aims of this work include: 1) Determine the survival and growth dynamics of *E. coli* isolated from different host and environmental sources in freshwater beach sand; 2) Define the accessory genomes of *E. coli* capable of long-term beach survival through comparative genomics; 3) Determine if freshwater beach sand is selecting for specific genotypes.

Chapter 2 describes the results of several microcosm studies that were conducted in effort to identify *E. coli* capable of survival in the beach sand and to determine the factors that contribute to its survival. This experiment assessed the ability for the various phylotypes of *E. coli* to survive in beach sand, with Clermont phylotyping employed to characterize *E. coli* populations from the beginning and end of the microcosm experiment. This work advanced the understanding of *E. coli* survival in beach sand, providing additional evidence that some strains of *E. coli* are better suited for survival in the secondary environment than others. This work was published in Applied and Environmental Microbiology in 2021.

The key results included:

1.) *E. coli* survival is not linked to isolate source, as members of populations isolated from each source tested (beach sand, gull waste, and sewage) were capable of long-term survival in beach sand for over 45 days in native sand and 96 days in autoclaved sand.
2.) Nutrient availability and competition appear to be major determinants of the survival of *E. coli* in beach sand, as populations in autoclaved sand, which had high nutrients and low

competition, survived better than those seeded into native sand, which had high competition and low levels of available nutrients.

3.) *E. coli* that belong to phylogroups A and B1 increased in frequency by the end of the microcosm experiment; this suggests that members of these phylotypes are better suited for survival in the beach environment compared to members of other phylotypes.

Chapter 3 describes the novel approach used to identify the accessory genes associated with long-term environmental survivors. In this work populations of *E. coli* from the beginning and end of the microcosm experiment were compared using shot-gun sequencing and genetic mapping. Briefly, population level composite samples were formed from combining isolates from each source (sand, sewage, and gull) and timepoint (beginning and end of microcosm deployment) separately, and sequenced with Illumina short-read technology. These short-reads were then mapped to a scaffold of known accessory genes (from 21 isolates used in the microcosm experiment and nine previously sequenced environmentally-associated cryptic clade isolates). Isolates capable of long-term survival and their associated genes increased in frequency by the end of the experiment as less genetically adapted *E. coli* perished. This increase in survivor gene frequency was detected through comparing the mapping coverages of input and output populations, with genes that were enriched >25% in output sequencing samples considered survival-associated.

Key results included:

1.) A total of 198 functions were enriched in surviving populations from each sand, sewage, and gull output samples, with 59 of these enriched to a greater extent. The majority of these enriched functions were associated with metabolic enzymes and cellular transport.
2.) There were several sets of survival associated functions that were either widely-distributed among several phylogroups (i.e., all non-clade members), with a limited number narrowly distributed among certain phylogroups (i.e., A, B1, and cryptic clade), suggesting that there are multiple strategies for environmental survival.
3.) There were several survival enriched pathway modules that aid in adaption to secondary environments that were enriched among non-clade members, including betaine biosynthesis, which aids in osmotic regularity, and GABA biosynthesis and the GABA shunt, which offers a means to produce intracellular carbon and nitrogen.

Chapter 4 provides an overview summary of the work completed. It discusses how this work has fulfilled the needs of the research community and provides a reflection on the current state of beach monitoring, with recommendations for future work to be done on this system.

CHAPTER 2
Selective survival of *Escherichia coli* phylotypes in freshwater beach sand

Abstract

Escherichia coli is used as an indicator of fecal pollution at beaches despite evidence of long-term survival in sand. This work investigated the basis for survival of *E. coli* through field microcosm experiments and phylotypic characterization of more than >1400 *E. coli* isolated from sand, sewage, and gulls, enabling identification of long-surviving populations and environmental drivers of their persistence. Microcosms containing populations of *E. coli* from each source (n=176) were buried in the backshore of Lake Michigan for 45 & 96 days under several different nutrient treatments, including unaltered native sand, sterile autoclaved sand and baked, nutrient depleted sand. Availability of carbon and nitrogen, and competition with the indigenous community were major factors that influenced *E. coli* survival. *E. coli* Clermont phylotypes B1 and A were the most dominant phylotypes surviving seasonally (>6 weeks), regardless of source and nutrient treatment, whereas cryptic clade and D/E phylotypes survived over winter (>300 days). Autoclaved sand, presumably supplying nutrients through increased availability, promoted growth, and the presence of the indigenous microbial community reduced this effect. Screening of 849 sand *E. coli* from four freshwater beaches demonstrated that B1, but also D/E, were the most common phylotypes recovered. Analysis by qPCR for the Gull2, Lachno3 and HB human markers demonstrated only 25% of the samples had evidence of gull waste and none of the samples had evidence of human waste. These findings suggest prevalence of *E. coli* in the sand could be attributed more to long-term surviving populations than to new fecal pollution.

Importance

Fecal pollution monitoring still relies upon the enumeration of *E. coli*, despite the fact that this organism can survive for prolonged periods and has been shown to be easily transported from sand into surrounding waters through waves and runoff, thus no longer representing recent fecal pollution events. Here, we experimentally demonstrate that regardless of host source, certain genetically distinct subgroups, or phylotypes, survive longer than others under conditions typical of Great Lakes beach sites. We found nutrients were a major driver of survival and could actually promote growth, and the presence of native microorganisms modulated these effects. These insights into the dynamics and drivers of survival will improve the interpretation of *E. coli* measurements at beaches and inform strategies that could focus on reducing nutrient inputs to beaches or maintaining a robust natural microbiome in beach sand.

Introduction

Escherichia coli is commonly used in microbial source tracking, and is a Gram-negative, facultative anaerobe with a biphasic lifecycle that alternates between the guts of warm-blooded animals and the external environment (Kabler, Clark, & Geldreich, 1963; Savageau, 1983). Its primary habitat is within the biofilms of the mucus layer of the digestive tract of a host (Blount, 2015; Chang et al., 2004), and is shed into secondary environments through fecal matter at a density of $10^6 – 10^9$ CFU per g human waste (Savageau, 1983). It has also been shown to be deposited at beaches through gull waste at a density of 10^5-10^9 CFU per g (Alderisio & DeLuca, 1999; Fogarty, Haack, Wolcott, & Whitman, 2003; Gould & Fletcher, 1978). Because *E. coli* is a common inhabitant of the gut in humans and animals, which can also harbor pathogens, the EPA has recommended the use of *E. coli* as a fecal indicator (Dufour & Ballentine, 1986), with the

enumeration of this organism being the most common method to determine recreational water quality at beaches across the Great Lakes (U.S. Environmental Protection Agency, 2000). However, relying on *E. coli* levels to demonstrate recent fecal pollution events does not account for biphasic lifecycle of *E. coli*, in which it can spend its life in not only a primary host-associated habitat, but also in secondary environmental habitats, such as beach sand (Savageau, 1983). Evidence suggests that some strains of *E. coli* are suited for life in their secondary habitat, as sand-associated populations have been identified through routine beach monitoring (Alm et al., 2003; R. L. Whitman & Nevers, 2003), and have been determined to be genetically distinct from host-associated strains through REP-PCR DNA fingerprinting (Beversdorf et al., 2007; Ishii, Hansen, Hicks, & Sadowsky, 2007; Tymensen et al., 2015), multilocus enzyme electrophoresis (MLEE) and multilocus sequence typing (MLST) (Walk et al., 2007), and DNA microarrays (Oh et al., 2012).

E. coli has been commonly isolated at significant levels in beach sand and has been suggested to become part of the native community (D.D. Cloutier & McLellan, 2017; R. Whitman et al., 2014). These sand-associated populations of *E. coli* have been shown to impact water quality, as they are transported into local water through wave action (Brown & Boehm, 2016; Vogel et al., 2016), resulting in elevated levels of *E. coli* in recreational water without recent fecal pollution inputs. The accuracy of *E. coli* as a fecal indicator at beaches has been suggested to be improved if *E. coli* associated with long-term survival in sand were able to be differentiated (Fujioka et al., 2015). One method used to differentiate *E. coli* is phylogenomics, which uses genetic distance to describe the substructure of populations. Early phylogenetic analysis of *E. coli* identified four main phylogroups A, B1, B2 and D (Herzer, Inouye, Inouye, & Whittam, 1990; Ochman & Selander, 1984; Selander et al., 1987), with groups B2 and D at the

basal position in the phylogenetic tree (Touchon et al., 2009), and groups A and B1, which are considered sister groups, evolving later (Lecointre, Rachdi, Darlu, & Denamur, 1998). Additional phylogenetic analysis identified group C as a sister group of B1 (Clermont et al., 2011; Moissenet et al., 2010) and group F as a sister group to B2 (Clermont et al., 2011; Jaureguy et al., 2008). Phylogroup E is a sister group to D (Clermont, Christenson, Denamur, & Gordon, 2013) and is not differentiated in the multiplex Clermont phylotyping method used in this study. Further phylotypic diversity was discovered through assessing the population genetics of *E. coli* isolated from freshwater sand. Cryptic clades I-V were characterized, with clades III, IV, and V suggested to be environmentally adapted because of a relatively higher frequency of these clade members recovered from beach sand samples compared to a near absence of them in host-associated populations (Luo et al., 2011; Walk et al., 2009).

Analysis has shown that the distribution of phylogroups among *E. coli* sources is non-random. Members of phylogroup B1 were found to be more transient in humans than members of phylogroups A and B2 (Martinson et al., 2019), suggesting its members are more likely to be recovered from secondary habitats. In humans, phylogroups A (40.5%) and B2 (25.5%) were the most common, followed by B1 and D being less common (17% each); whereas in animals, B1 is the most common (41%) followed by A (22%), B2 (21%), and D (16%) (Carlos et al., 2010; Tenaillon et al., 2010). Previous work captured a snapshot of *E. coli* populations at six freshwater beaches and demonstrated that B1 was the most common phylotype recovered in the sand (56%), followed by groups A and D (23% and 15%, respectively), with B2 (6%) being the least frequently recovered (Walk et al., 2007). This variance in the distribution of phylotypes among sources suggests that some phylotypes are associated with particular hosts and environmental niches.

In effort to explore in-depth the ability of *E. coli* to survival in beach sand and test the hypothesis that certain phylotypes survived preferentially over others, we conducted a series of microcosm experiments with and without the native microbial community, and with a range of nutrient conditions. Our objective was to understand the potential and timeframe of survival for *E. coli* using isolates (i) obtained from the beach and already selected for in the secondary environment, and (ii) from gulls, a common fecal pollution source at beaches, using culture and Clermont phylotyping to assess shifts in the surviving *E. coli* population. We found that survival was associated with specific phylotypes rather than isolation from the primary or secondary habitats and was driven by nutrient availability and the absence of the native microbial community. Understanding which *E. coli* phylotypes can preferentially accumulate in the sand during long term survival could be useful to identify chronic *E. coli* reservoirs at beaches and improved interpretation of beach water quality that may be affected by delivery of *E. coli* from beach sand.

Results

Survival of *E. coli* from different sources

E. coli from each source and treatment were able to survive in the microcosm experiments for more than six weeks and in some cases nearly 1 year. Populations of *E. coli*, which represented a range of phylotypes, were incubated in microcosms containing native sand (i.e., collected from the beach and unaltered). Cell densities were found to steadily decrease by 2-3 orders of magnitude over 45 days (Figure 2.1), with exponential decay rates of -0.138, -0.104, and -0.122 day^{-1} for sand, sewage, and gull isolates, respectively (Table S2.1). In contrast, these same populations of *E. coli* incubated in autoclaved sand increased overall by one order of

magnitude by the end of the incubation (96 days) with exponential decay rates of -0.007, 0.009, and -0.014. Throughout the initial incubation period *E. coli* levels for each source remained within one order of magnitude of each other within each treatment, with initial concentrations of 1.38×10^7, 1.27×10^7, and 1.10×10^7 CFU/100 g sand in autoclaved sand, sewage, and gull treatments, and concentrations of 1.70×10^7, 1.23×10^7, and 4.47×10^7 after 96 days. Concentrations declined after the year-long incubation, with 1.12×10^6, 1.10×10^6, and 4.0×10^5 CFU/100 g sand in sand, sewage, and gull survivors.

In an effort to identify isolates that are capable of survival over winter at the beach, autoclaved and native microcosms (in duplicate) were left buried for 320 (native treatment), and 355 days (autoclaved treatment), in microcosms constructed to allow for air and water flow. Isolates from each source survived in the autoclaved treatment with final average concentrations of 1.15×10^6, 1.10×10^6, and 4.00×10^5 CFU/ 100 g sand in sand, sewage, and gull isolate collections; signifying only 1-2 orders of magnitude loss of *E. coli*. In contrast, *E. coli* in the native treatment had a six order of magnitude decrease over the year-long experiment, with a low number of residual isolates, 1.55×10^1 and 1.11×10^1 CFU/100g sand, recovered from sewage and gull microcosms native sand treatments, respectively (n= 34 isolates recovered, 6 combined microcosms).

Effect of nutrient availability on the survival of sand isolates

Additional microcosm experiments using nutrient-limited sand and *E. coli* isolated from the sand confirmed that survival varied based on sand treatment and nutrient availability. In this experiment, *E. coli* isolates originally collected from the sand were incubated in the nutrient-limited baked sand and survived poorly, with four orders of magnitude loss in cell densities, whereas isolates seeded into native sand decreased only two orders of magnitude. In contrast,

isolates seeded into autoclaved sand grew and increased two orders of magnitude (Figure 2.2). The exponential die-off coefficients for the native, autoclaved, and baked treatments were -1.75, 2.26, and -3.33 day^{-1}, respectively (Table S2.1). Results were comparable to those of the source microcosm experiment, with survival (or growth) somewhat more robust during this deployment. The nutrient concentrations determined in each treatment mirrored the growth or decay dynamics (Table 2.1). Nitrogen concentrations were found to have the strongest correlation to decay rates and higher total nitrogen correlated to slower decay. At the final timepoint, *E. coli* densities in the 90% baked 10% autoclaved treatment showed growth (decay constant of 1.93) and reached concentration levels within an order of magnitude of the 100% autoclaved sand treatment (Figure 2.2, Table S2.1). In contrast, the native sand amended with 10% autoclaved sand showed a reduction in cell densities and was nearly identical to the native sand treatment with no amendment.

Microbial community profiles within the microcosm experiments

Analysis of 16S rRNA gene sequencing revealed a wide range of richness and diversity in the microbial communities within the sand treatments of the microcosm experiments. Simpson's Diversity Index score for samples collected from the beginning of the microcosm experiment were 0.99, 0.55, and 0.25 for native, baked, and autoclaved treatments, respectively; shifting to 0.96, 0.89, 0.86 by the end of the experiment, with a moderate shift in the composition over the course of the experiment that included increases in *Methylophilaceae*, a common aquatic betaproteobacteria found in sediments in oligotrophic environments (Newton, Jones, Eiler, McMahon, & Bertilsson, 2011), and *Pseudomonaceae* (Chistoserdova, 2015) (Figure 2.3). The *E. coli* signal could be tracked in each microcosm in relation to the background microbial community within the 16S rRNA gene sequence data. The native sand obtained from Bradford

Beach had residual culturable *E. coli*, but levels were approximately 4 orders of magnitude lower than the inoculum of *E. coli* and was not detected in the top 50 ASVs from sequencing data. The average relative abundance of sequences annotated as *Escherichia/Shigella* in the community was 0.31 % in the native sand input and 0.016% in the output, with a concurrent 2 order of magnitude reduction in cultured *E. coli* levels over the course of the experiment. In contrast, a similar amount of *E. coli* was introduced into the baked sand and autoclaved microcosms, but the *E. coli* signal was much larger averaging 19% and 68% of the community, respectively. The microbial community for the baked and autoclaved sand treatments likely represent residual DNA, although the shifts in the community demonstrate regrowth of some members. The autoclaved sand treatments showed robust growth of *E. coli* with a one to two order of magnitude increase. However, by the end of the experiment, certain members of the microbial community increased, as seen by the low relative abundance of *Escherichia/Shigella* sequences, which comprised 0.04 and 2.12% of the community of the autoclaved and baked sand treatments, respectively.

Distribution of phylotypes among sources

Clermont phylotyping of *E. coli* from each freshwater beach sand (n=879), sewage (n=186), and gull waste (n=361) revealed that all isolates tested fell into one of seven potential phylogroup designations: A, A/C, B1, B2, cryptic clade, D/E, F, or unknown. All phylotypes were recovered from each source, however their distribution varied based on isolation source. A chi-square goodness of fit test determined that the phylotype distributions between isolates from sand and sewage were not significantly different, as they were both dominated by phylotype B1, averaging 42 and 53% relative abundance, respectively, and phylotype D/E averaging 18 and 20%, respectively (individual sources detailed in Table S2.2). Phylotypes A, B2, and F were

minor components (<20% collectively) of sand and sewage *E. coli* populations. This was in contrast to the phylotype distribution of the gull isolates, which were more evenly distributed and significantly different from the sand and sewage assemblages (Chi-square, $P < 0.01$, chi = 115; 76.7 respectively) (Table S2.2).

Phylotype profile shifts in microcosm experiments

Significant shifts in phylotype distributions ($p < 0.01$) were observed after 45 days in the native sand microcosms seeded with sand, sewage, and gull isolates, with phylotypes B1 and A increasing in relative abundance, and phylogroups A/C and F decreasing regardless of the original source of the *E. coli* populations (Figure 2.4 a-b). These same isolates from sand, gull and sewage collections incubated in autoclaved sand microcosms showed significant, yet not as pronounced, changes in phylotype distribution, with notable increases in the relative abundance in phylogroups A and B1. However, in the nutrient microcosm experiment only a slight shift in these phylotypes was observed in the sand isolate collections (Figure 2.4 c-d). Interestingly, the A/C phylotype remained relatively stable in the autoclaved sand microcosms, unlike the native sand microcosm, where it decreased. Characterization of isolates recovered at one year from the native sand treatments demonstrated phylotypes B1, D/E, and B2 comprised the sewage survivors (n=29), and phylotypes clade, D/E, and very low numbers of B1 comprised the gull survivors (n=15) (Figure 2.4c).

Phylotype growth in minimal media

The growth of isolates from each phylogroup in minimal media R2A, R2A media diluted 1:10, and a rich media MUG media demonstrated that there was no significant difference in the carrying capacity, growth rate constant, or generation times between the phylogroups within each

media type. However, the carrying capacity, or the total number of organisms that can be supported, varied among media treatments (Figure S2.1).

Distribution of phylotypes at freshwater beaches and detection of human and gull markers

An extensive survey of *E. coli* isolated from sand demonstrated that phylogroup B1 was the most common phylotype recovered, dominating seven of ten sand samples collected May through October of 2008-2019, with phylogroups F and D/E also present at notable frequencies. Clade/E, A/C and B2 were more minor contributors to the populations (Figure 2.5). Chi-square goodness of fit test determined that the distribution of phylotypes among urban and rural beaches differed significantly (Chi-square, $P < 0.01$, chi = 62.2). Although both types of beaches were dominated by B1, urban beaches had more members of phylogroups B2 and Clade/E than at rural beaches (Figure 2.5). Those *E. coli* collected after the winter season in February (n=49) and April (n=118) of 2019, showed that phylotypes were more diverse and February patterns were dominated by F and D/E phylotypes, rather than B1. Interesting, April patterns closely mirrored what has been recovered in gulls (Figure 2.4). The *E. coli* levels in the winter samples were low with 3 CFU/100 g sand in the February sample, and 25 CFU/100 g in the spring April sample.

To determine if recent fecal pollution could be detected in conjunction with isolation of *E. coli* from sand, qPCR was conducted on a subset of samples using fecal indicator markers such as human *Lachnospiraceae* (Lachno3) and human *Bacteroides* (HB), and the gull marker (Gull2). Human markers were not detected in any of the samples analyzed. However, the gull marker was detected in 11/44 samples tested. This included 1/9 sites from each of Kohler-Andrea and Point Beaches (collected 6/21/16), 2/3 sites on only one of the 4 survey dates from Braford Beach (5/30/18), as well as multiple sites on each of the four survey dates collected from Atwater

Beach (5/30/18, 7/11/18, 7/16/18, and 7/22/19). The geometric mean of *E. coli* levels in samples positive for the gull marker were four-fold higher than samples with no marker detected (Table S2.3). Overall, the urban beach sites Atwater and Bradford had evidence of fecal pollution in 23% of samples, and the rural sites Kohler-Andrea and Point Beaches had evidence in 11% of samples. These results suggest that gull waste may be contributing significantly to the *E. coli* burden in the sand.

Discussion

E. coli reservoirs and potential sources

E. coli can be introduced into beach sand through many routes, including bird and animal waste, sewage, and run-off (Blyton et al., 2015; Lee et al., 2006; R. L. Whitman & Nevers, 2003). While there is strong evidence that *E. coli* reservoirs in beach sand impact water quality, it is unclear how much might be attributed to repeated deposition compared with long term survival (R. L. Whitman & Nevers, 2003). Previous work demonstrated B1 was the most common phylotype isolated from freshwater beach sand (Walk et al., 2007), water (White et al., 2011), and soil (Nandakafle et al., 2017), suggesting *E. coli* within this lineage may become naturalized (Solo-Gabriele et al., 2016). However, the B1 phylotype is also found in animals, including birds (Blyton et al., 2015; Carlos et al., 2010; Ishii et al., 2007). Gull waste has been identified as a major contributor of fecal pollution at Great Lakes beaches (D.D. Cloutier & McLellan, 2017; Fogarty et al., 2003; Staley & Edge, 2016). We found *E. coli* present in sand in the absence of the Gull2 marker in 34 of 44 sand samples tested with no samples positive for human markers, which is consistent with past reports by our laboratory (D.D. Cloutier & McLellan, 2017). In this work, we also demonstrated that the phylotype profiles in gulls are

different than what is recovered from sand in most cases, offering further support that the *E. coli* reservoirs are a result of long-term survival. We tested a range of isolates from the primary habitat of gulls as well as isolates that had already been selected for in the secondary habitats of sand and sewage in an effort to determine if prolonged survival was a uniform characteristic of certain phylotypes regardless of isolation sources. We further explored the decay rates under actual beach conditions and characterized the strains that were capable of surviving over winter.

E. coli phylotypes in primary and secondary habitats and survival in beach microcosm experiments

E. coli belongs to one of eight phylotypes*:* A, B1, B2, C, D, E, F, and cryptic clade (I-V), and have been shown to vary in behavior and lifestyle (Johnson, Gordon, & Bauer, 2002; Walk et al., 2009). In each source tested (sand, sewage, and gull waste) all phylotypes were recovered. Populations from secondary habitats (beach sand and sewage) had similar phylotype distributions, dominated by phylotype B1, whereas gulls had a more even distribution of phylotypes and included higher proportions of A/C and B2, the latter of which is common in primary habitats, including human and animal hosts (Stoppe et al., 2017; Tenaillon et al., 2010; White et al., 2011).

Results from the microcosm experiments support the hypothesis that some phylotypes are better adapted for secondary habitats over primary habitats. After six weeks of incubation the relative proportion of B1 as well as A phylotypes increased regardless of original source, while members of phylotypes B2 and D/E, which are considered host-associated specialists (Johnson et al., 2002), did not survive well in the six week microcosm deployment. Previous work supports the idea that phylotypes B1 and A are strong survivors in the external environment, as they were the most common phylotypes recovered from dairy manure lagoons (Howard et al., 2017) and

microcosms with filtered estuary water showed that phylogroup B1 persisted longer than the other phylogroups (Berthe et al., 2013). Additionally, members of phylogroup A, isolated from raw sewage, were found to survive chlorine stress better than other phylogroups (Zhi et al., 2016).

Phylogroups A and B1 have been shown to cluster mostly on one branch when assessed with MLST, and phylogroups B2, D/E, and F cluster on a separate branch, with cryptic clade sequence types set apart on their own branch deep within the tree (Walk et al., 2009). Interestingly, A/C, which is a sister group to A, did not survive as well as A and B1. Our results demonstrated that original isolation from primary or secondary habitats did not affect the relative enrichment of A and B1, suggesting that there is ongoing flux between host and the secondary habitat as opposed to long term naturalization of select strains within a lineage. This is consistent with past research that shows *E. coli* B1 sequence types from water and sediment could not be distinguished from B1 isolated from various hosts using MLST (Clermont et al., 2011; Tymensen et al., 2015). Because A and B1 are clearly phylogenetically distinct from other phylotypes, they may have unique phenotypic traits that make them better suited for the external portion of the *E. coli* biphasic lifestyle.

B1 and A phylotypes survive seasonally, but not over winter

Here we show that on a seasonal scale (~6-8 weeks) B1 and A phylotypes preferentially survive over others, but not on an annual basis. At the ~1 year time point, isolates from each source were easily recovered from the autoclaved sand treatment as levels were only reduced 1 order of magnitude from the original inoculum. From the native sand treatments, there were minimal residual strains in only two of the three source microcosms. Interestingly, the majority of these phylotypes were B2, D/E, and Clade/E, with members of the B1 phylotype found only in

the sewage source microcosm (Figure 2.4). Results from the winter sampling of *E. coli* at beaches mirrored the results of the year-long microcosm experiments. While it is difficult to draw conclusions from a limited number of strains, taken together, these results support the idea that strains within B2, cryptic clade, and D/E, which are lineages that are distinct from the A and B1 branches of the phylogenetic tree (Walk et al., 2007), can survive for more than one year. Therefore, it may be that select strains of these host-associated groups have the ability to become naturalized, i.e., become long term replicating members of the sand microbial community, rather than B1 and A that appear to be suited for seasonal survival.

Factors that influence the survival of *E. coli* in beach sand

Previous work has demonstrated that high nutrient levels at beach sites were associated with high *E. coli* burdens (Danielle D Cloutier, Alm, & McLellan, 2015). Here, results from the native and autoclaved sand microcosms suggested that the ability to survive is dependent on both nutrient availability and competition with the native community. Interestingly, the *E. coli* in native sand and native sand amended with 10% autoclaved sand (providing a nutrient source) had nearly identical decay rates, while the *E. coli* in baked sand amended with 10% autoclaved sand grew nearly to the levels close to what was found in microcosms with 100% autoclaved sand. While these experiments cannot distinguish if the native community was competing for nutrients, or if exclusion or predation played a role, it is clear that the indigenous microbial community in sand is a major modulator of *E. coli* survival. Previous work showed that predation and competition have negative effects on *E. coli* survival, while increased nutrients are associated with increased survival (Wanjugi, Fox, & Harwood, 2016), and that *E. coli* not only survived but grew in sand amended with nutrient sources such as stormwater runoff, plankton, and autoclaved sand (Alm et al., 2006; Beversdorf et al., 2007; M. N. Byappanahalli et al.,

2006). It has also been observed that moisture content can affect bacterial growth and survival, with the bacterial load the highest in the moist sand of the berm (Beversdorf et al., 2007; Muruleedhara N. Byappanahalli, Whitman, Shively, Sadowsky, & Ishii, 2006; Ishii et al., 2010). The microcosm experiments performed here are distinct from previous reports (with the exception of Alm et al. 2006, which tested a specific strain of *E. coli*) where experiments were carried out under actual beach conditions for long periods of time and were matched with phylotyping that demonstrated a shift to B1 and A phylotypes under an array of conditions. It is unclear what critical nutrients the autoclaved sand provided to promote growth, i.e., carbon or nitrogen or both; however, knowledge of specific nutrients that drive prolonged survival would be important to uncover to better understand these dynamics.

By evaluating growth in various minimal and rich media, we showed that members of all phylogroups can grow at low nutrient concentrations. Previous work also shows that the growth rate of *E. coli* does not vary based on phylogenetic association (Ingle et al., 2011). These finding suggest that levels of *E. coli* found in beach sand are a function of traits involved in survival, for example, resistance to stressors or metabolic traits involved in homeostasis, rather than traits associated with growth in suboptimal conditions.

Phylotype patterns at beaches and implications for beach monitoring

The phylotype patterns recovered from beach sand sampled during the swimming season included a high relative proportion of B1 phylotype and resembled a mixture of gull and the sand profiles we originally characterized in the microcosm experiments. The phylotype distribution differed between urban and rural beaches, where phylogroups B2 and D/E were more common at urban beaches and in general, and sites that tested positive for the gull marker generally contained more B2 or D/E phylotypes. The April collection also closely resembled what was

recovered from gulls. This suggests that urban beaches may have more constant inputs. However, we noted that B2, D/E, and cryptic clade were enriched in the beach profiles compared to gulls, and these were the same groups that survived, albeit at low level, over winter. Taken together, it seems the major culturable burden may come from seasonal survivors, but over an annual cycle, very long term survivors can add to the overall beach burden, raising questions about their capability to take up enough nutrients to actually grow rather than simply survive. Understanding the population dynamics and survival characteristics can improve the interpretation of beach monitoring results, as a high level of B1 and A may be an indication of contamination from seasonal survival in beach sand, rather than from recent fecal input. The time scale of the microcosm experiments suggests that while there is survival that could impact water quality results during the beach season, *E. coli* reservoirs in sand do not appear robust from season to season.

Methods

Experimental Overview

Experiments were designed to characterize *E. coli* capable of long-term survival at the beach and to determine the factors that allow for their survival. Microcosm experiments were employed to compare the survival of *E. coli* from various sources under different nutrient conditions, including unaltered native sand, autoclaved sand, and nutrient depleted baked sand, with phylotyping used to characterize shifts in populations after incubation. Microbial communities in each treatment were assessed through 16S rRNA gene sequencing and culturing for *E. coli*. Surveys were conducted at four beach sites to determine if phylotype profiles matched patterns found in microcosms, with gull and human fecal pollution markers measured

concurrently with *E. coli* levels. Taken together, these experiments were able to characterize *E. coli* capable of long-term survival in beach sand.

Isolate Collection

E. coli was isolated from freshwater beach sand, gull waste that was observed to be freshly deposited, and wastewater influent from multiple dates (Table S2.2). Gull samples were included as they are documented to be a potential contributor to fecal pollution at beaches along the Great Lakes (D.D. Cloutier & McLellan, 2017; Converse et al., 2012; R. L. Whitman & Nevers, 2003) and around the world (Brown & Boehm, 2016; Goodwin, Gruber, Vondrak, & Crumpacker, 2016). *E. coli* isolated from freshwater beach sand was collected from the berm of the beach, where the sand receives intermittent moisture through wave action, at 3-4 sites 10m apart per beach (Table S2.2). Samples were collected from top 6 cm of the berms of Bradford and Atwater Beach in Milwaukee, WI; Kohler-Andrae Park in Sheboygan, WI; and Point Beach in Manitowoc, WI. Additional samples were collected from Bradford and Atwater Beaches in February (n=49) and April (n=118) of 2019, with the February sample recovered from under the snow and ice buildup on the beach in the approximate area of the berm. Due to low *E. coli* concentrations winter samples from each beach were combined to form composite samples. The samples were collected in sterile Whirl-Pak bags and transported on ice then processed using the method described by Boehm et al (Getrich et al., 2009), in which the samples were then eluted in sterile water (1:10) and shaken by inversion for two minutes, then filtered onto a 0.45 µm pore size nitrocellulose filter (Millipore, Billerica, MA). Filters were transferred to modified mTEC agar plates, and incubated according to the USEPA Method 1603 for 18 hr (U.S. Environmental Protection Agency, 2000). After incubation colonies were streaked-out onto mTEC agar plates

and incubated for 18 hr. Isolated colonies were then picked into 96-well micro-titer plates and stored in a 25% glycerol 75% MUG media solution at -80°C until needed.

Microcosm Experiments

Microcosms were used to conduct *in situ* studies on the survival of *E. coli* in beach sand. The microcosm design was adapted from Alm et al, (Alm et al., 2006), and were constructed from pieces of polyvinyl chloride pipe that were 9 cm in length and 5 cm in diameter, and sealed with polyvinyl chloride knockout test caps which had approximately thirty 1 mm diameter holes drilled through them. The end caps were lined with two 0.22 µm filters on the inner sides of the caps, which prevented bacteria and other microbes from entering or leaving the microcosms while allowing for oxygen and moisture to enter. The microcosms were filled with various sand treatments, seeded with *E. coli* then sealed with silicone sealant, and buried 0.5 m deep in the backshore sand of Lake Michigan in Milwaukee, Wisconsin, USA. The native sand treatment mimicked the conditions at the beach and consisted of sand collected directly from the beach with its native microbial community and available nutrient content intact. The autoclaved treatment consisted of moist sand collected from the beach that was autoclaved for 1 hr, which inactivated the native microbial community, releasing nutrients from biological constituents and making organic nutrients highly available (Table 2.1). Lastly the baked sand treatment created a nutrient-limited environment and consisted of sand that was baked in a muffle furnace for 3 hr at 550 °C, then washed with sterile MiliQ water, and autoclaved to sterilize. Isolates used in the experiments were grown for 18 hr in MUG media, washed with DNA-free sterile water 3X, then diluted to a final concentration of 10^6 cells/ml through using OD values.

The source microcosm experiment entailed pooling populations of 176 isolates from each gull, sewage, and beach sand by source and then inoculating the pooled isolates into native and autoclaved sand. The nutrient microcosm experiment was inoculated with the beach sand isolate collection. Native sand microcosms were buried for 45 days and autoclaved microcosms were buried for 96 days, with an additional set of microcosms left for 320 and 360 days to test long-term survivability. The following summer season, a nutrient microcosm experiment was conducted in which the same population of beach isolates (n=176) were seeded into native, autoclaved, and baked sand, and buried at the beach for 56 days. Included in this experiment were two nutrient dose treatments, in which native and baked sand were seeded with 10% by weight autoclaved sand, which acted as a nutrient source. Comparison between the nutrient spiked treatments to non-spike treatments were used to determine the effects that competition has on *E. coli* survival in beach sand. Isolates from the end of each experiment were recovered for phylotyping, stored in 96-well microtiter plates in a 25% glycerol 75% MUG solution in a -80°C freezer.

Phylotypic Characterization

Phylotyping was used to characterize the populations of *E. coli* in each of the sources tested and was used to assess the populations of *E. coli* which survived the microcosm experiments. The Clermont phylotyping method was used to classify samples into one of seven groups: A, A/C, B1, B2, D/E, F, or Cryptic Clade (Clermont et al., 2013). Template for this reaction was prepared by growing cultures for ~18 hr, then diluting 1:20 with sterile water, followed a 10 min at 100°C in a thermocycler to lyse the cells. The template was then used in the multiplex PCR as described in (Olivier Clermont et al., 2013), with the PCR product then

analyzed on a 2% agarose gel, and visualized under UV at 320 nm after staining in 500 mL of 0.5 mg/mL of Ethidium bromide in 1XTAE for 15 min.

Nutrient Analysis

The levels of total carbon, total nitrogen, organic carbon, organic nitrogen, and total phosphorus were measured in each treatment of the nutrient microcosm. Each sand sample was dried in a drying oven overnight at 60°C prior to processing. Carbon and nitrogen measurements were made with a Carlo-Erba NA-1500 CNS elemental autoanalyzer (Haak-Buchler Instruments, Saddlebrook, NJ) and determined with an acetanilide standard. Total phosphorus for each sample was processed through the methods described by (Ruban et al., 1998). Briefly 200 g of sand was combusted with Mg $(NO_3)_2$ for 2 h, followed by a 16-hr digestion in 1 N HCl. Sand extracts were then diluted and analyzed using the ascorbic acid phosphomolybdate method.

Nutrient Growth Experiment

A nutrient growth experiment was conducted in effort to compare the effects that nutrient availability has on various phylotypes. A subset of isolates from each phylotype (n= 8-12) were grown in MUG for ~18 hr then inoculated into three different nutrient conditions in triplicate; an optimal growth media, MUG (Thermo Fisher Scientific, NH); a nutrient limited media, R2A; and an extreme nutrient limited media, 1:10 diluted R2A. The growth of the isolates was then tracked through measuring the optical density of each culture at a wavelength of 600 nm on a BioTek plate reader (VT, USA), for 36 hours.

16S rRNA gene sequencing and qPCR analysis

Each sample was analyzed in duplicate. DNA was extracted from nitrocellulose filters that processed 100 ml of sand eluent that was frozen at -80 °C and crushed manually using a sterile spatula. DNA extraction was completed with a Fast DNA Spin kit for soil, according to

the manufacturer's instructions (MP Biomedicals, Solon OH). The microbial communities of sand samples from within each microcosm experiment were assessed through 16S rRNA gene sequencing of the V4V5 hypervariable region, using protocols developed at the Josephine Bay Paul Center at the Marine Biological Laboratory, Woods Hole, MA (Nelson, Morrison, Benjamino, Grim, & Graf, 2014). The qPCR assays were carried out on sand samples from the beach surveys using an ABI StepOne real-time PCR system with TaqMan hydrolysis probe chemistry (Applied Biosystems, Foster City CA). The qPCR assays conducted were based on those ran by Cloutier 2016 and included human-associated *Lachnospiraceae* (Lachno3) (S. Feng, Bootsma, & McLellan, 2018), human-associated *Bacteroides* (HF 183) (Bernhard & Field, 2000; Kildare et al., 2007), and gull-associated *Catellicoccus marimammalium* (Gull2) (Lu, Santo Domingo, Lamendella, Edge, & Hill, 2008). Samples were run in duplicate, with standard curves ran in triplicate, made through six serial dilutions of 1:10 of a 1.5 x 10^6 linearized plasmid containing a target sequence. PCR cycling was performed as the following: 2 min at 50 °C, 10 min at 95 °C, 40 cycles of 95 °C for 15 s and 60 °C for 1 min (D.D. Cloutier & McLellan, 2017). The copy number was then converted to CN/100 g of sand.

Data analysis

All statistical analysis was conducted in R version 3.5.1 (R Core Team, 2018) using R core packages. Phylotypic characterization of habitats, and phylotype shifts during microcosm experiments were characterized through chi-square analysis with an acceptable significance level of $p<0.01$. The package growthcurvr was used to assess growth in the nutrient limitation experiment (Sprouffske & Wagner, 2016), with ANOVA employed in assessing the results, with an acceptable significance level of $p<0.05$. Spearman's rank correlation coefficient was used to determine any correlations between the nutrients measured and the final population

concentrations with an acceptable significance level of r^s <0.6. Lastly the 16S rRNA gene sequence data was processed, with the reads trimmed using cut-adapt v2.10 (Martin, 2011); forward and reverse reads merged using PEAR v1.10.12 (Zhang, Kobert, Flouri, & Stamatakis, 2014); taxonomic determination was conducted with DADA2 v1.16 (Callahan et al., 2016), with phyloseq v1.22.3 used to visualizations (McMurdie & Holmes, 2013), and statistical analysis complete with the R program VEGAN v2.5.6.

Acknowledgments

This work was supported by the University of Wisconsin Sea Grant Program grant #NA10OAR4170070. We would like to thank the members of the McLellan lab and Newton lab for their assistance in microcosm deployment and recovery, and for strain isolation, specifically Shuchen Feng, Adelaide Roguet, Lou LaMartina, Katie Alexander and Sage Legault.

CHAPTER 2: Tables and Figures

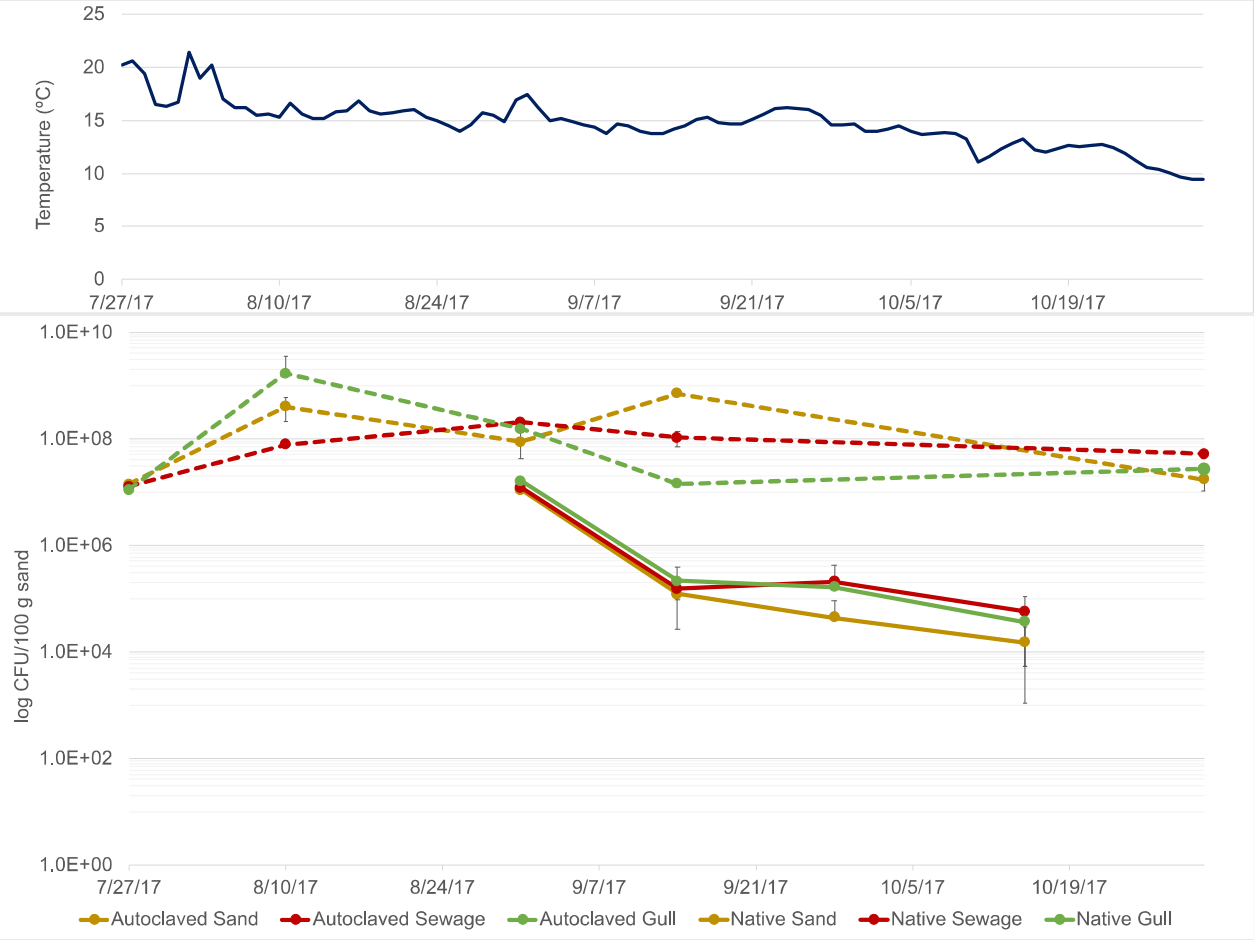


Figure 2.1. Temperature and *E. coli* levels in source microcosms. Top panel displays internal temperature in microcosm experiments measured hourly through an *in situ* temperature probe. Temperatures fluctuated, ranging 9-22 °C. Bottom panel displays *E. coli* levels recovered from native sand (solid lines) and autoclaved sand (dashed lines) inoculated with 176 *E. coli* isolates from each source. Microcosms were buried in sand at the shoreline and continually wetted by wave action. Native sand treatments were deployed for 45 days and autoclaved sand treatments 96 days. Microcosms were analyzed in triplicate for each time point for each source. Control microcosms with no seeded *E. coli* were negative. Isolate survival varied based more on treatment rather than source.

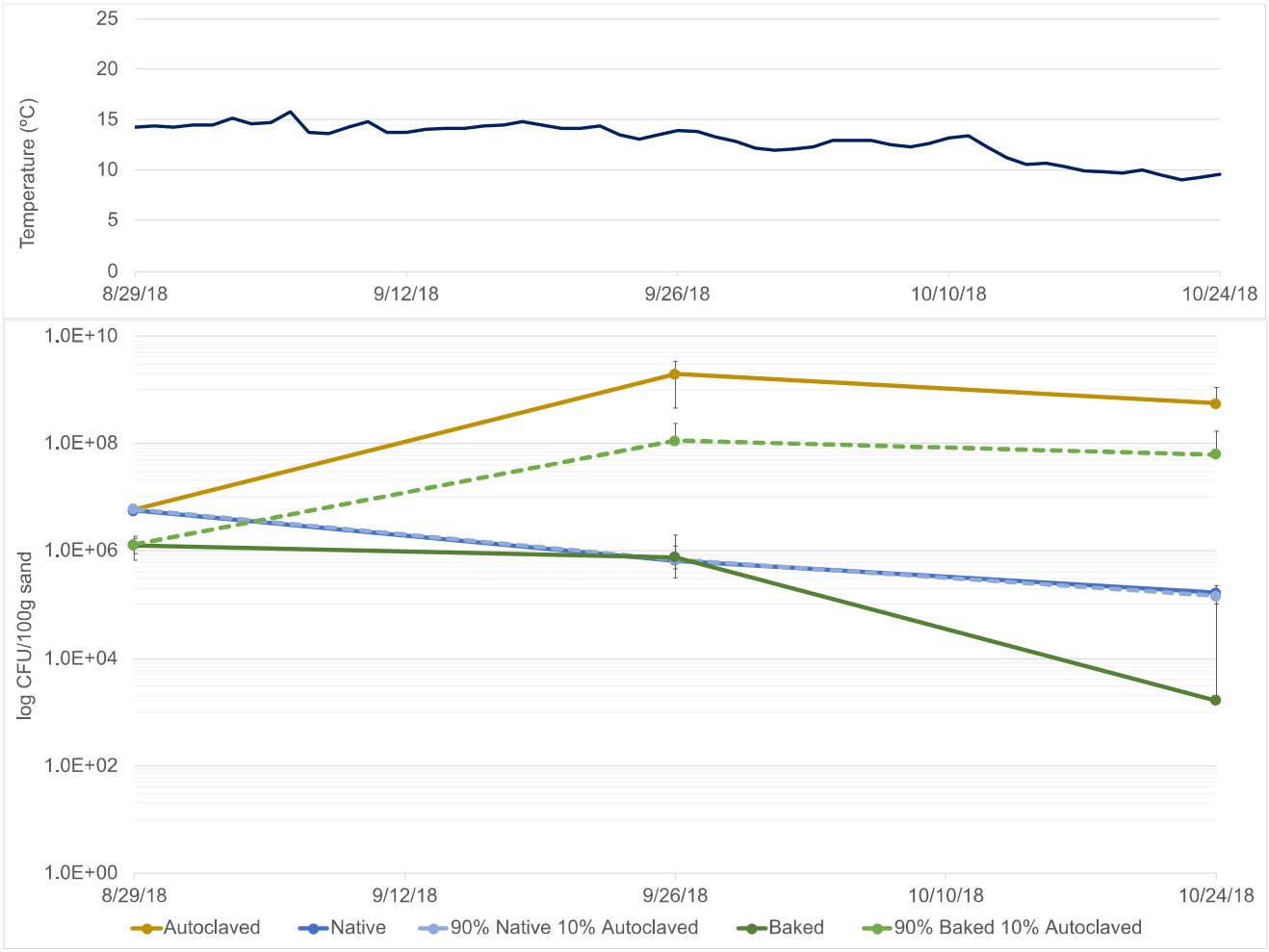


Figure 2.2. Temperature and population counts in nutrient microcosms. Top panel shows the internal temperature in microcosm experiments measured hourly through a deployable temperature prob. Temperatures fluctuated, ranging 8-16 °C. Bottom panel displays population counts of *E. coli* recovered throughout the 56 day nutrient microcosm experiment, with nutrient dose treatments noted by dashed lines. The same set of isolates originating from beach sand were used in each treatment. Nutrients appear to be correlated with *E. coli* population survival and growth.

Table 2.1. Nutrient content and Spearman's rank of nutrient treatments in microcosm experiments.

Sand Treatment	Total Carbon (ppm)	Total Carbon SD	Organic Carbon (ppm)	Organic Carbon SD	Total Nitrogen (ppm)	Total Nitrogen SD	Total Phosp-horus (ppm)	Total Phosph-orus SD
Native Blank	17600	NA	5660	NA	19	NA	13.7	NA
Autoclaved Blank	45800	NA	23700	NA	36.6	NA	18.6	NA
Baked Blank	30700	NA	14900	NA	19.5	NA	16.6	NA
Native	17700	506	151	2.81	13.7	4.60	9.77	1.43
Autoclaved	57700	19800	12810	17900	64.1	21.3	15.6	0.46
Baked	28900	803	676	134	BLD	NA	16.4	1.73
90% Native 10% Autoclaved	21000	3990	12800	7540	12.7	5.40	12.8	2.47
90% Baked 10% Autoclaved	34500	1480	8470	8640	23	12.6	13.9	0.86
Spearman's Rank Correlation Coefficient	0.59		0.05		0.74		0.00	

[1]Spearman's rank correlation coefficient was calculated using nutrient measurements from the output of the experiment and the percent of *E. coli* surviving in each of the inoculated treatments BLD signifies a value bellow the limit of detection. Nutrients were measured in duplicate for inoculated samples and on single blank samples. Organic nitrogen in each sample was below the limit of detection of 10 ppm.

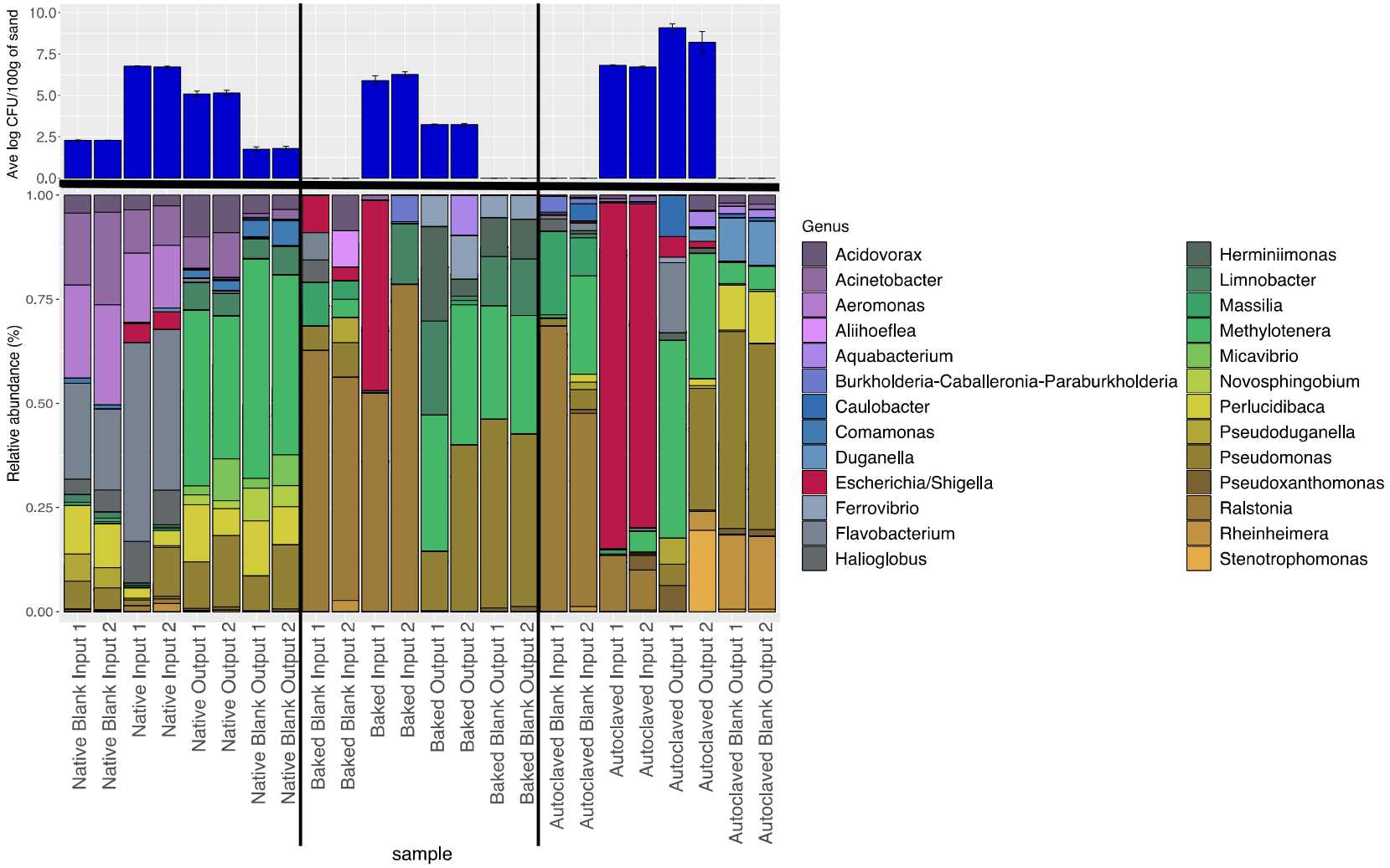


Figure 2.3. Comparison of 16S rRNA gene composition and *E. coli* counts in nutrient microcosm treatments. Top) Concentrations of *E. coli* in microcosm treatments, determined through direct plate counts of samples Bottom) Top 50 ASV in each samples, organized by treatment from initial blank, initial input, final output, to final blank

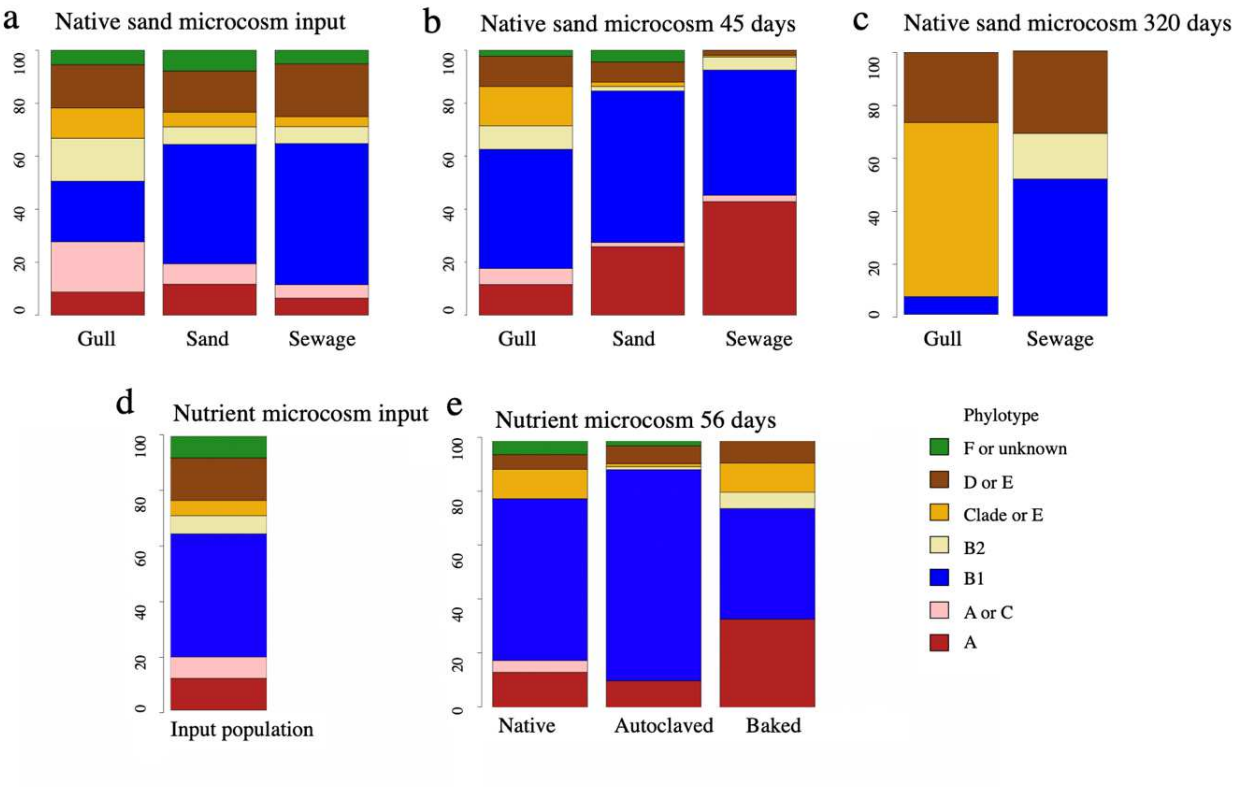


Figure 2.4. Comparison of phylotype distribution in sources at the start and end of microcosm deployments with gull, sand, and sewage *E. coli* isolates. a) phylotype profiles of sources in native sand treatments at the start of the deployment, panels a, b, and c are presented in Figure 2.4 but shown here as reference; b) phylotype profiles of isolates in native sand treatments at 45 days (n=182 for gull and sand sources, n=161 for sewage sources); c) phylotype profiles of isolates at 320 days (sewage n=29 and gulls n=15); d) Phylotype profile of sand isolates used as input in the nutrient microcosm experiment; e) Phylotype profile of the nutrient microcosm outputs.

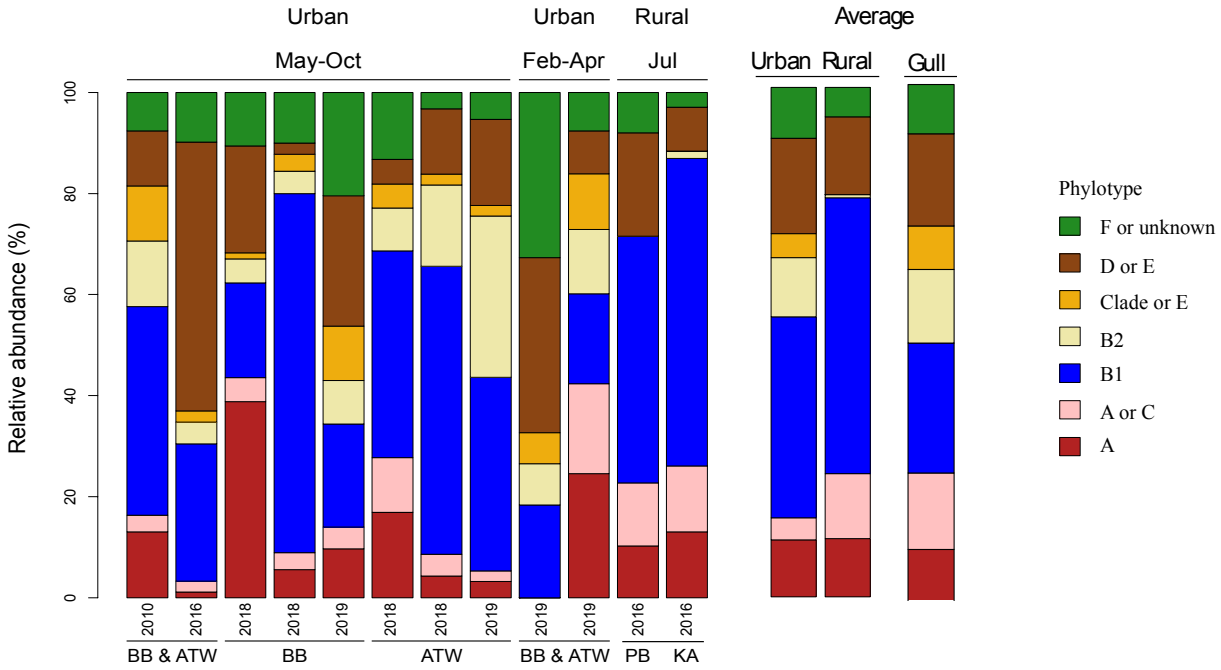


Figure 2.5. Phylotype distributions of *E. coli* isolates collected from beaches along the western shores of Lake Michigan, 2008-2019. Isolates were collected from the berm of the beach and characterized through Clermont phylotyping. Urban beaches included Atwater Beach (ATW) and Bradford Beach (BB) in metro Milwaukee, Wisconsin. Rural beaches included Pont Beach (PB) in Manitowoc Wisconsin, and Kohler-Andre Beach (KA) in Sheboygan, Wisconsin. Comparison of urban and rural beaches, found in the last two columns, excludes samples collected during the winter, BB & ATW Feb 2019 and BB & ATW April 2019. Average phylotype distribution of *E. coli* from gull waste (See Table S2.2) included for comparisons.

CHAPTER 3
The genetic determinates of *E. coli* survival in beach sand

Abstract

Escherichia coli contain a high level of genetic diversity and are generally associated with the guts of warm-blooded animals, but have also been isolated from secondary habitats outside a host. Here we employed a novel approach that combined *in situ* microcosm experiments with population level genomic analysis to identify accessory genes associated with survival within the beach sand environment. In this work, *E. coli* originally isolated from sand, sewage, or gull waste (n=564 total) were seeded into sand, sealed in microcosm chambers, and buried for 45 days in the backshore beach of Lake Michigan. Survival-associated functions were identified by comparing the genetic composition of populations from the beginning and end of the microcosm experiment. Short-read sequences from the pool of isolates were mapped to a scaffold of accessory genes from 21 *E. coli* genomes. We found that environmental survival was associated with a wide variety of factors distributed across most phylogroups, with the majority corresponding to metabolism enzymes and transport proteins. Of the 414 unique functions identified, only a small fraction were exclusive to a specific phylotype. Several pathway modules enriched in surviving populations included those for betaine biosynthesis, which produces an osmo-protectant, as well as the GABA biosynthesis and the GABA (gamma-Aminobutyrate) shunt pathway, which aid in nutrient use versatility. Overall, these results demonstrate that the genetic flexibility within this species allows for survival in the environment for extended periods.

Importance

Escherichia coli is commonly used as an indicator of recent fecal pollution in recreational water despite its known ability to form host-independent populations in secondary environments, such as beach sand. These long-term survivors form reservoirs and can be introduced into the

water column through wave action or run-off during precipitation events, thereby impacting the perception of local water quality. Current beach monitoring methods cannot differentiate these long-term environmental survivors from *E. coli* derived from recent fecal input, resulting in inaccurate monitoring results and unnecessary beach closures. This work identified the genetic factors that are associated with long-term environmental-survivors, providing insight into the mechanistic basis for *E. coli* accumulation in beach sand. A greater understanding of the intrinsic ability and conditions that promote long term survival of *E. coli* will improve interpretations of beach water quality assessments using this indicator.

Introduction

Escherichia coli is a gram-negative bacterial species that is best known for its association with the gut of warm-blooded animals. However, *E. coli* is also commonly found in secondary habitats, such as water, sediment, and soil (Jang et al., 2017; Savageau, 1983). The ability for *E. coli* to survive in these secondary habitats varies greatly based on its genetic composition. Initially *E. coli* populations in secondary environments were thought to decay quickly and were characterized by a net negative growth rate (Savageau, 1983). However, additional analysis has revealed that certain strains are capable of long-term survival in several secondary environments, including warm and temperate soils (Brennan, Abram, Chinalia, Richards, & O'Flaherty, 2010; Muruleedhara N. Byappanahalli et al., 2006), sediments (Solo-Gabriele, Wolfert, Desmarais, & Palmer, 2000), and beach sand (Alm et al., 2006; Beversdorf et al., 2007; M. N. Byappanahalli et al., 2006; Hartz et al., 2008; Palmer, Law, & Soupir, 2020; Praveena, Shamira, Ismail, & Aris, 2016).

It can be difficult to identify these long-term survivors in the environment since urban, agricultural, and even pristine areas can have a constant input of fecal pollution from multiple sources that include wildlife. These inputs create heterogeneous populations that contain long-term environmental survivors and recently deposited isolates from fecal contamination (such as gull waste) and run-off. The dynamic nature of these environmental populations was observed in previous work that showed that *E. coli* with various survival capabilities can co-exist in the same secondary environment, as seen in estuary water samples that contained both survival-associated genotypes with survival rates > 14 days and host-associated genotypes that had a significantly lower survival rate, of < 2 days (Berthe et al., 2013).

The berm of the beach is known to be one of these highly dynamic environments, with robust *E. coli* populations consistently observed. Previous work has demonstrated *E. coli* concentrations to be constantly 10 -1,000 higher in sand (weight to volume) than in paired water samples (Alm et al., 2006; Solo-Gabriele et al., 2000; Staley et al., 2015; R. L. Whitman & Nevers, 2003), with a similar trend seen when comparing to backshore beach sand populations (Danielle D Cloutier et al., 2015). This is significant as these berm reservoirs are known to impact the perception of local water quality through their reintroduction into the water column via wave action (Safaie, Weiskerger, Nevers, Byappanahalli, & Phanikumar, 2021; Vogel et al., 2016). These relocated sand populations can result in an overestimation of fecal contamination and unnecessary beach closures, as current beach monitoring methods cannot differentiate these long-term environmental survivors from recent fecal input.

E. coli can inhabit a large number of environmental niches due to its expansive pangenome, which consists of the total collection of all genes within a species. The pangenome is comprised of a ‘core genome’, coding for basic cellular processes found in all members of the

species, which is estimated to contain 1,500 – 2,200 genes (Gordienko, Kazanov, & Gelfand, 2013; Kaas, Friis, Ussery, & Aarestrup, 2012; Lukjancenko, Wassenaar, & Ussery, 2010; Rasko et al., 2008; Touchon et al., 2009), and an 'accessory genome' that contains genes present in some but not all members of a species (Medini et al., 2005). Due to its highly adaptive nature, the pangenome of *E. coli* has been estimated to be infinite in size (Lapierre & Gogarten, 2009), as the number of accessory niche-associated genes discovered increases as isolates from more diverse environments are sequenced. It is this variation within the accessory genome that allows *E. coli* to be associated with multiple habitats. Variance in gene ownership results in the differential survival of members of the same species within a niche, with certain genotypes being more fit for survival.

Several ecotypes of *E. coli* have been identified through comparing the distribution of accessory genes among isolates from various sources. For example, a comparison of commensal and pathogenic whole genome sequences (n=55) revealed that certain metabolic capabilities corresponded to particular intestinal and extraintestinal niches, with the majority of accessory genes identified belonging to pathways for the utilization of niche-associated nutrient sources (Monk et al., 2013). A similar niche-associated trend was also observed in the distribution of plasmid-born genes in *E. coli* isolated from livestock farms and wastewater treatment influent (Shaw et al., 2021). Some limited work has gone into identifying the beach ecotype genes that allow for the survival of *E. coli*. Targeted PCR screening of over 200 strains revealed that select accessory genes involved in iron acquisition, complement resistance/surface exclusion, and biofilm formation were enriched in isolates collected from surface water and sediment compared to those collected from host-associated habitats (Tymensen et al., 2015). Additionally, a pangenomic comparison of environmentally-associated isolated cryptic clade members to

commensal and pathogenic strains revealed a set of environment specific genes for diol utilization, an energy substrate, and lysozyme production, which allows for the hydrolysis of cell walls (Luo et al., 2011). Collectively this work suggests that *E. coli* that are able to survive for prolonged periods in the environment contain accessory genes distinct from those strains that are generally more host-associated.

In addition to the ownership of ecotype related genes, certain phylotypes of *E. coli* are more commonly found in the environment, including the cryptic clades or the phylogroup B1. Members of the cryptic clades are recovered from open-environments such as freshwater beach sand rather than from hosts (Walk et al., 2009) and are genetically and phenotypically distinct from the main subgroups of *E. coli* (A, B1, B2, C, D, and E) (Luo et al., 2011; Touchon et al., 2009). Laboratory experiments have demonstrated that members of these cryptic clades are able to replicate at lower temperatures and form biofilms more readily than host-associated strains (Ingle et al., 2011), while members of the B1 phylogroup have been shown to persist for longer in freshwater beach sand (Rumball, Mayer, & McLellan, 2021) and estuary water (Berthe et al., 2013) compared to members of other phylogroups. In general, B1 has been the most common phylotype recovered from several host-independent environments including freshwater beach sand (Rumball et al., 2021; Walk et al., 2007), surface water (White et al., 2011), dairy manure lagoons (Howard et al., 2017), and soil (Nandakafle et al., 2017). Collectively these results suggest that members of the cryptic clades and B1 are genetically predisposed for survival in select secondary environments.

The aim of this work was to examine the genetic factors in *E. coli* linked to survival in beach sand. We employed an *in situ* population level approach, as previous work had only focused on screening for specific genes or examining a small number of isolates from select

phylogroups, with experiments conducted mainly under laboratory conditions. We used *E. coli* originally isolated from sand, sewage, and gull waste and seeded these populations (n=188 per original source) into sand and buried them in microcosms in the backshore beach of Lake Michigan for 45 days. Results showed an overall average 3-log reduction in *E. coli* densities, with a concurrent shift in the population from a mix of phylotypes to a greater dominance of A and B1 phylotypes in the surviving populations (Rumball et al., 2021). In this work, a broad array of genes associated with surviving populations were identified, with approximately half of these common to isolate collections from all three sources, across multiple phylotypes, demonstrating the widespread nature of this survival ability.

Results

Identification of genes enriched in microcosm survivors

We examined the gene content of *E. coli* that demonstrated long-term survival in beach sand by identifying loci that increased in relative frequency in populations recovered from the end of the microcosm experiment (T45, microcosm output) compared to the initial (T0, microcosm input) population. We expected that the T0 population contains the entire pangenome, with the isolates recovered after the 45-day deployment to contain a subset of the pangenome, with the accessory genes of the surviving population enriched in relative abundance. We identified specific loci associated with survival by comparing the average depth of coverage of short sequence reads mapped to a scaffold of accessory genes from 21 *E. coli* genomes (Figure 3.1, Figure S3.1). The strains used for the scaffold were highly representative of our isolate collection, as an average of 90% of the total reads from the composite samples of pooled isolates were mapped (Table S3.1). We examined three separate isolate collections and found

1,765, 2,024 and 1,777 loci were enriched by >25% in sand, gull, and sewage output isolate collections, respectively, with 991, 1,763 and 357 of these loci enriched >50% in survivors. Overall, a total of 4,229 unique loci were associated with long-term surviving populations (Table 3.1).

Functions Associated with Environmental Survival

The loci enriched >25% in the microcosm output samples corresponded to 440 unique functions, with a total of 198 of these functions shared across sand, gull, and sewage output isolate pools (Table 3.1, Figure S3.2). A total of 59 of these functions were enriched >50%. The enriched functions that were shared among all three isolate pools were considered survival-associated as they are widely distributed across isolate collections (i.e., populations) from the three sources. Protein family classification of the 198 survival-associated functions revealed that the majority were associated with metabolic enzymes and transporters (Table 3.2). The enriched metabolism enzymes were largely associated with carbohydrate and amino acid processing, while the enriched transporter functions corresponded to membrane transport (including ABC transporters, phosphotransferase systems and bacterial secretion) and signal transduction functions (including two component-systems). For these functions, the specific loci that were enriched were examined (Figure 3.2). The sand and gull microcosm output populations had a higher number of survival-associated genes that were not necessarily shared with sewage populations (Figure 3.3, Figure S3.2).

Association of enriched survival genes with specific phylotypes

Members of certain phylotypes, such as B1, A, and the cryptic clades, have been noted to survive for longer in secondary habitats than other phylotypes. In these microcosm experiments, B1 and A increased in relative abundance by the end of the experiment (Rumball et al., 2021).

This led us to assess if the survival enriched functions were associated with any particular phylotypes. There was a small subset of 16 functions that were associated with a combination of A, B1, and/or cryptic clade members (Table 3.3), that were mainly associated with transport, replication and repair mechanisms. It was also observed that several phylotype specific accessory genes from microcosm input A, B1, and the cryptic clades members were not enriched in the microcosm output, demonstrating that not all members of these groups are capable of long-term environmental survival.

The majority of the enriched functions were not associated with a particular phylotype but were distributed among members of multiple phylogroups. However, only 25% of all survival enriched genes were found in members of B2, and these loci were not exclusive to B2, suggesting that the accessory pangenome of the surviving populations were most frequently found in non-B2 phylotypes. There was also a set of survival enriched functions (n=13) that were previously considered environmentally associated based on the genomic analysis of cryptic clade members isolated from the same freshwater beach sand ecosystem (Table S3.2 and S3.3). These functions were largely associated with transport and metabolic enzymes and were associated with multiple phylotypes in addition to the cryptic clades.

Identification of complete or nearly complete pathway modules

Individual gene functions within a pathway coordinate to form functional units called modules. We found several complete or nearly complete (missing one gene) modules enriched in microcosm output samples (Table 3.4, Table S3.4). These included the module for betaine biosynthesis, which was enriched >25% in sewage output samples and enriched >50% in sand and gull survivors; as well as two related modules for GABA biosynthesis and the GABA shunt, which where enriched >25% in sand and sewage survivors, and enriched >50% in gull survivors.

Additionally, there was one module that was only enriched in sand and gull microcosm output populations (>50%), which coded for histidine degradation.

In effort to determine if there were any phylotype associations among the survival-associated modules, the genomes of 33 *E. coli* from various phylogroups and sources were assessed, including those used to form the mapping scaffold and those used in the 2011 Luo study that compared host-associated and cryptic clade members (Table S3.4). It was determined that the modules for betaine biosynthesis, GABA biosynthesis, and the GABA shunt were generally more associated with non-clade phylotype members. It was also observed that the GABA biosynthesis module was not detected in the B2 strains included in this analysis. Lastly it was determined that the sand and gull associated module for histidine degradation was less widely distributed and was associated with members of some of the more uncommon phylogroups including E, F, and the cryptic clades.

Discussion

Population genomics approach to identify survival associated accessory genes

The ability of *E. coli* to survive in secondary environments varies based on the presence of survival-related genes that enable this organism to overcome periodic selection events (i.e., shifts in osmatic pressure or changes in nutrient availability). The genes associated with survival were identified in this study through a novel, large scale population genomics approach, with the characterization of a pool of ~564 isolates from the beginning and end of microcosm deployments. This work provided a realistic simulation of the fate of *E. coli* in the beach environment with populations from sources that are known to impact water quality (sand, sewage, and gull waste). *E. coli* isolates capable of long-term survival increased in relative frequency by the end of the microcosm experiment. By using isolates from three different

sources in the microcosm experiments, we were able to initially determine that *E. coli* could survive for weeks in beach sand, regardless of the original source. Therefore, it did not appear that isolates obtained from a primary environment (gull waste) are less suited for long term survival than isolates obtained from a secondary environment where selection has already occurred to some degree (beach sand or sewage). In this study, the redundant identification of loci enriched in output samples from all three isolate collections supports this conclusion.

Distribution of survival-associated functions across phylotypes

Analysis of enriched loci revealed that multiple functions were associated with surviving strains, with various levels of distribution, some being common among all or large phylogenetic subsets (i.e., shared among non-clade phylogroup members) while others were more limited to members of a specific subset of phylogroups (i.e., members of A, B1, and/or the cryptic clades). These findings were consistent with the previous phylotypic assessment of the microcosm survivors, which demonstrated that select members from each of the phylogroups were capable of long-term environmental survival in beach sand (Rumball et al., 2021). Previous to this work, it had been unclear if members of all phylogroups were capable of environmental survival, as it was undetermined if their presence in the environment was due exclusively to recent fecal deposits. However, the results of this genomic analysis combined with previous environmental observations suggest that members of all phylogroups can survive for prolonged periods in secondary environments if they contain certain functions.

Though select members of all phylogroups are capable of initial environmental survival, there are a higher proportion of some phylogroups, such as B1 and the cryptic clades, that have been shown to have enhanced environmental survival and secondary habitat associations. For instance, it was observed that the frequency of A and B1 phylogroup members increased in

populations by the end of the microcosm experiment, signifying that members from these groups had out-survived and/or out-propagated members from other phylogroups. This increased ability to survive is consistent with the previous observation that B1 is the most common phylotype recovered from multiple secondary environments, such as in the beach sand (Rumball et al., 2021; Walk et al., 2007). In this analysis functions that could be contributing to this differential survival were identified, with several survival enriched functions for cellular transport, replication and repair, and metabolism enzymes exclusively associated with A, B1, and/or cryptic clade members.

This distribution of functions is somewhat consistent with the phylogeny of *E. coli*, as A and B1 are sister groups (Jaureguy et al., 2008; Lecointre et al., 1998), thus are expected to contain some related genetic advantages. However, members of the cryptic clades have been shown to share several survival capabilities with members of A and B1, despite not being demonstrated to exchange genetic material with the main phylogenetic groups of *E. coli* and to be genetically distinct (Luo et al., 2011; Walk et al., 2009). This suggests that these functions may have been selected for independently in the A/B1 members compared to those belonging to the cryptic clades, with the ownership of the functions potentially providing survival advantages in secondary environments.

The relationship between the ownership of these functions and enhanced survival was further supported by the fact that not all members of the survival-associated phylogroups (A/B1/cryptic clade) increased in frequency by the end of the microcosm experiment. This was observed as several genes exclusive to the A/B1/cryptic clade members were detected in the initial isolate collections but were not enriched in the surviving populations. This indicated that there were intra-phylotype differences among members of these groups, such as gene ownership,

that were affecting the ability for prolonged environmental survival. This intra-phylotype variation has been previously observed to affect host-specificity, with the presence or absence of specific genes shown to have a greater effect on habitat range than those differences related to the clonal frame (White et al., 2011).

It has been suggested that intra-phylogroup variation may allow for a wider range of phylogroups to survive in secondary environments than expected. For example, despite their known association and high abundance in host-associated habitats, members of phylogroups D/E have been shown to be consistently recovered at low frequencies at multiple freshwater beaches (Rumball et al., 2021; Walk et al., 2007), with select members recovered during winter sampling and after 320 days of incubation in sand microcosms (Rumball et al., 2021). Further intra-phylotype assessment of these groups are needed to determine the factors that allow for enhanced survival.

There was also a set of survival-associated genes (Table S3.2) that were previously proposed as being environmentally-related, as they were identified through comparing the genomic content of cryptic clade strains that were mainly collected from the freshwater beach sand ecosystem, to human-associated pathogenic and commensal strains from phylogroups A, B1, and B2 (Luo et al., 2011). These shared survival related functions were largely associated with metabolism and cellular transport, and were found distributed in members belonging to phylogroups other than the cryptic clades. Interestingly, many of the functions identified in the assessment of theses cryptic clade members, that were previously described as environmentally-associated, were not enriched in the surviving populations from this study. This may be a result of these functions being more associated with the phylogeny of the cryptic clades rather than

their ability for survival, or may be a result of the low number of cryptic clade isolates included in this study, which resulted in a low level of gene enrichment.

Phylogenetic associations of survival enriched pathway modules

Upon further inspection of the survival-associated functions it was determined that there were several pathway modules that were widely distributed across non-cryptic clade members (A, B1, D, and F), and included those for betaine biosynthesis, GABA biosynthesis, and the GABA shunt pathways (Table 3.4). These survival-enriched modules coded for functions that aid in environmental survival. For example, the betaine biosynthesis pathway produces compatible solutes in response to changing osmotic conditions (Eshoo, 1988; Landfald & Strom, 1986).This pathway is expressed under both aerobic and anaerobic conditions (Landfal & Strom, 1986), thus allows its owner to adapt to secondary environments outside of the gut. The other generally distributed enriched pathways for GABA biosynthesis and the GABA shunt modules play an important role in pH homeostasis and intracellular nutrient recycling and nutrient use flexibility. These pathways allow for putrescine, a molecule that typically plays a role in cell proliferation and normal cell growth in *E. coli* (Tabor & Tabor, 1984), to be used as a carbon and nitrogen source during times of nutrient limitation (Kurihara et al., 2005, 2008). The precursor molecules in these pathways are produced by a wide range of organisms (including bacteria, plants, and animals) and are highly prevalent when cells are proliferating under stressful conditions (Tabor & Tabor, 1984), resulting in a high concentration of this compound in secondary environments. The ownership of these modules allows *E. coli* to utilize these abundant biomolecules for survival, aiding in the transition from a state of cellular growth and division to a state of survival and adaption. Combined, these non-clade enriched modules aid in the survival of non-clade

strains of *E. coli* through offering means to adapt to the dynamic nature of secondary environments.

There were further phylo-specific patterns in the distribution of survival related functions and the B2 phylotype. Overall, 3/4 of the survival related functions were not found in the members of B2 assessed, with the survival enriched GABA biosynthesis module absent in each of the B2 members assessed (0/5) (Table 3.4). The distribution of these pathways reflects the phylogeny of *E. coli*, as members of B2 are genetically distinct from the other phylogroups, as it holds the most basal position on the phylogenetic tree, compared to A & B1 which branch to form separate sister groups (Jaureguy et al., 2008; Lecointre et al., 1998), and the cryptic clades which are distant monophyletic groups (Luo et al., 2011). This phylogenetic distribution is consistent with the known patterns of environmental survivorship capabilities of these groups. Members of the more distant branches containing A/B1 and the cryptic clades have been shown to be more environmentally hardy (Rumball et al., 2021) and are more associated with secondary environments (Walk et al., 2007), compared with those that belong to the B2 & D/E phylogenetic branch, which have been shown to be generally more host-associated (Gordon, 2003), as well as recovered less frequently and shown to survive for shorter periods in secondary environments (Rumball et al., 2021; Walk et al., 2007). Recent pangenomic analysis reflects these differences in behavior, with members of B2 shown to contain exclusive metabolic genes that enable this strain to colonize the intestinal mucosa more readily than members of other phylogroups (Fang et al., 2018), compared to the survival enriched genes identified in this study that enable adaption to dynamic secondary environments.

The genetic relationship between gull and sand *E. coli* populations

While nearly 50% of the survival enriched functions were found in each source assessed (Table 3.1), we did note that the gull output samples had a broader array of survival enriched functions than the surviving sand or sewage isolate collections, with more functions exclusive to the gull survivors compared to the sand or sewage populations (Figure S3.2). It was also observed that the gull and sand surviving populations had significantly more functions in common than sewage and sand samples, suggesting that gull waste may be contributing more than sewage to bacterial burdens in the sand. This distribution also suggests that gulls may be acting as a vector of transport for those *E. coli* that are able to survive in the environment.

It should be noted that there may be more survival-associated functions then what was determined through this study, as the methods used relied upon observing shifts in populations throughout the experiment. For example, sand populations may have shifted less throughout the microcosm experiment then members from the other sources, as the initial population was composed of isolates randomly collected from the beach, which may have contained a large proportion of *E. coli* capable of long-term survival. Therefore, the distribution of this population may not change much throughout incubation, which would cause survival related genes in this group to remain unidentified.

Overall, our findings demonstrate that the ability to survive in secondary environments is impacted by gene ownership, with the distribution of survival-related functions shown to vary, including those more generally distributed (i.e., common among all non-clade) and those exclusive to more specific sets of phylogroups (i.e., members of A, B1, and/or the cryptic clades). In order to determine the true impact of these pathways on survival more work needs to be done, such as comparing the survival ability of those that own these pathways to those that do not. Once this work has been completed, pathway genes most associated with survival can then

be used as an indicator of long-term environmental survivors, offering a means to potentially improve current water quality monitoring techniques.

Methods

Microcosm experiment

A microcosm experiment, previously described in Rumball, Mayer, & Mclellan, 2021, was conducted during the summer of 2017, in which 188 *E. coli* isolates from each freshwater beach sand, sewage, and gull waste were seeded into beach sand at a concentration of 10^7 CFU per 100 g of sand, sealed in microcosm chambers (in triplicate), and buried in the backshore of the beach at Lake Michigan for 45 days. *E. coli* densities decreased from 10^7 CFU per 100 g of sand to an average of 10^4 CFU per 100 g of sand in this time frame (Rumball et al., 2021). Upon termination of the experiment *E. coli* were re-isolated from the microcosms by eluting 45 g of sand in 450 ml of sterile water. These samples were then shaken by inversion for 2 min and processed through 0.45 µm pore size nitrocellulose filters in aliquots (Millipore, Billerica, MA) and grown on modified mTEC (membrane-thermotolerant *E. coli*) at 35°C for 2 hr, then 44.5°C for 10 hr. Distinctly isolated colonies (~376 from each source) were then picked into MUG media (Thermo Fisher Scientific, NH), stored in three 96 -well plates, and grown at 35°C for ~12 hr. The samples were then stored in a 25% glycerol 75% MUG media solution at -80°C until sequencing preparation was conducted.

DNA Isolation & sequencing

Sequencing was conducted on a set of individual isolates from the microcosm experiment (n=14; Table S3.3) and on triplicate composite samples formed from 188 isolates from initial microcosm populations and 376 from each of the final populations of sand, sewage, and gull microcosm treatments (Table S3.1). Prior to DNA extraction, samples for sequencing were cultured from freezer stocks in MUG media at 35°C for ~12hr. Composite population samples were formed by pooling equal concentrations of each cultured isolate for each source and

timepoint, which was determined through OD readings at a wavelength of 600 nm on a BioTek plate reader (VT, USA.). DNA was extracted from the isolate and composite samples using phenol-chloroform extractions as recommended by the Joint-Genome Institute (Feil, Feil,& Copeland, 2012). The majority of the DNA sequencing was conducted at the Josephine Bay Paul Center at the Marine Biological Laboratory (Woods Hole, MA, USA), using the NextSeq Illumina[R] platform, with the exception of the individual isolate genomes for samples DC5 C10, DC5 C4, DC5 E12, and Gull1 B2, which were sequenced at the UWM School of Freshwater Sciences (Milwaukee, WI, USA) using the PacBio RSII platform.

Bioinformatic analysis

In an effort to enhance the identification of environmentally associated genes, several genome sequences from NCBI were downloaded and included in this analysis. These comprised of environmentally-associated cryptic clade members and host-associated commensal and pathogenic B2 genomes (Table 3.4, Table S3.3). These cryptic clade genomes were included in the mapping scaffold that was used to identify survival-related genes, with the B2 genomes included in the analysis of survival-associated gene modules across phylotypes. Prior to analysis, low-quality reads were removed from the sequenced individual isolates and composite samples with the 'iu-filter-quality-minoche' program in the illumine-utils v1.4.1 (Eren, Vineis, Morrison, & Sogin, 2013). The scaffold used for composite sample mapping was composed of 5/21, 6/21, 1/21, 1/21, and 8/21 isolates belonging to phylogroups A, B1, D, F, and cryptic clades (Table S3.3). The accessory genome from this set of isolates was identified with the Anvio pangenome Snakemake workflow (Eren et al., 2015; Köster & Rahmann, 2012), with those genes present in <20 of the genomes considered accessory. These accessory genes were then concatenated and used as a scaffold for the mapping of composite population samples with the Anvio metagenome Snakemake workflow (Eren et al., 2015; Köster & Rahmann, 2012). Through the Anvio interface both the pangenomic and metagenomic datasets were annotated with the COG database

(Tatusov, Galperin, Natale, & Koonin, 2000) using blastp (Altschul, Gish, Miller, Myers, & Lipman, 1990).

Genes associated with long-term surviving populations were identified through comparing the genomic content of composite samples from the beginning and end of the microcosm experiment. Strains that were better suited for environmental survival increased in prevalence by the end of the microcosm experiment, which resulted in a higher number of associated short-reads in the composite samples from the end of the experiment, compared to those from the initial timepoint. These differences were observed by mapping the composite samples to the scaffold of accessory genes with the Bowtie2 (Langmead & Salzberg, 2012) extension of the Anvio metagenomics snakemake workflow, with the mapping results normalized to the total number of reads in each quality filtered composite sample file. Those functions that had >25% average coverage in output composite samples compared to those from the input were considered survival-associated, with an additional cut-off of >50% enrichment used to assess the variation in the depth of coverage between sources. Those enriched functions that were shared among each sand, gull, and sewage output populations were considered strongly survival-associated.

The functions of these survival enriched genes were identified through GhostKOALA (KEGG Orthology and Links Annotation) (Kanehisa, Sato, & Morishima, 2016), using BRITE and pathway module analysis. The phylotype analysis of enriched pathway modules was conducted by assessing the presence of survival-related modules in the genome isolates used to form the scaffold, as well as those B2 genomes downloaded from NCBI with GhostKoala. Additional phylogenetic relationships among survival-associated functions were identified through assessing distribution of the enriched survival-associated gene functions in the Anvio pangenome.

Acknowledgements

This work was supported by the University of Wisconsin Sea Grant Program grant #NA10OAR4170070 and was conducted at the University of Wisconsin-Milwaukee School of Freshwater Sciences. We would like to thank the members of the McLellan lab and Newton lab for their assistance in microcosm deployment and recovery, and for strain isolation, specifically Katie Alexander, Adelaide Roget, Shuchen Feng, HannahRose Mayer, and Brigid Meyers. We also thank Dr. Hilary Morrison and Angie Schmoldt for their assistance with the sequencing of samples.

CHAPTER 3: Tables and Figures

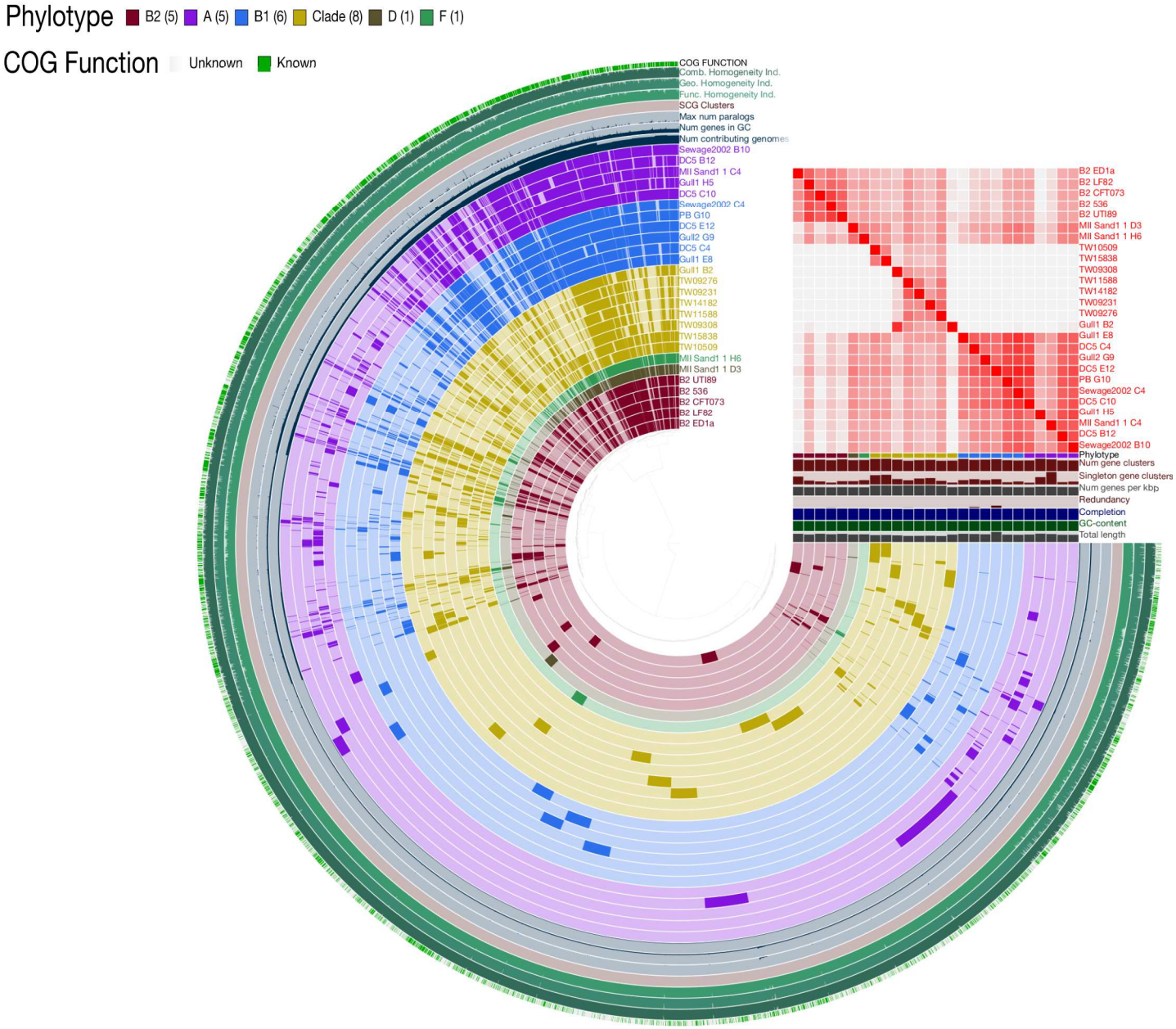


Figure 3.1. Pangenomic analysis of accessory genes from 25 *E. coli* genomes. Each row represents a different *E. coli* genome, with shaded regions around each ring indicating the ownership of a gene. There are fewer accessory genes belonging to member of B2 than the other phylogroups. These genomes were used to form the scaffold for mapping composite samples from the microcosm experiment.

Table 3.1. The number of survival enriched genes and corresponding unique functions in microcosm output samples.

Enriched > 25%		
Source isolate collection	# Loci Match[1]	# Unique Functions[2]
Sand	1765	356
Gull	2024	326
Sewage	1777	306
Total unique	4229	440
Total shared among all sources	76	198
Enriched >50%		
Source	# Loci Match	# Unique Functions
Sand	991	306
Gull	1763	118
Sewage	357	300
Total unique	2499	403
Total shared among all sources	16	59

[1] # Loci Match is the total number of scaffold genes enriched in each output source, including those not annotated with KEGG; [2]# Unique Functions is the number of non-redundant KEGG functions associated with the matched loci, and does not include those that were unable to be annotated.

Table 3.2. BRITE classification of functions enriched >25% in each sand, sewage, and gull microcosm output.

BRITE[1] category	# of KO
Protein families: metabolism	
ko01000 Enzymes	87
ko01001 Protein kinases	4
ko01002 Peptidases and inhibitors	6
ko01003 Glycosyltransferases	2
ko01005 Lipopolysaccharide biosynthesis proteins	3
ko01011 Peptidoglycan biosynthesis and degradation proteins	1
ko01004 Lipid biosynthesis proteins	1
ko01007 Amino acid related enzymes	3
Protein families: genetic information processing	
ko03000 Transcription factors	5
ko03019 Messenger RNA biogenesis	1
ko03012 Translation factors	1
ko03110 Chaperones and folding catalysts	2
ko04131 Membrane trafficking	1
ko03032 DNA replication proteins	1
ko03036 Chromosome and associated proteins	1
ko03400 DNA repair and recombination proteins	4
ko03029 Mitochondrial biogenesis	1
Protein families: signaling and cellular processes	
ko02000 Transporters	63
ko02044 Secretion system	9
ko02022 Two-component system	3
ko02035 Bacterial motility proteins	6
ko04147 Exosome	1
ko02048 Prokaryotic defense system	9

[1] BRITE is a functional hierarchy of KO calls and is used to determine protein family classification. The majority of the survival-enriched KO are associated with metabolic enzymes & transporters involved in signaling and cellular processes; detailed annotation can be found in supplemental dataset 1.

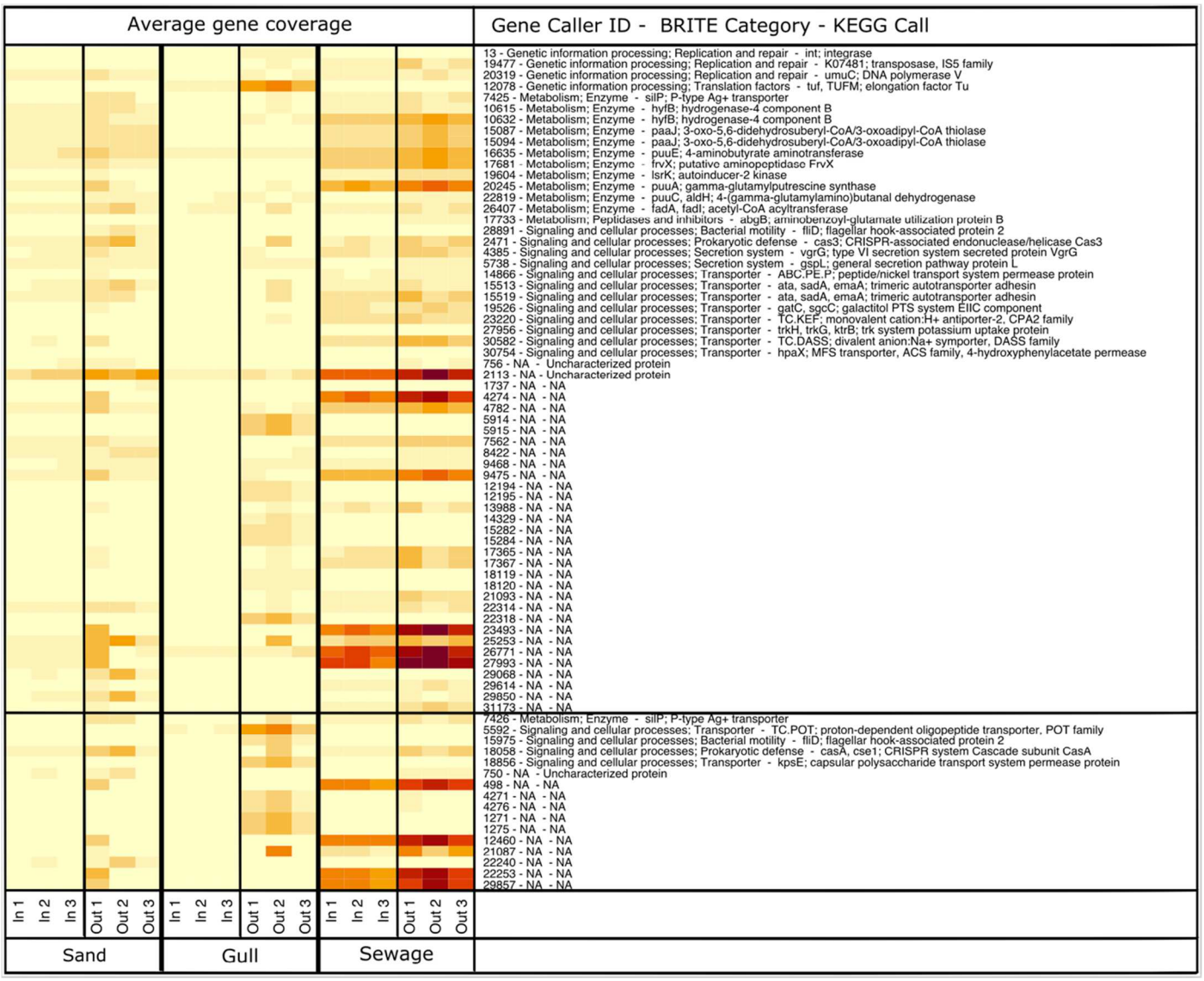


Figure 3.2 Heatmap of the average coverages of shared enriched loci. 76 loci (from the total 198 shared functions) were enriched >25% across all output sources, with 15 enriched >50% (those functions below the horizontal black line)

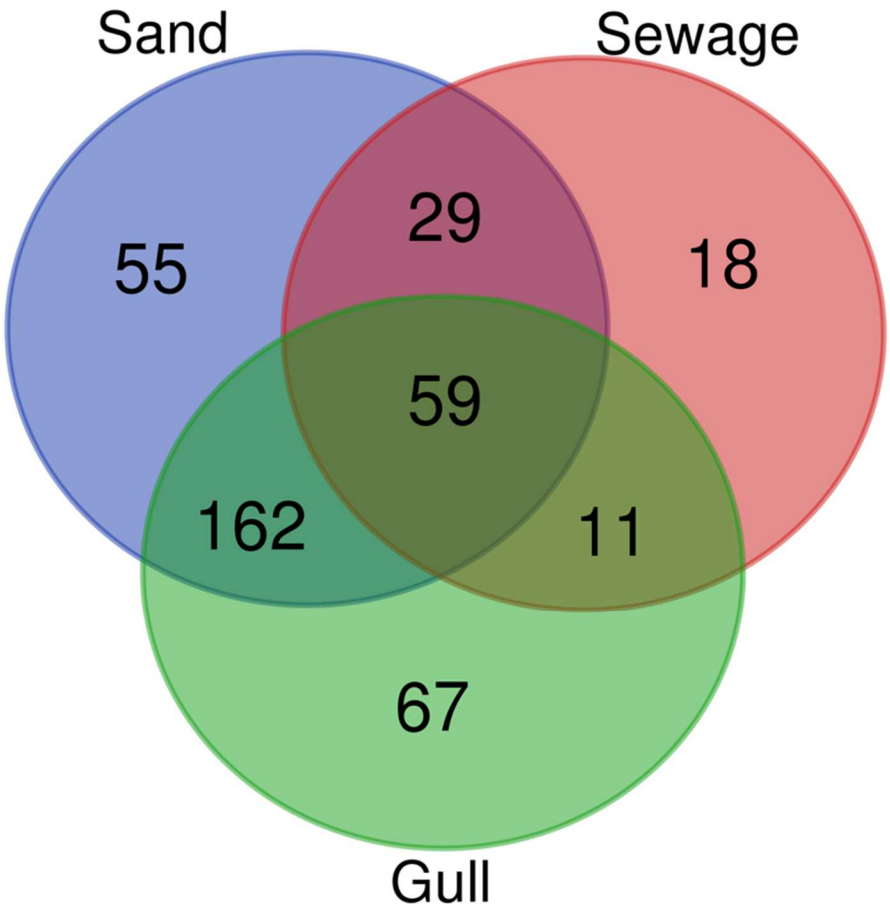


Figure 3.3. The source distribution of functions enriched >50% in the microcosm output. This analysis included only those genes that were able to be annotated with KEGG. The gull samples overall have the highest number of enriched genes, and had the highest number of functions in common with sand populations.

Table 3.3. Phylotype associations of survival-enriched genes.

Associated[1] Groups	Number of genes	KO #	Protein Family	Family Category	KO Function
A B1 Clade	6	K07481	Genetic information and processing	Replication and repair	transposase, IS5 family
		K07484	Genetic information and processing	Replication and repair	transposase
		K00276	Metabolism	Enzymes	AOC3, AOC2, tynA; primary-amine oxidase
		K00074	Metabolism	Enzymes	paaH, hbd, fadB, mmgB; 3-hydroxybutyryl-CoA dehydrogenase
		K07452	Signaling and cellular processes	Prokaryotic defense system	mcrB; 5-methylcytosine-specific restriction enzyme B
		K19147	Signaling and cellular processes	Prokaryotic defense system	mcrC; 5-methylcytosine-specific restriction enzyme subunit McrC
B1	4	K03654	Genetic information and processing	DNA repair and recombination proteins	recQ; ATP-dependent DNA helicase RecQ
		K02335	Genetic information and processing	DNA replication, repair, and recombination proteins	polA; DNA polymerase I
		K07452	Metabolism/Signaling and cellular processes	Prokaryotic defense system	mcrB; 5-methylcytosine-specific restriction enzyme B
		K11192	Signaling and cellular processes	Transport	murP; N-acetylmuramic acid PTS system EIICB component
A	2	K20543	Signaling and cellular processes	Transport	bcsC; cellulose synthase operon protein C
		K03455	Signaling and cellular processes	Transport	TC.KEF; monovalent cation:H+ antiporter-2, CPA2 family
Clade	2	K03455	Signaling and cellular processes	Transport	TC.KEF; monovalent cation:H+ antiporter-2, CPA2 family
		K07679	Signaling and cellular processes	Two-component system	evgS, bvgS; two-component system, NarL family, sensor histidine kinase EvgS
A Clade	1	K23254	Metabolism	Enzymes	silP; P-type Ag+ transporter

B1 Clade	1	K03498	Signaling and cellular processes	Transport	trkH, trkG, ktrB; trk system potassium uptake protein

[1] The phylotype-associations of the enriched genes were identified through assessing the distribution of the survival enriched genes in the isolate pangenome.

Table 3.4 Distribution of enriched pathway modules.

Genome Name	Source	Phylotype	Luo et al. 2011 study	M00555 Betaine biosynthesis (Enriched completely in all)	M00136 GABA biosynthesis (missing 1 KO in all)	M00027 GABA shunt (missing 1 KO in all)	M00045 Histidine degradation (missing 1 KO Sand & Gull output)
DC5_B12	freshwater beach sand	A	NO	YES	YES	YES	NO
DC5_C10*	freshwater beach sand	A	NO	YES	YES	YES	NO
DC5_C4*	freshwater beach sand	B1	NO	YES	YES	YES	NO
DC5_E12*	freshwater beach sand	B1	NO	YES	YES	YES	NO
Gull1_E8	gull waste	B1	NO	YES	YES	YES	NO
Gull1_H5	gull waste	A	NO	YES	YES	YES	NO
Gull2_G9	gull waste	B1	NO	YES	YES	YES	NO
MII_Sand1_1_C4	freshwater beach sand - microcosm output	A	NO	YES	YES	YES	NO
MII_Sand1_1_D3	freshwater beach sand - microcosm output	D	NO	YES	YES	YES	NO
MII_Sand1_1_H6	freshwater beach sand - microcosm output	F	NO	YES	YES	YES	YES
PB_G10	freshwater beach sand	B1	NO	YES	YES	YES	NO
Sewage2002_B10	sewage	A	NO	YES	YES	YES	NO
Sewage2002_C4	sewage	B1	NO	YES	YES	YES	NO
Gull1_B2*	gull waste	clade or E	NO	NO	NO	YES	YES
TW09231	freshwater beach	clade III	YES	NO	YES	YES	NO
TW09276	freshwater beach	clade III	YES	NO	YES	NO	NO
TW09308	freshwater beach	clade V	YES	NO	NO	YES	YES
TW10509	feces	clade I	YES	YES	YES	NO	NO
TW11588	soil	clade IV	YES	NO	YES	YES	NO
TW14182	freshwater beach	clade IV	YES	NO	NO	YES	NO
TW15838	freshwater sediment	clade I	YES	YES	YES	NO	NO
ED1a	Human feces-commensal	B2	YES	NO	NO	YES	NO
CFT073	Human blood - pathogenic (ExPEC)	B2	YES	YES	NO	YES	NO
LF82	Ileal mucosa - pathogenic (AIEC)	B2	NO	YES	NO	YES	NO
UTI89	Human urine - pathogenic (ExPEC)	B2	YES	YES	NO	YES	NO
536	Human unrine - Pathogenic (ExPEC)	B2	YES	YES	NO	YES	NO
MG1555	Human feces - commensal	A	YES	YES	YES	YES	NO
HS	Human feces - commensal	A	YES	YES	YES	YES	NO
SE11	Human feces - commensal	B1	YES	YES	YES	YES	NO

IAI1	Human feces - commensal	B1	YES	YES	YES	YES	NO
Sakai	Human feces - pathogenic (EHEC)	E	YES	YES	YES	YES	NO
EDL933	Ground beef - pathogenic (EHEC)	E	YES	YES	YES	YES	NO
UMN026	Pathogenic (UPEC)	D	NO	YES	YES	YES	NO
IAI39	Pathogenic (ExPEc)	F	NO	YES	YES	YES	YES

* Sequenced with PacBio

CHAPTER 4
Summary

Research topic overview

The results from this work demonstrate that *E. coli* isolated from multiple sources are capable of long-term environmental survival, and that this survival is associated with several functions, some of which are widely distributed across phylogeny (i.e., exclusive to non-clade members), while others are more niche and associated with a specific subset of phylogroups (i.e., A, B1, and/or cryptic clade members). These long-surviving populations have been previously observed through routine monitoring (Alm et al., 2006; Hartz et al., 2008; R. L. Whitman & Nevers, 2003), with microcosm experiments demonstrating that select strains of *E. coli* are capable of long-term environmental survival (Alm et al., 2006; Hartz et al., 2008). These long-surviving *E. coli* were originally determined to be genetically distinct from those host-associated populations using low-resolution forms of analysis based on genetic markers including DNA microarrays (Kon et al., 2007; Oh et al., 2012), DNA fingerprinting (Ishii et al., 2007; R. L. Whitman & Nevers, 2003) and REP-PCR (Beversdorf et al., 2007); and were further characterized by phylogeny using Clermont phylotyping (Gordon et al., 2008; Walk et al., 2007, 2009). Previous to this study, exploration into the genetic determinants of survival had remained limited. Only a small number of environmentally-associated cryptic clade isolates were assessed through whole genome sequencing (Luo et al., 2011), with no work done on the genetic factors associated with survival in larger populations that would mimic recently deposited *E. coli* in secondary environments.

This project expanded beyond the limitations of previous work by experimentally assessing the genetic factors associated with survival using an unprecedented number of isolates (n = ~600) from a unique set of sources that mimic *E. coli* deposition at the beach (sand, sewage, and gull waste). This work offers the most comprehensive description of the genetic

determinants that are associated with survival and includes phylogenetic and population level genomic assessment. This genomic analysis expanded beyond members of the previously assessed cryptic clades, but rather took a holistic approach by assessing the genomic contents of whole populations consisting of many phylotypes. This allowed for survival associated pathways to be identified, which is key to developing methods to differentiate long-term surviving *E. coli* from those that are associated with recent fecal pollution. The identification of survival-associated genes can allow for the development of molecular tools that can differentiate potential-long term survivors from host-associated strains, which would help improve the accuracy of current beach monitoring methods.

New perspectives on *E. coli* survival

This work has provided unprecedented insight into the survival of *E. coli* at beaches through combining classical microbiology techniques, such as cell culturing and microcosms, with high-resolution molecular biology tools, such as multiplex PCR and next-generation DNA sequencing. The work outlined in chapters two and three were presented as two publications; the first, which focused on assessing the phylotypic composition of long-term survivors and the factors that influence their survival, was published in Applied Environmental Microbiology in 2021; and the second, which focused on the genetic determinants of survival with an assessment of pathways enriched in long-term survivors, is being prepared for submission in the summer of 2021. Additionally, a preliminary MLST analysis of portions of the rpoS and uidA genes of the ~600 isolates from the microcosm input was included in appendix C. This work demonstrated that isolates from sand and gull clustered more closely than those from sewage, with members of B1 and A phylogroups more widely distributed across the phylogenetic tree than members of B2,

D/E, and the cryptic clades. Together these results describe a phylogenetic and accessory gene profile of *E. coli* capable of long-term environmental survival.

This demonstration of the capability of *E. coli* to survive for long periods in the environment brings into question the accuracy of *E. coli* as a fecal indicator at beaches. Previous to this work it had been unknown if *E. coli* survival was limited to a specific subset of strains from a particular source or phylogeny, or how different environmental factors affect these abilities. The work done in chapter two addresses these short-comings through assessing the survival of *E. coli* in beach sand using *in situ* approaches including microcosm experiments and beach surveys, with additional in-lab assessments of survival related phylogeny with Clermont phylotyping, bacterial community, and nutrient analysis. It was determined that *E. coli* from each source tested (sand, sewage, and gull waste) were capable of long-term survival in the beach sand, and that those *E. coli* that belonged to phylotypes A and B1 survived longer than members of other phylogroups in the microcosm experiment, with B1 being the most common phylotype recovered from a survey of beach sand from multiple freshwater beaches. Nutrient availability and competition were shown to be the biggest drivers of survival, with isolates capable of overcoming these limitations able to survive in sand for extended periods of time (8-14 weeks). These findings allow for an improved interpretation of beach monitoring results, as a high prevalence of B1 and A phylotypes may suggest a beach is affected by long-term survivors rather than *E. coli* from recent fecal pollution. This work also suggests that nutrient mitigation strategies, which limit nutrient pollution at beaches, may help control bacterial growth and improve the water quality at chronically contaminated beaches. Collectively this work suggests that employing these types of population assessment techniques can improve the interpretation of

monitoring results and allow the opportunity for more site-specific tailored solutions to be developed to improve water quality.

Despite the past decades' characterization of sand associated *E. coli* populations, the genetic traits that allow for their survival here had remained unknown. The work presented in chapter three has provided unprecedented insight into the genetics of survival. Populations of *E. coli* from before and after the microcosm experiment (as described above) were analyzed using next-generation DNA sequencing. Through the use of Anvio and GhostKoala, it was determined that there were a total of 198 functions enriched >25% in all output populations assessed, with 59 of these enriched >50%. These enriched functions were largely associated with metabolic enzymes and transporters. It was determined that different groups of *E. coli* had various survival strategies, with a subset of functions found widely-distributed among non-clade members, and a smaller separate set of functions associated with a smaller subset of survival associated phylotypes including A, B1, and the cryptic clades. The widely distributed survival-associated genes coded for functions that aid in general environmental survival including osmotic regulation via betaine biosynthesis and nutrient recycling through GABA biosynthesis and the GABA shunt, while the phylogroup-associated functions were generally associated with more specific strategies, largely metabolic enzymes, transporters, and prokaryotic defense mechanisms. This work is significant as these survival-associated genes can be used as markers to identify *E. coli* capable of long-term survival; which can be useful when assessing the cause of chronically contaminated beaches.

Interpretation of the research findings in the context of previous work

Although *E. coli* is one of the most well-studied organism on earth, with a robust library of associated research, a relatively small amount of work has gone into understanding its

behavior in secondary environments. Thus far the results of this study appear to be congruent with the findings of the research community. For example, Walk *et al.* 2007 determined B1 to be the most common phylotype recovered from freshwater beaches along the eastern shores of Lake Michigan. In combination with this study, which focused on beach samples collected from the Western shores of Lake Michigan that showed members of B1 dominating each site of the beach surveys conducted, a potential lake wide phylotype distribution pattern was demonstrated. This secondary-environment phylogroup association has been proposed to be partially due to the known behavior of members of this group in hosts, with members of B1 demonstrated to leave the gut of hosts more readily than members of other phylogroups (Martinson et al., 2019). Our findings are complementary to this idea, as it showed B1 to be more cosmopolitan and more genetically adapted to survival in secondary environments than members of other groups. This wide range of habitats may be a result of the wide genomic variance within this phylogroup, which was seen in the preliminary MLST analysis in Appendix D. B1 members were the most widely distributed phylogroup in the tree, indicating it is the most genetically variable phylogroup. This wide range of genotype varieties provides enhanced odds of being able to endure a selection event.

Additionally, this work is impactful as it assessed the population dynamics of *E. coli* under real environmental conditions, in contrast to previously conducted microcosm experiments that were limited to focusing on the survival of single isolates under more controlled setting such as in-lab environments or in less representative medias such as in autoclaved sand. This type of population level research is important, as multiple sources have been determined to be contributing to *E. coli* loads at the beach such as sewage (Goodwin, Matragrano, Wanless, Sinigalliano, & LaGier, 2009; Korajkic et al., 2011; Yamahara, Sassoubre, Goodwin, & Boehm,

2012) and gull waste (Converse et al., 2012; Fogarty et al., 2003). Through this work some of the potential short comings of modern water quality monitoring were observed, as it was demonstrated that *E. coli* from multiple sources were able to survive in the beach sand. As survival was observed to vary based on phylotype association, further phylogenetic assessment of populations of *E. coli* from sources of contamination may be useful in assessing their potential impact on the local sand *E. coli* populations and perceived water quality.

Though previous work assessing the genetic factors that allow for environmental survival had remained limited and focused only on cryptic clade members, the results from this wider assessment of survival were consistent. For example, the previous comparison of whole genomes of clade members isolated from freshwater beach sand to host-associated strains revealed that genes that allow for nutrient access and utilization were associated with environmental survivors, such as those for lysozyme production and the complete pathway for diol utilization (an energy substrate) (Luo et al., 2011). This work revealed a somewhat similar pattern of gene ownership in surviving populations, with the majority of survival enriched functions identified as being related to metabolism enzymes and cellular transport. Interestingly the majority of these survival-associated functions were associated with members from each phylogroup assessed, demonstrating that metabolism flexibility is important to the survival of all phylogroups of *E. coli* in secondary environments. Previous work has shown that the ownership of certain metabolic genes allows for the adaption to particular environmental niches (Alteri & Mobley, 2012; Monk et al., 2013; White et al., 2011), with this project being the first of its kind to study this niche and its genetic associations extensively.

Additionally, there were several pathways that were enriched in survivors that were exclusively found in non-clade members. This demonstrates that long-term environmental

survival is not limited to cryptic clade members, but also includes members of other phylogroups containing certain functions. This variation in the distribution of these genes among initial and surviving populations suggests that the beach is selecting for specific ecotypes that can withstand the environmental conditions here. Similar ecotype assemblies have been observed in other microbial populations, such as with *Prochlorococcus,* which had been shown to have distinct environmental distribution patterns within the species (Delmont & Eren, 2018). Once these beach sand ecotype specific genes have been verified experimentally to contribute to survival, they could be used to make markers of naturalized *E. coli*, similar to how other molecular markers such as GULL2 can differentiate gull waste related *Lachnospiraceae* from other strains (S. Feng et al., 2018).

Collectively, the findings of this project and of previous work supports the idea of environmentally associated *E. coli*, which are genetically suited for survival in secondary environments. These results challenge the suitability of the use of *E. coli* as an indicator of recent fecal pollution. As demonstrated not all *E. coli* in a population behave the same, with their survival ability generally being associated with phylotype membership and accessory gene ownership. *E. coli* monitoring results can still be useful in detecting acute pollution events; however, a closer assessment of populations is key to interpreting monitoring data, which could be done through phylogenetic assessment or using survival-associated gene markers.

Limitations of this work

Though this work has provided key insights into understanding *E. coli* survival in beach sand, it does have its limitations. Several of these limitations are associated with the methods used for isolate collection. For instance, though an unprecedented number of isolates were assessed in this work, the origin of these samples were limited to beaches on the western shores

of Lake Michigan. It is unknown if the patterns of survival observed here would be the same for beaches in other regions without further assessment conducted. Secondly this work relied upon typical *E. coli* phenotypes for species identification. For instance, to isolate *E. coli* from sand microbial communities the differential media modified mTEC was utilized, which identifies *E. coli* through a colorimetric change that detects the break-down of beta-glucuronidase, a sugar used exclusively by Escherichia members. While it is assumed that all wild-type *E. coli* can break down this sugar, mutants deficient in this ability have been detected sparsely (P. Feng, Lum, Chang2, & Hartman, 1991). Limiting the selection of isolates to only those that have this assumed universal genotype could potentially result in a loss of full genetic representation in this study, as atypically presenting *E. coli* were not included. Lastly the sample collection protocol did not account for those *E. coli* that may be in a persister state, in which the organism is in a dormant, viable but not culturable state (VBNC). Previous work has suggested that in harsh environments, such as at the beach, many bacteria may be in this state. These isolates would largely be unaccounted for in this study, which relied upon re-growing and culturing *E. coli* from sand samples. This especially could have had an impact on the samples collected during times of excessive environmental stress, such as those collected during winter or at the end of the year-long incubation study.

Another limitation of this work is the current narrow view of the environmental functions of the proposed survival-associated pathways. Survival-associated pathways were determined by observing an increase in frequency of them among microcosm survivors. However, it is unknown if the presence of these pathways themselves are aiding in survival or if they are simply a relic carried by strains that can survive. An *in situ* approach such as transcriptomics or gene knock-out experiments would need to be conducted to assess if these genes are being expressed

and aiding in survival in the environment. This would provide deeper insight into survival-associated gene expression and could offer an explanation into how isolates survive long-term in the beach sand. However, it should be noted that transcriptomics has its own limitations, as the complexity of environmental samples can interfere with the analysis of results, with complicating factors including extracellular mRNA from past gene expression and a complex microbial community.

Recommendations for further studies

Although this study has provided evidence that *E. coli* is capable of long-term survival in the beach sand, the ecological relevance of these findings still needs to be put in context. For instance, it needs to be determined if these proposed survival-associated genes are truly functioning to assist in survival or if they are a relic of phylotype association or other factors. The survival performance of the isolates that have these pathways should be compared to those that do not contain them through the use of *in situ* or laboratory microcosms. Additionally, the distribution of these genes should be assessed further in environmental samples, with correlations of *E. coli* prevalence and environmental conditions to be considered. As previously mentioned, transcriptomics is another method that could be deployed to explore the ecological relevance of survival associated genes. This would assess gene expression, opposed to gene presence/absence and abundances, which was used in this study. Assessing gene expression of *E. coli* during times of environmental stress would reveal what genes and functions are assisting in survival and would determine if the proposed environmentally-associated genes are truly aiding in survival.

Additionally, an in-depth analysis of the patterns of nutrient loading and distribution at beaches should be conducted. As seen in this study, nutrient availability limits the growth and

survival of *E. coli,* with increased nutrient levels resulting in increased growth. One common event that effects nutrient deposition at the beach is rain, which has been demonstrated to increase levels of *E. coli* and nutrients at the beach through run-off (Dada & Hamilton, 2016). Though some research has gone into assessing *E. coli* levels in beach water post-rain, little attention has been paid to how this affects *E. coli* levels in the sand and how rain-associated nutrient deposition can influence their survival here. This work would be useful in determining if a beach would benefit from being equipped with run-off mitigation strategies such as rain-gardens and drainage areas, which can decrease nutrient deposition.

Conclusion

This work has provided robust evidence to prove the existence of environmentally-associated populations of *E. coli*. It was demonstrated that *E. coli* from various sources are capable of long-term survival in beach sand, with an increase in the frequency of A & B1 phylogroup members by the end of the microcosm experiment, suggesting that *E. coli* belonging to these phylogroups are better suited for survival than members of other phylogroups. Genetic analysis revealed there to be 198 survival-enriched functions shared among sand, sewage, and gull output samples, with some functions being more widely distributed (i.e., among non-clade members) while other were associated with a more specific subset of phylogroups (i.e., A, B1, and cryptic clade members). Enriched functions were largely associated with metabolic enzymes and transporter proteins, suggesting that an enhanced ability to take up and use nutrients aids in survival here. It can be concluded that the universal cut-off values used for beach monitoring need to be modified to account for factors aside from fecal pollution that can increase *E. coli* levels at beaches; which would include wave action, geological/ anthropomorphic structures, nutrient input, and native sand populations (Strayer & Findlay, 2010). Interpretation of water monitoring results without

the consideration of these factors for each site could result in an over-estimation of fecal contamination at some beaches, and lead to unnecessary beach closures. The survival-associated genes identified through this work offer a starting point for the development of molecular-tools to differentiate long-term environmental surviving *E. coli* from those that are more fecal related.

REFERENCES

Alderisio, K. A., & DeLuca, N. (1999). Seasonal enumeration of fecal *coli*form bacteria from the feces of ring- billed gulls (Larus delawarensis) and Canada geese (Branta canadensis). *Applied and Environmental Microbiology*, *65*(12), 5628–5630. https://doi.org/10.1128/aem.65.12.5628-5630.1999

Alm, E. W., Burke, J., & Hagan, E. (2006). Persistence and Potential Growth of the Fecal Indicator Bacteria, *Escherichia coli*, in Shoreline Sand at Lake Huron. *J. Great Lakes Res. Internat. Assoc. Great Lakes Res*, *32*, 401–405. Retrieved from https://s3.amazonaws.com/academia.edu.documents/45411342/Persistence_and_Potential_Growth_of_the_20160506-12150-1e0xxck.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1520193165&Signature=rr0wqaN5YknoA5oA549iNYLQbBo%3D&response-content-disposition=inline

Alm, E. W., Burke, J., & Spain, A. (2003). Fecal indicator bacteria are abundant in wet sand at freshwater beaches. *Water Research*, *37*, 3978–3982. https://doi.org/10.1016/S0043-1354(03)00301-4

Alteri, C. J., & Mobley, H. L. T. (2012). *Escherichia coli* physiology and metabolism dictates adaptation to diverse host microenvironments. *Current Opinion in Microbiology*, *15*(1), 3–9. https://doi.org/10.1016/j.mib.2011.12.004

Arnold, B. F., Wade, T. J., Benjamin-Chung, J., Schiff, K. C., Griffith, J. F., Dufour, A. P., … Colford, J. M. (2016). Acute gastroenteritis and recreational water: Highest burden among young US children. *American Journal of Public Health*, *106*(9), 1690–1697. https://doi.org/10.2105/AJPH.2016.303279

Bernhard, A. E., & Field, K. G. (2000). A PCR assay To discriminate human and ruminant feces on the basis of host differences in Bacteroides-Prevotella genes encoding 16S rRNA. *Applied and Environmental Microbiology*, *66*(10), 4571–4574. https://doi.org/10.1128/aem.66.10.4571-4574.2000

Berthe, T., Ratajczak, M., Clermont, O., Denamur, E., & Petit, F. (2013). Evidence for coexistence of distinct *Escherichia coli* populations in various aquatic environments and their survival in estuary water. *Applied and Environmental Microbiology*, *79*(15), 4684–4693. https://doi.org/10.1128/AEM.00698-13

Beversdorf, L. J., Bornstein-Forst, S. M., & McLellan, S. L. (2007). The potential for beach sand to serve as a reservoir for *Escherichia coli* and the physical influences on cell die-off. *Journal of Applied Microbiology*, *102*(5), 1372–1381. https://doi.org/10.1111/j.1365-2672.2006.03177.x

Blount, Z. D. (2015). The unexhausted potential of *E. coli*. *ELife*, *4*, e05826. https://doi.org/10.7554/eLife.05826

Blyton, M. D. J., Pi, H., Vangchhia, B., Abraham, S., Trott, D. J., Johnson, J. R., & Gordon, D. M. (2015). Genetic Structure and Antimicrobial Resistance of *Escherichia coli* and Cryptic Clades in Birds with Diverse Human Associations. *Applied and Environmental Microbiology*, *81*(15), 5123–5133. https://doi.org/10.1128/AEM.00861-15

Brennan, F. P., Abram, F., Chinalia, F. A., Richards, K. G., & O'Flaherty, V. (2010). Characterization of environmentally persistent *Escherichia coli* isolates leached from an irish soilv. *Applied and Environmental Microbiology*, *76*(7), 2175–2180. https://doi.org/10.1128/AEM.01944-09

Brown, K. I., & Boehm, A. B. (2016). Transport of Fecal Indicators from Beach Sand to the Surf Zone by Recirculating Seawater: Laboratory Experiments and Numerical Modeling. *Environmental Science & Technology*, *50*(23), 12840–12847.

https://doi.org/10.1021/acs.est.6b02534
Byappanahalli, M. N., Whitman, R. L., Shively, D. A., Ting, W. T. E., Tseng, C. C., & 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*, *4*(3), 313–320. https://doi.org/10.2166/wh.2006.518
Byappanahalli, Muruleedhara N., Whitman, R. L., Shively, D. A., Sadowsky, M. J., & Ishii, S. (2006). Population structure, persistence, and seasonality of autochthonous *Escherichia coli* in temperate, coastal forest soil from a Great Lakes watershed. *Environmental Microbiology*, *8*(3), 504–513. https://doi.org/10.1111/j.1462-2920.2005.00916.x
Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869
Carlos, C., Pires, M. M., Stoppe, N. C., Hachich, E. M., Sato, M. I. Z., Gomes, T. A. T., … Ottoboni, L. M. M. (2010). *Escherichia coli* phylogenetic group determination and its application in the identification of the major animal source of fecal contamination. *BMC Microbiology*, *10*, 161. https://doi.org/10.1186/1471-2180-10-161
Chang, D.-E., Smalley, D. J., Tucker, D. L., Leatham, M. P., Norris, W. E., Stevenson, S. J., … Conway, T. (2004). Carbon nutrition of *Escherichia coli* in the mouse intestine. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(19), 7427–7432. https://doi.org/10.1073/pnas.0307888101
Chaudhuri, R. R., & Henderson, I. R. (2012). The evolution of the *Escherichia coli* phylogeny. *Infection, Genetics and Evolution*, *12*(2), 214–226. https://doi.org/10.1016/J.MEEGID.2012.01.005
Chiang, S. M., Dong, T., Edge, T. A., & Schellhorn, H. E. (2011). Phenotypic diversity caused by differential RpoS activity among environmental *Escherichia coli* isolates. *Applied and Environmental Microbiology*, *77*(22), 7915–7923. https://doi.org/10.1128/AEM.05274-11
Chistoserdova, L. (2015, July 26). Methylotrophs in natural habitats: current insights through metagenomics. *Applied Microbiology and Biotechnology*. Springer Verlag. https://doi.org/10.1007/s00253-015-6713-z
Clermont, O., Christenson, J. K., Denamur, E., & Gordon, D. M. (2013). The Clermont *Escherichia coli* phylo-typing method revisited: Improvement of specificity and detection of new phylo-groups. *Environmental Microbiology Reports*, *5*(1), 58–65. https://doi.org/10.1111/1758-2229.12019
Clermont, O., Olier, M., Hoede, C., Diancourt, L., Brisse, S., Keroudean, M., … Denamur, E. (2011). Animal and human pathogenic *Escherichia coli* strains share common genetic backgrounds. *Infection, Genetics and Evolution*, *11*(3), 654–662. https://doi.org/10.1016/j.meegid.2011.02.005
Cloutier, D.D., & McLellan, S. L. (2017). Distribution and differential survival of traditional and alternative indicators of fecal pollution at freshwater beaches. *Applied and Environmental Microbiology*, *83*(4). https://doi.org/10.1128/AEM.02881-16
Cloutier, Danielle D, Alm, E. W., & McLellan, S. L. (2015). Influence of Land Use, Nutrients, and Geography on Microbial Communities and Fecal Indicator Abundance at Lake Michigan Beaches. *Applied and Environmental Microbiology*, *81*(15), 4904–4913. https://doi.org/10.1128/AEM.00233-15
Converse, R. R., Kinzelman, J. L., Sams, E. A., Hudgens, E., Dufour, A. P., Ryu, H., … Wade, T. J. (2012). Dramatic Improvements in Beach Water Quality Following Gull Removal.

Environmental Science & Technology, 120905140233000. https://doi.org/10.1021/es302306b

Dada, A. C., & Hamilton, D. P. (2016). Predictive Models for Determination of *E. coli* Concentrations at Inland Recreational Beaches. *Water, Air, & Soil Pollution*. https://doi.org/10.1007/s11270-016-3033-6

Delmont, T. O., & Eren, A. M. (2018). Linking pangenomes and metagenomes: the *Prochlorococcus* metapangenome. *PeerJ*, *6*, e4320. https://doi.org/10.7717/peerj.4320

Dufour, A. P., & Ballentine, R. K. (1986). *Ambient Water Quality Criteria for Bacteria*. *Environmental Protective Agency*. Washington DC. Retrieved from https://www.epa.gov/sites/production/files/2019-03/documents/ambient-wqc-bacteria-1986.pdf

EPA, O. of R. & D. U. (2011). New Medium For The Enumeration And Subculture Of Bacteria From Potable Water.

Eren, A. M., Esen, Ö. C., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L., & Delmont, T. O. (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*. https://doi.org/10.7717/peerj.1319

Eshoo, M. W. (1988). *lac Fusion Analysis of the bet Genes of Escherichia coli: Regulation by Osmolarity, Temperature, Oxygen, Choline, and Glycine Betaine*. *JOURNAL OF BACTERIOLOGY* (Vol. 170). Retrieved from http://jb.asm.org/

Fang, X., Monk, J. M., Mih, N., Du, B., Sastry, A. V., Kavvas, E., … Palsson, B. O. (2018). *Escherichia coli* B2 strains prevalent in inflammatory bowel disease patients have distinct metabolic capabilities that enable colonization of intestinal mucosa. *BMC Systems Biology*, *12*(1), 1–10. https://doi.org/10.1186/s12918-018-0587-5

Farrell, M. J., & Finkel, S. E. (2003). The Growth Advantage in Stationary-Phase Phenotype Conferred by rpoS Mutations Is Dependent on the pH and Nutrient Environment. *JOURNAL OF BACTERIOLOGY*, *185*(24), 7044–7052. https://doi.org/10.1128/JB.185.24.7044-7052.2003

Feng, P., Lum, R., Chang2, G. W., & Hartman, A. (1991). Identification of uidA Gene Sequences in 3-D-Glucuronidase-Negative *Escherichia coli*. *APPLIED AND ENVIRONMENTAL MICROBIOLOGY*, *57*(1), 320–323.

Feng, S., Bootsma, M., & McLellan, S. L. (2018). Human-Associated *Lachnospiraceae* Genetic Markers Improve Detection of Fecal Pollution Sources in Urban Waters. *Applied and Environmental Microbiology*, *84*(14). https://doi.org/10.1128/AEM.00309-18

Fogarty, L. R., Haack, S. K., Wolcott, M. J., & Whitman, R. L. (2003). Abundance and characteristics of the recreational water quality indicator bacteria *Escherichia coli* and enterococci in gull faeces. *Journal of Applied Microbiology*, *94*(5), 865–878. https://doi.org/10.1046/j.1365-2672.2003.01910.x

Fujioka, R. S., Solo-Gabriele, H. M., Byappanahalli, M. N., & Kirs, M. (2015). U.S. recreational water quality criteria: A vision for the future. *International Journal of Environmental Research and Public Health*, *12*(7), 7752–7776. https://doi.org/10.3390/ijerph120707752

Getrich, M., McGee, C., Madison, M., Ferguson, D., Goodwin, K. D., Lee, C. M., … Cao, Y. (2009). Faecal indicator bacteria enumeration in beach sand: a comparison study of extraction methods in medium to coarse sands. *Journal of Applied Microbiology*, *107*(5), 1740–1750. https://doi.org/10.1111/j.1365-2672.2009.04440.x

Goodwin, K. D., Gruber, S., Vondrak, M., & Crumpacker, A. (2016). Watershed Assessment with Beach Microbial Source Tracking and Outcomes of Resulting Gull Management.

Environmental Science & Technology, *50*(18), 9900–9906. https://doi.org/10.1021/acs.est.6b02564

Goodwin, K. D., Matragrano, L., Wanless, D., Sinigalliano, C. D., & LaGier, M. J. (2009). A Preliminary Investigation Of Fecal Indicator Bacteria, Human Pathogens, And Source Tracking Markers In Beach Water And Sand, *2*(4).

Gordienko, E. N., Kazanov, M. D., & Gelfand, M. S. (2013). Evolution of pan-genomes of *Escherichia coli*, Shigella spp., and Salmonella enterica. *Journal of Bacteriology*, *195*(12), 2786–2792. https://doi.org/10.1128/JB.02285-12

Gordon, D. M. (2003). The Influence of Ecological Factors on the Distribution and the Genetic Structure of *Escherichia coli*. https://doi.org/10.1128/ecosalplus.6.4.1

Gordon, D. M., Clermont, O., Tolley, H., & Denamur, E. (2008). Assigning *Escherichia coli* strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method. *Environmental Microbiology*, *10*(10), 2484–2496. https://doi.org/10.1111/j.1462-2920.2008.01669.x

Gould, D. J., & Fletcher, M. R. (1978). Gull droppings and their effects on water quality. *Water Research*, *12*(9), 665–672. https://doi.org/10.1016/0043-1354(78)90176-8

Graciaa, D. S., Cope, J. R., Roberts, V. A., Cikesh, B. L., Kahler, A. M., Vigar, M., … Hlavsa, M. C. (2018, August 1). Outbreaks Associated with Untreated Recreational Water — United States, 2000–2014. *American Journal of Transplantation*. Blackwell Publishing Ltd. https://doi.org/10.1111/ajt.15002

Halliday, E., McLellan, S. L., Amaral-Zettler, L. A., Sogin, M. L., & Gast, R. J. (2014). Comparison of bacterial communities in sands and water at beaches with bacterial water quality violations. *PLoS ONE*, *9*(3). https://doi.org/10.1371/journal.pone.0090815

Hartz, A., Cuvelier, M., Nowosielski, K., Bonilla, T. D., Green, M., Esiobu, N., … Rogerson, A. (2008). Survival Potential of *Escherichia coli* and Enterococci in Subtropical Beach Sand: Implications for Water Quality Managers. *Journal of Environmental Quality*, *37*(3), 898–905. https://doi.org/10.2134/jeq2007.0312

Herzer, P. J., Inouye, S., Inouye, M., & Whittam, T. S. (1990). Phylogenetic Distribution of Branched RNA-Linked Multicopy Single-Stranded DNA among Natural Isolates of *Escherichia coli*. *Journal of Bacteriology*, *127*(11), 6175–6181. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC526797/pdf/jbacter00165-0011.pdf

Howard, K. J., Martin, E., Gentry, T., Feagley, S., & Karthikeyan, R. (2017). Effects of Dairy Manure Management Practices on *E. coli* Concentration and Diversity. *Water, Air, & Soil Pollution*, *228*(1), 4. https://doi.org/10.1007/s11270-016-3182-7

Ingle, D., Clermont, O., Skurnik, D., Denamur, E., Walk, S. T., & Gordon, D. M. (2011). Biofilm Formation by and Thermal Niche and Virulence Characteristics of Escherichia spp. *Applied and Environmental Microbiology*, *77*(8), 2695–2700. Retrieved from https://aem.asm.org/content/aem/77/8/2695.full.pdf

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. https://doi.org/10.1021/ES0623156

Ishii, S., Yan, T., Vu, H., Hansen, D. L., Hicks, R. E., & Sadowsky, M. J. (2010). Factors Controlling Long-Term Survival and Growth of Naturalized *Escherichia coli* Populations in Temperate Field Soils. *Microbes and Environments*, *25*(1), 8–14. https://doi.org/10.1264/jsme2.ME09172

Jang, J., Hur, H. G., Sadowsky, M. J., Byappanahalli, M. N., Yan, T., & Ishii, S. (2017,

September 1). Environmental *Escherichia coli*: ecology and public health implications—a review. *Journal of Applied Microbiology*. Blackwell Publishing Ltd. https://doi.org/10.1111/jam.13468

Jaureguy, F., Landraud, L., Passet, V., Diancourt, L., Frapy, E., Guigon, G., … Brisse, S. (2008). Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains. *BMC Genomics*, *9*(1), 560. https://doi.org/10.1186/1471-2164-9-560

Johnson, J. R., Gordon, D. M., & Bauer, S. (2002). The genetic structure of *Escherichia coli* populations in primary and secondary habitats. *Microbiology*, *148*(5), 1513–1522. https://doi.org/10.1099/00221287-148-5-1513

Kaas, R. S., Friis, C., Ussery, D. W., & Aarestrup, F. M. (2012). Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes. *BMC Genomics*, *13*(1), 577. https://doi.org/10.1186/1471-2164-13-577

Kabler, P. W., Clark, H. F., & Geldreich, E. E. (1963). Sanitary Significance Of Coliform And Fecal Coliform Organisms In Surface Water.

Kildare, B. J., Leutenegger, C. M., McSwain, B. S., Bambic, D. G., Rajal, V. B., & Wuertz, S. (2007). 16S rRNA-based assays for quantitative detection of universal, human-, cow-, and dog-specific fecal Bacteroidales: A Bayesian approach. *Water Research*, *41*(16), 3701–3715. https://doi.org/10.1016/j.watres.2007.06.037

Kon, T., Weir, S. C., Trevors, J. T., Lee, H., Champagne, J., Meunier, L., … Masson, L. (2007). Microarray Analysis of *Escherichia coli* Strains from Interstitial Beach Waters of Lake Huron (Canada). *APPLIED AND ENVIRONMENTAL MICROBIOLOGY*, *73*(23), 7757–7758. https://doi.org/10.1128/AEM.01333-07

Korajkic, A., Brownell, M. J., & Harwood, V. J. (2011). Investigation of human sewage pollution and pathogen analysis at Florida Gulf coast Beaches. *Journal of Applied Microbiology*, *110*(1), 174–183. https://doi.org/10.1111/j.1365-2672.2010.04869.x

Köster, J., & Rahmann, S. (2012). Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Kurihara, S., Oda, S., Kato, K., Kim, H. G., Koyanagi, T., Kumagai, H., & Suzuki, H. (2005). A novel putrescine utilization pathway involves γ-glutamylated intermediates of *Escherichia coli* K-12. *Journal of Biological Chemistry*, *280*(6), 4602–4608. https://doi.org/10.1074/jbc.M411114200

Kurihara, S., Oda, S., Tsuboi, Y., Hyeon, G. K., Oshida, M., Kumagai, H., & Suzuki, H. (2008). γ-glutamylputrescine synthetase in the putrescine utilization pathway of *Escherichia coli* K-12. *Journal of Biological Chemistry*, *283*(29), 19981–19990. https://doi.org/10.1074/jbc.M800133200

Landfald, B., & Strom, A. R. (1986). Choline-Glycine Betaine Pathway Confers a High Level of Osmotic Tolerance in *Escherichia coli*. *JOURNAL OF BACTERIOLOGY* (Vol. 165). Retrieved from http://jb.asm.org/

Lapierre, P., & Gogarten, J. P. (2009, March 1). Estimating the size of the bacterial pan-genome. *Trends in Genetics*. Elsevier Current Trends. https://doi.org/10.1016/j.tig.2008.12.004

Lecointre, G., Rachdi, L., Darlu, P., & Denamur, E. (1998). *Escherichia coli* molecular phylogeny using the incongruence length difference test. *Molecular Biology and Evolution*, *15*(12), 1685–1695. https://doi.org/10.1093/oxfordjournals.molbev.a025895

Lee, C. M., Lin, T. Y., Lin, C.-C., Kohbodi, G. A., Bhatt, A., Lee, R., & Jay, J. A. (2006). Persistence of fecal indicator bacteria in Santa Monica Bay beach sediments. *Water Research*, *40*(14), 2593–2602. https://doi.org/10.1016/j.watres.2006.04.032

Lu, J., Santo Domingo, J. W., Lamendella, R., Edge, T., & Hill, S. (2008). Phylogenetic diversity and molecular detection of bacteria in gull feces. *Applied and Environmental Microbiology*, *74*(13), 3969–3976. https://doi.org/10.1128/AEM.00019-08

Lukjancenko, O., Wassenaar, T. M., & Ussery, D. W. (2010). Comparison of 61 sequenced *Escherichia coli* genomes. *Microbial Ecology*, *60*(4), 708–720. https://doi.org/10.1007/s00248-010-9717-3

Luo, C., Walk, S. T., Gordon, D. M., Feldgarden, M., Tiedje, J. M., & Konstantinidis, K. T. (2011). Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *Proceedings of the National Academy of Sciences*, *108*(17), 7200–7205. https://doi.org/10.1073/pnas.1015622108

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10. https://doi.org/10.14806/ej.17.1.200

Martins, M. T., Rivera, I. G., Clark, D. L., Stewart, M. H., Wolfe, R. L., & Olson2, A. B. H. (1993). Distribution of uidA Gene Sequences in *Escherichia coli* Isolates in Water Sources and Comparison with the Expression of ,B-Glucuronidase Activity in 4-Methylumbelliferyl-3-D-Glucuronide Media. *APPLIED AND ENVIRONMENTAL MICROBIOLOGY*, *59*(7), 2271–2276.

Martinson, J. N. V, Pinkham, N. V, Peters, G. W., Cho, H., Heng, J., Rauch, M., … Walk, S. T. (2019). Rethinking gut microbiome residency and the Enterobacteriaceae in healthy human adults. *The ISME Journal*, *13*(9), 2306–2318. https://doi.org/10.1038/s41396-019-0435-7

McMurdie, P. J., & Holmes, S. (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE*, *8*(4), e61217. https://doi.org/10.1371/journal.pone.0061217

Medini, D., Donati, C., Tettelin, H., Masignani, V., & Rappuoli, R. (2005). The microbial pan-genome. *Current Opinion in Genetics & Development*, *15*(6), 589–594. https://doi.org/10.1016/J.GDE.2005.09.006

Mohiuddin, M. M., Salama, Y., Schellhorn, H. E., & Golding, G. B. (2017). Shotgun metagenomic sequencing reveals freshwater beach sands as reservoir of bacterial pathogens. *Water Research*, *115*, 360–369. https://doi.org/10.1016/j.watres.2017.02.057

Moissenet, D., Salauze, B., Clermont, O., Bingen, E., Arlet, G., Denamur, E., … Vu-Thien, H. (2010). Meningitis Caused by *Escherichia coli* Producing TEM-52 Extended-Spectrum -Lactamase within an Extensive Outbreak in a Neonatal Ward: Epidemiological Investigation and Characterization of the Strain. *Journal of Clinical Microbiology*, *48*(7), 2459–2463. https://doi.org/10.1128/JCM.00529-10

Monk, J. M., Charusanti, P., Aziz, R. K., Lerman, J. A., Premyodhin, N., Orth, J. D., … Palsson, B. Ø. (2013). Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(50), 20338–20343. https://doi.org/10.1073/pnas.1307797110

Nandakafle, G., Seale, T., Flint, T., Nepal, M., Venter, S. N., & Brozel, V. S. (2017). Distribution of Diverse *Escherichia coli* between Cattle and Pasture. *Microbes and Environments*, *32*(3), 226–233. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5606692/pdf/32_226.pdf

Nelson, M. C., Morrison, H. G., Benjamino, J., Grim, S. L., & Graf, J. (2014). Analysis, Optimization and Verification of Illumina-Generated 16S rRNA Gene Amplicon Surveys. *PLoS ONE*, *9*(4), 94249. https://doi.org/10.1371/journal.pone.0094249

Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D., & Bertilsson, S. (2011). A Guide to the Natural History of Freshwater Lake Bacteria. *Microbiology and Molecular Biology Reviews*, *75*(1), 14–49. https://doi.org/10.1128/mmbr.00028-10

Noble, R. T., Weisberg, S. B., Leecaster, M. K., Mcgee, C. D., Ritter, K., Walker, K. O., & Vainik, P. M. (2003). Comparison Of Beach Bacterial Water Quality Indicator Measurement Methods.

Ochman, H., & Selander, R. K. (1984). Standard reference strains of *Escherichia coli* from natural populations. *Journal of Bacteriology*, *157*(2), 690–693. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6363394

Oh, S., Buddenborg, S., Yoder-Himes, D. R., Tiedje, J. M., & Konstantinidis, K. T. (2012). Genomic Diversity of Escherichia Isolates from Diverse Habitats. *PLoS ONE*, *7*(10), e47005. https://doi.org/10.1371/journal.pone.0047005

Palmer, J. A., Law, J. Y., & Soupir, M. L. (2020). Spatial and temporal distribution of *E. coli* contamination on three inland lake and recreational beach systems in the upper Midwestern United States. *Science of the Total Environment*, *722*, 137846. https://doi.org/10.1016/j.scitotenv.2020.137846

Praveena, S. M., Shamira, S. S., Ismail, S. N. S., & Aris, A. Z. (2016). Fecal indicator bacteria in tropical beach sand: Baseline findings from Port Dickson coastline, Strait of Malacca (Malaysia). *Marine Pollution Bulletin*, *110*(1), 609–612. https://doi.org/10.1016/j.marpolbul.2016.06.024

R Core Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.r-project.org/

Rasko, D. A., Rosovitz, M. J., Myers, G. S. A., Mongodin, E. F., Fricke, W. F., Gajer, P., … Ravel, J. (2008). The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *Journal of Bacteriology*, *190*(20), 6881–6893. https://doi.org/10.1128/JB.00619-08

Ruban, V., López-Sánchez, J. F., Pardo, P., Rauret, G., Muntauc, H., & Quevauviller, P. (1998). Selection and evaluation of sequential extraction procedures for the determination of phosphorus forms in lake sediment. Retrieved from http://pubs.rsc.org/en/content/articlepdf/1999/em/a807778i

Rumball, N. A., Mayer, H. R. C., & McLellan, S. L. (2021). Selective Survival of *Escherichia coli* Phylotypes in Freshwater Beach Sand. *Applied and Environmental Microbiology*, *87*(4). https://doi.org/10.1128/AEM.02473-20

Safaie, A., Weiskerger, C. J., Nevers, M. B., Byappanahalli, M. N., & Phanikumar, M. S. (2021). Evaluating the impacts of foreshore sand and birds on microbiological contamination at a freshwater beach. *Water Research*, *190*, 116671. https://doi.org/10.1016/j.watres.2020.116671

Savageau, M. A. (1983). *Escherichia coli* habitats, cell types, and molecular mechanisms of gene control. *American Naturalist*, *122*(6), 732–744. https://doi.org/10.1086/284168

Selander, R., Caugant, D., & Whittam, T. (1987). Genetic structure and variation in natural populations of *Escherichia coli*. In F. C. Neidhardt (Ed.), *Escherichia coli and Salmonella typhimurium : cellular and molecular biology* (pp. 1625–1648). Washington DC: American Society for Microbiology. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Genetic+stucture+and+variation+in+natural+populations+of+Escherichia+coli#0

Shaw, L. P., Chau, K. K., Kavanagh, J., AbuOun, M., Stubberfield, E., Gweon, H. S., …

Stoesser, N. (2021). Niche and local geography shape the pangenome of wastewater- and livestock-associated Enterobacteriaceae. *Science Advances*, *7*(15), eabe3868. https://doi.org/10.1126/sciadv.abe3868

Shuval Hillel Shuval Fredy, H. (2003). Estimating the global burden of thalassogenic diseases: human infectious diseases caused by wastewater pollution of the marine environment.

Solo-Gabriele, H. M., Harwood, V. J., Kay, D., Fujioka, R. S., Sadowsky, M. J., Whitman, R. L., … Brandão, J. C. (2016). Beach sand and the potential for infectious disease transmission: observations and recommendations. *Journal of the Marine Biological Association of the United Kingdom*, *96*(01), 101–120. https://doi.org/10.1017/S0025315415000843

Solo-Gabriele, H. M., Wolfert, M. A., Desmarais, T. R., & Palmer, C. J. (2000). Sources of *Escherichia coli* in a coastal subtropical environment. *Applied and Environmental Microbiology*, *66*(1), 230–237. https://doi.org/10.1128/AEM.66.1.230-237.2000

Sprouffske, K., & Wagner, A. (2016). Growthcurver: an R package for obtaining interpretable metrics from microbial growth curves. *BMC Bioinformatics*, *17*(1), 172. https://doi.org/10.1186/s12859-016-1016-7

Staley, Z. R., & Edge, T. A. (2016). Comparative microbial source tracking methods for identification of fecal contamination sources at Sunnyside Beach in the Toronto region area of concern. *Journal of Water and Health*, *14*(5), 839–850. https://doi.org/10.2166/wh.2016.296

Staley, Z. R., Vogel, L., Robinson, C., & Edge, T. A. (2015). Differential occurrence of *Escherichia coli* and human Bacteroidales at two Great Lakes beaches. *Journal of Great Lakes Research*, *41*(2), 530–535. https://doi.org/10.1016/J.JGLR.2015.03.006

Stecher, G., Tamura, K., & Kumar, S. (2020). Molecular evolutionary genetics analysis (MEGA) for macOS. *Molecular Biology and Evolution*, *37*(4), 1237–1239. https://doi.org/10.1093/molbev/msz312

Stoppe, N. de C., Silva, J. S., Carlos, C., Sato, M. I. Z., Saraiva, A. M., Ottoboni, L. M. M., & Torres, T. T. (2017). Worldwide Phylogenetic Group Patterns of *Escherichia coli* from Commensal Human and Wastewater Treatment Plant Isolates. *Frontiers in Microbiology*, *8*, 2512. https://doi.org/10.3389/fmicb.2017.02512

Strayer, D. L., & Findlay, S. E. G. (2010). Ecology of freshwater shore zones. *Aquatic Sciences*, *72*(2), 127–163. https://doi.org/10.1007/s00027-010-0128-9

Tabor, C. W., & Tabor, H. (1984). Polyamines. *Annual Review of Biochemistry*, *53*(1), 749–790. https://doi.org/10.1146/annurev.bi.53.070184.003533

Takahashi, H., Takayanagi, Y., Fujita, N., Tanaka, K., & Ishihama, A. (2006). Heterogeneity of the principal sigma factor in *Escherichia coli*: the rpoS gene product, sigma 38, is a second principal sigma factor of RNA polymerase in stationary-phase *Escherichia coli*. *Proceedings of the National Academy of Sciences*, *90*(8), 3511–3515. https://doi.org/10.1073/pnas.90.8.3511

Tenaillon, O., Skurnik, D., Picard, B., & Denamur, E. (2010). The population genetics of commensal *Escherichia coli*. *Nature Reviews Microbiology*, *8*. https://doi.org/10.1038/nrmicro2298

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., … Denamur, E. (2009). Organised Genome Dynamics in the *Escherichia coli* Species Results in Highly Diverse Adaptive Paths. *PLoS Genetics*, *5*(1), e1000344. https://doi.org/10.1371/journal.pgen.1000344

Tymensen, L. D., Pyrdok, F., Coles, D., Koning, W., McAllister, T. A., Jokinen, C. C., …

Neumann, N. F. (2015). Comparative accessory gene fingerprinting of surface water *Escherichia coli* reveals genetically diverse naturalized population. *Journal of Applied Microbiology*, *119*(1), 263–277. https://doi.org/10.1111/jam.12814

U.S. Environmental Protection Agency. (2000). Improved Enumeration Methods for the Recreational Water Quality Indicators: Enterococci and *Escherichia coli*. Retrieved from https://www.epa.gov/sites/production/files/2018-12/documents/improved-enumeration-methods-enterococci.pdf

US EPA, O. (n.d.). About the BEACH Act.

Vogel, L. J., O'Carroll, D. M., Edge, T. A., & Robinson, C. E. (2016). Release of *Escherichia coli* from Foreshore Sand and Pore Water during Intensified Wave Conditions at a Recreational Beach. *Environmental Science and Technology*, *50*(11). https://doi.org/10.1021/acs.est.6b00707

Walk, S. T., Alm, E. W., Calhoun, L. M., Mladonicky, J. M., & Whittam, T. S. (2007). Genetic diversity and population structure of *Escherichia coli* isolated from freshwater beaches. *Environmental Microbiology*, *9*(9), 2274–2288. https://doi.org/10.1111/j.1462-2920.2007.01341.x

Walk, S. T., Alm, E. W., Gordon, D. M., Ram, J. L., Toranzos, G. A., Tiedje, J. M., & Whittam, T. S. (2009). Cryptic lineages of the genus Escherichia. *Applied and Environmental Microbiology*, *75*(20), 6534–6544. https://doi.org/10.1128/AEM.01262-09

Wanjugi, P., Fox, G. A., & Harwood, V. J. (2016). The Interplay Between Predation, Competition, and Nutrient Levels Influences the Survival of *Escherichia coli* in Aquatic Environments. *Microbial Ecology*, *72*(3), 526–537. https://doi.org/10.1007/s00248-016-0825-6

White, A. P., Sibley, K. A., Sibley, C. D., Wasmuth, J. D., Schaefer, R., Surette, M. G., … Neumann, N. F. (2011). Intergenic sequence comparison of *Escherichia coli* isolates reveals lifestyle adaptations but not host specificity. *Applied and Environmental Microbiology*. https://doi.org/10.1128/AEM.05909-11

Whitman, R., Harwood, V. J., Edge, T. A., Nevers, M., Byappanahalli, M., Vijayavel, K., … Solo-Gabriele, H. M. (2014). Microbes in Beach Sands: Integrating Environment, Ecology and Public Health. *Re/Views in Environmental Science and Bio/Technology*, *13*(3), 329–368. https://doi.org/10.1007/s11157-014-9340-8

Whitman, R. L., & Nevers, M. B. (2003). Foreshore sand as a source of *Escherichia coli* in nearshore water of a Lake Michigan beach. *Applied and Environmental Microbiology*, *69*(9), 5555–5562. https://doi.org/10.1128/AEM.69.9.5555-5562.2003

Yamahara, K. M., Sassoubre, L. M., Goodwin, K. D., & Boehm, A. B. (2012). Occurrence and Persistence of Bacterial Pathogens and Indicator Organisms in Beach Sand along the California Coast. https://doi.org/10.1128/AEM.06185-11

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, *30*(5), 614–620. https://doi.org/10.1093/bioinformatics/btt593

Zhi, S., Banting, G., Li, Q., Edge, T. A., Topp, E., Sokurenko, M., & Scott, C. (2016). Evidence of Naturalized Stress-Tolerant Strains of *Escherichia coli* in, *82*(18), 5505–5518. https://doi.org/10.1128/AEM.00143-16.Editor

APPENDIX A.
Supplemental Material for Chapter 2

Supplemental Table 2.1. Exponential decay rates and temperatures throughout microcosm experiments[1].

Isolate Source	Experimental Conditions	Duration (days)	Exponential Decay Rate 4 weeks	Exponential Decay Rate Phylotype Experiment	R^2 Phylotype Experiment	Exponential Decay Rate Over-Winter	R^2 Over-winter	Max Temp (°C)	Min Temp (°C)	Mean Temp (°C)	Mean Temp Over-winter (°C)
Deployment 1: Source Microcosms											
Sand	Native sand	45	-0.198	-0.138	0.83	NA	NA	17.5	11.06	14.49	7.66
	Autoclaved sand	96	0.065	-0.007	0.021	-0.007	0.56	21.47	9.49	14.69	7.66
Sewage	Native sand	45	-0.145	-0.104	0.72	-0.042	0.9	17.5	11.06	14.49	7.66
	Autoclaved sand	96	0.098	0.009	0.1	-0.007	0.67	21.47	9.49	14.69	7.66
Gull	Native sand	45	-0.163	-0.122	0.82	-0.044	0.89	17.5	11.06	14.49	7.66
	Autoclaved sand	96	0.093	-0.014	0.07	-0.009	0.58	21.47	9.49	14.69	7.66
Deployment 2: Nutrient Microcosms											
Sand	Native sand	56	-0.076	-1.75	0.98	NA	NA	15.87	8.54	12.83	NA
	Autoclaved sand	56	0.207	2.26	0.55	NA	NA	15.87	8.54	12.83	NA
	Baked Sand	56	-0.018	-3.33	0.81	NA	NA	15.87	8.54	12.83	NA
	Baked Control - in lab	56	-0.245	-4.55	0.92	NA	NA	15.87	8.54	12.83	NA
	90% Native 10% Autoclaved	56	-0.077	-1.86	0.99	NA	NA	15.87	8.54	12.83	NA
	90% Baked 10% Autoclaved	56	0.159	1.93	0.64	NA	NA	15.87	8.54	12.83	NA

[1] Decay rates were calculated as the slope of a linear fit of the sampling timepoints, with x =time and y=ln(relative concentration), where the relative concentration is C/C0 (C= final concentration; C0 = initial concentration), with R^2 representing the linear fit to the data. The temperature was measured hourly with *in situ* temperature sensors.

Supplemental Table 2.2. Phylotype distribution of *E. coli* isloates from freshwater sand, sewage, and gull waste.

			***Escherichia coli* phylotypes**							
Source	**Sample Location and Date**	**No. of Isolates**	**A**	**A or C**	**B1**	**B2**	**D or E**	**F**	**Clade or E**	**Unknown**
Sand	Bradford and Atwater* Beaches 10/2/08 - 9/21/10	92	13.04	3.26	41.30	13.04	10.87	2.17	10.87	5.43
	Bradford Beach 7/16/18 - 7/22/19	93	9.68	4.30	20.43	8.60	25.81	16.13	10.75	4.30
	Bradford Beach 5/30/18	85	38.82	4.71	18.82	4.71	21.18	9.41	1.18	1.18
	Bradford Beach 7/11/2018	90	5.56	3.33	71.11	4.44	2.22	3.33	3.33	6.67
	Bradford and Atwater Beaches 10/14/16	92	1.09	2.17	27.17	4.35	53.26	3.26	2.17	6.52
	Atwater Beach 7/16/19 - 7/22/19	94	3.19	2.13	38.30	31.91	17.02	4.26	2.13	1.06
	Atwater Beach 7/11/18	93	4.30	4.30	56.99	16.13	12.90	3.23	2.15	0.00
	Atwater Beach 5/30/18	83	16.87	10.84	40.96	8.43	4.82	6.02	4.82	7.23

	Point Beach* 7/6/16	88	10.23	12.50	48.86	0.00	20.45	4.55	0.00	3.41
	Kohler Andre 7/6/16	69	13.04	13.04	60.87	1.45	8.70	1.45	0.00	1.45
	Sand Total	879	11.26	5.80	42.09	9.67	18.09	5.46	3.87	3.75
Gull Waste	Bradford Beach 5/25/02 - 5/21/02	93	16.13	30.11	9.68	13.98	5.38	0.00	20.43	4.30
	South Shore Beach* 5/15/02 - 7/22/02	91	1.10	7.69	36.26	18.68	27.47	3.30	2.20	3.30
	6th & Canal* 10/18/12	90	12.22	10.00	31.11	12.22	21.11	2.22	2.22	8.89
	6th & Canal 10/12/12	87	13.79	10.34	22.99	11.49	17.24	4.60	8.05	11.49
	Gull Total	361	10.80	14.68	24.93	14.13	17.73	2.49	8.31	6.93
Sewage	Sewage 2001* 43rd & Euclid	93	9.68	6.45	46.24	2.15	15.05	0.00	2.15	18.28
	Sewage 2002* 72nd & Center	93	2.15	2.15	43.01	8.60	18.28	1.08	4.30	20.43
	Sewage Total	186	6.41	5.13	53.21	6.41	19.87	0.64	3.85	4.49

*Isolates used in the microcsm expements

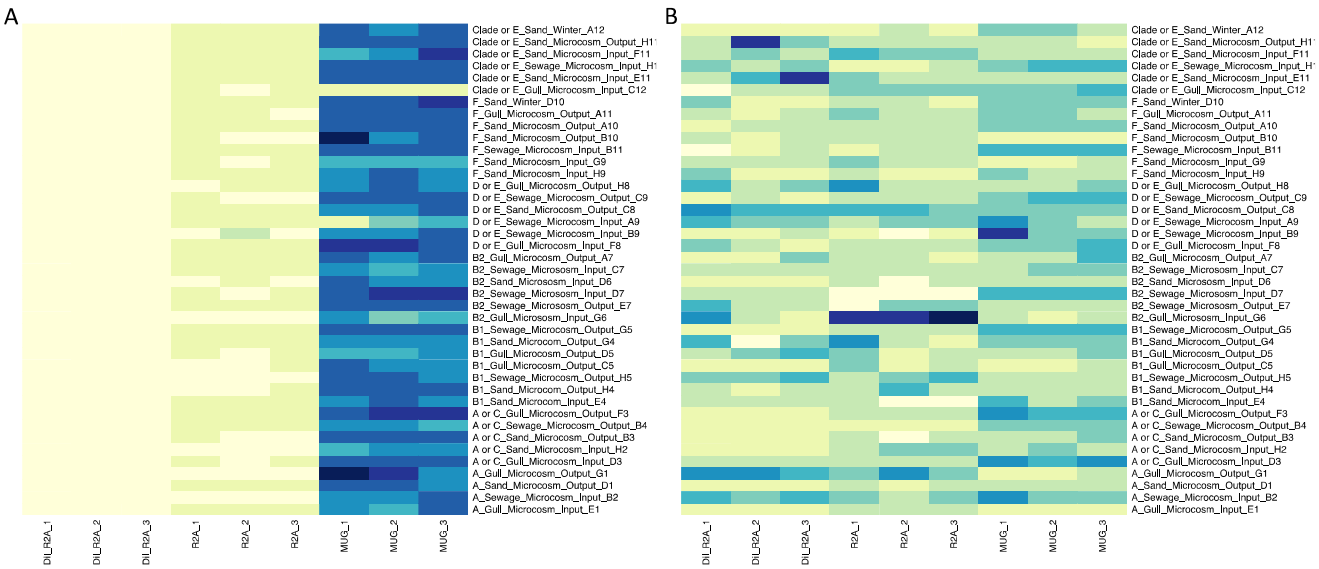
Supplemental Table 2.3. qPCR[1] and culturable *E. coli* results for human and gull markers in beach sand samples.

Sample Date	Site Code[2]	CFU/ 100 g Sand	Lachno3 Human *Lachnospiraceae* (CN/100 g)	HF183 Human *Bacteroides* (CN/100 g)	Gull2 *Catellicoccus* (CN/100 g)
6/21/16	PB CONC 1	0	0	0	0
	PB CONC 2	450	0	0	0
	PB CONC 3	450	0	0	0
	PB LTHS 1	1350	0	0	0
	PB LTHS 2	450	0	0	0
	PB LTHS 3	1350	0	0	0
	PB PICNIC 1	4050	0	0	0
	PB PICNIC 2	3150	0	0	163000
	PB PICNIC 3	450	0	0	0
6/21/16	KA NAC 1	0	0	0	0
	KA NAC 2	900	0	0	0
	KA NAC 3	5400	0	0	0
	KA NB 1	0	0	0	0
	KA NB 2	9900	0	0	0
	KA NB 3	0	0	0	50600
	KA SP 1	7200	0	0	0
	KA SP 2	3600	0	0	0
	KA SP 3	1350	0	0	0
5/30/18	BB 1	12150	0	0	158000
	BB 2	4050	0	0	0
	BB 3	72000	0	0	
5/30/18	ATW 1	NA	0	0	329000
	ATW 2	10350	0	0	0
	ATW 3	4950	0	0	161000
	ATW 4	14400	0	0	0
7/11/18	BB 1	3600	0	0	0
	BB 2	4500	0	0	0
	BB 3	68000	0	0	0
7/11/18	ATW 1	59000	0	0	75100
	ATW 2	141000	0	0	123000

	ATW 3	8550	0	0	0
	ATW 4	4050	0	0	0
7/16/19	BB 1	450	0	0	0
	BB 2	0	0	0	0
	BB 3	450	0	0	0
7/16/19	ATW 1	0	0	0	0
	ATW 2	1350	0	0	43900
	ATW 3	450	0	0	0
7/22/19	BB 1	4500	0	0	0
	BB 2	0	0	0	0
	BB 3	0	0	0	0
7/22/19	ATW 1	54000	0	0	0
	ATW 2	4010000	0	0	464000
	ATW 3	4500	0	0	159000

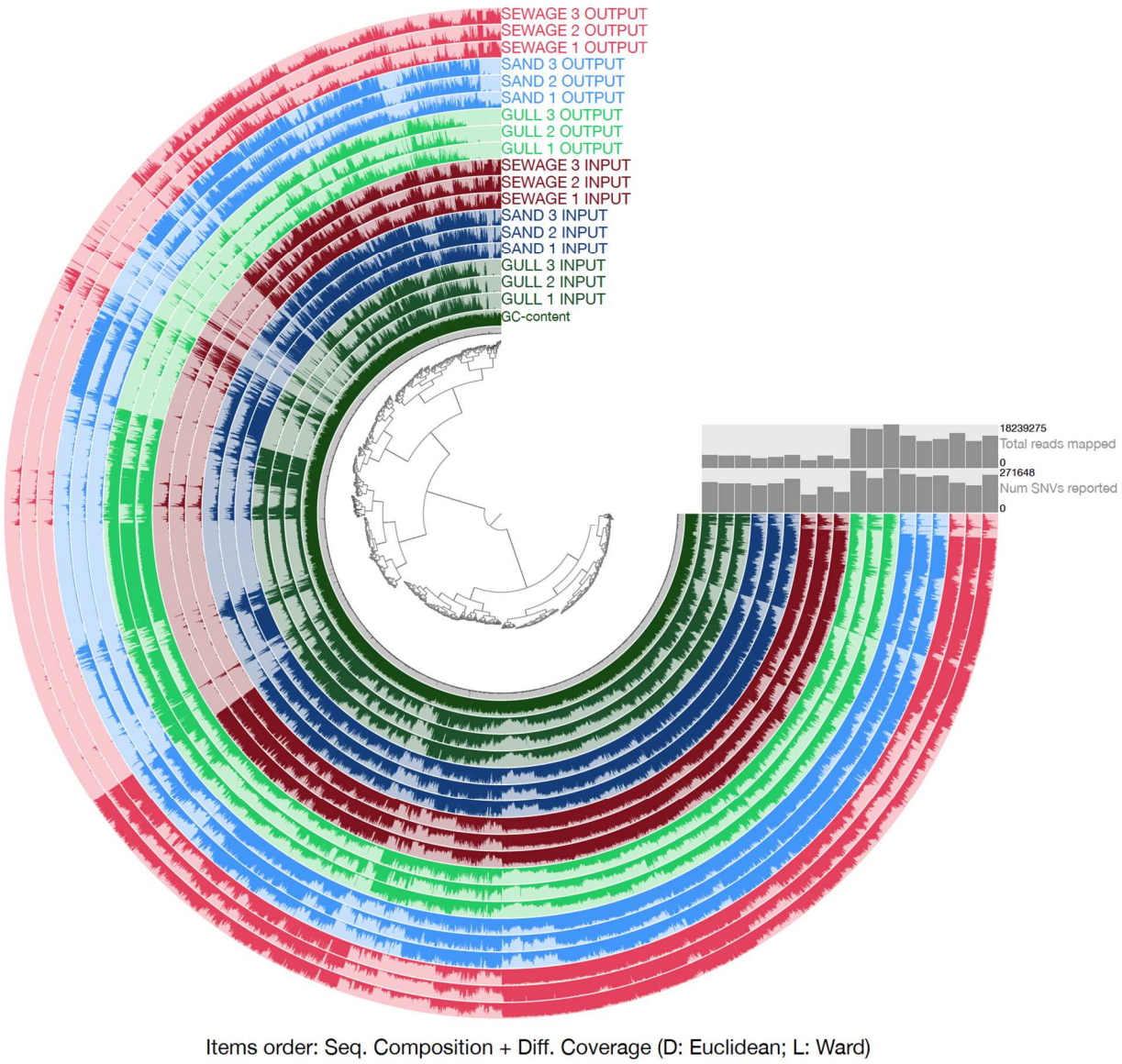
[1] CN concnetraions are the average of duplicate samples

[2]Each beach had multiple testing sites, indicated by the numeral following the site name



high values left) Carrying capacity, or the maximum cell density in growth experiment cultures. The more nutrient in the treatment, the higher the carrying capacity right) Growth rate constants, or the slope of the growth curve, the steeper the curve, the faster the growth, and higher the growth rate constant. No significant difference among phylogroups was observed (Repeated measures ANOVA, P = 0.05).

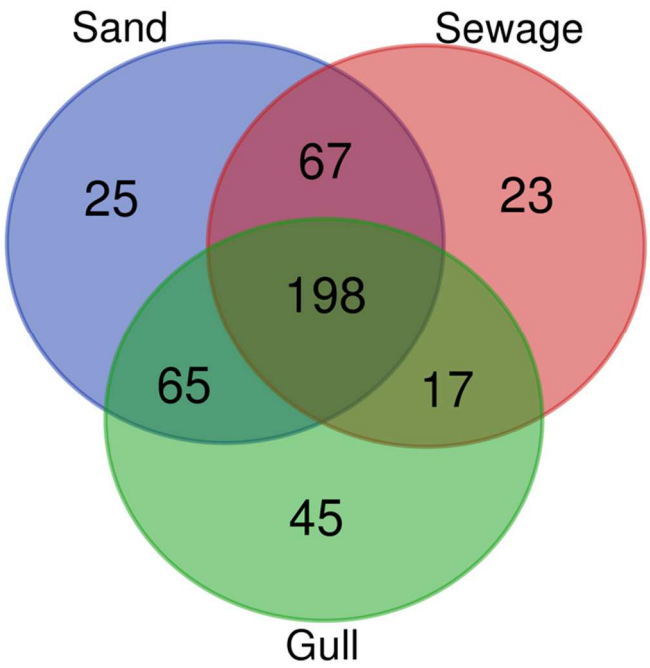
APPENDIX B.
Supplemental Material for Chapter 3



Supplemental Figure 3.1. Short reads from shotgun sequencing of microcosm isolate collections mapped to the accessory genomes of 21 isolates. Population level sequence data from inoculated and recovered isolate collections were mapped to the concatenated accessory genomes of 21 isolates. Each ring of the diagram represents one sample triplicate, with red, blue, and green shading representing the sample source as either from sewage, sand, or gull. Darker shades of these colors represent populations from the input of the experiment (pool of 188 *E. coli*), and lighter shades representing the surviving populations. The magnitude of the shading within each ring represents the depth of coverage that sample has for that specific gene.

Supplemental Table 3.1. Assessment of mapping efficiency.

Sample	Total number of reads	Percent reads mapped to total pangenome	Percent reads mapped to accessory pangenome
GULL_1_INPUT	2.07E+07	87.4	25.7
GULL_2_INPUT	1.92E+07	87.6	25.8
GULL_3_INPUT	1.94E+07	87.0	25.7
GULL_1_OUTPUT	6.80E+07	83.0	24.1
GULL_2_OUTPUT	6.67E+07	80.3	24.5
GULL_3_OUTPUT	7.41E+07	84.6	24.6
SAND_1_INPUT	1.85E+07	89.0	21.4
SAND_2_INPUT	1.99E+07	92.0	22.4
SAND_3_INPUT	2.40E+07	91.6	22.4
SAND_1_OUTPUT	6.36E+07	94.9	21.1
SAND_2_OUTPUT	5.25E+07	94.6	21.2
SAND_3_OUTPUT	5.63E+07	93.1	21.3
SEWAGE_1_INPUT	1.56E+07	88.1	20.4
SEWAGE_2_INPUT	2.39E+07	88.8	20.7
SEWAGE_3_INPUT	1.96E+07	88.4	19.1
SEWAGE_1_OUTPUT	6.86E+07	93.9	21.0
SEWAGE_2_OUTPUT	5.28E+07	94.8	21.0
SEWAGE_3_OUTPUT	6.40E+07	93.6	21.0



Supplemental Figure 3.2. The source distribution of functions enriched >25% in the microcosm output. This analysis included only those genes that were able to be annotated with KEGG, with the majority of enriched functions shared by all sources assessed.

Supplemental Table 3.2. KO Functions that wree enriched in all output sources and previsoul;y consider environmentally related based on cryptic clade genomic analysis

KO #	Protein Family	Family Category	KO Function
K02529	Genetic information processing	Transcription factors	lacI, galR; LacI family transcriptional regulator
K00130	Metabolism	Enzymes	betB, gbsA; betaine-aldehyde dehydrogenase
K13243	Metabolism	Enzymes	dos; c-di-GMP-specific phosphodiesterase
K00048	Metabolism	Enzymes	fucO; lactaldehyde reductase
K09471	Metabolism	Enzymes	puuB, ordL; gamma-glutamylputrescine oxidase
K03615	Metabolism	Enzymes	rnfC; Na+-translocating ferredoxin:NAD+ oxidoreductase subunit C
K12985	Metabolism	Enzymes	waaW; (galactosyl)LPS 1,2-glucosyltransferase
K16703	Metabolism	Enzymes	wcaL, amsK; colanic acid/amylovoran biosynthesis glycosyltransferase
K13979	Metabolism	Enzymes	yahK; alcohol dehydrogenase (NADP+)
K11911	Signaling and cellular processes	Secretion system	vasL; type VI secretion system protein VasL
K02511	Signaling and cellular processes	Transport	hpaX; MFS transporter, ACS family, 4-hydroxyphenylacetate permease
K00427	Signaling and cellular processes	Transport	lldP, lctP; L-lactate permease
K08173	Signaling and cellular processes	Transport	ydfJ; MFS transporter, MHS family, metabolite:H+ symporter

Supplemental Table 3.3. Description of genomes used in the pangenome analysis.

Genome Name	Source	phylo type	total length	total # genes	# core genes	# accessory genes[1]	# island genes
DC5_B12	freshwater beach sand	A	4626235	4390	3054	1336	68
DC5_C10*	freshwater beach sand	A	4764872	4415	3057	1358	63
DC5_C4*	freshwater beach sand	B1	5076009	4863	3067	1796	168
DC5_E12*	freshwater beach sand	B1	6453416	6427	4037	2390	165
Gull1_B2*	gull waste	clade or E	4758314	4451	3048	1403	63
Gull1_E8	gull waste	B1	5217237	5054	3044	2010	139
Gull1_H5	gull waste	A	5293412	5165	3063	2102	268
Gull2_G9	gull waste	B1	4922839	4722	3046	1676	131
MII_Sand1_1_C4	freshwater beach sand - microcosm output	A	5054306	4934	3059	1875	342
MII_Sand1_1_D3	freshwater beach sand - microcosm output	D	4807908	4422	3047	1374	101
MII_Sand1_1_H6	freshwater beach sand - microcosm output	F	4907297	4518	3067	1451	152
PB_G10	freshwater beach sand	B1	4716719	4431	3048	1383	65
Sewage2002_B10	sewage	A	4617972	4340	3044	1296	81
Sewage2002_C4	sewage	B1	4608924	4292	3041	1251	79
TW09231	freshwater beach	clade III	4016469	4559	3062	1497	173
TW09276	freshwater beach	clade III	3811947	4134	3037	1097	107
TW09308	freshwater beach	clade V	4248945	4714	3086	1628	147
TW10509	feces	clade I	4421679	4926	3064	1862	267
TW11588	soil	clade IV	3903243	4419	3235	1184	125
TW14182	freshwater beach	clade IV	3938151	4415	3052	1363	170
TW15838	freshwater sediment	clade I	4489542	5042	3079	1963	310

*Genomes were generated through PacBio sequencing (*)

[1]The # of accessory genes includes all accessory genes including island genes.

Supplemental Table 3.4. Distribution of enriched modules in output sources.

Cut-off category	M00555 Betaine biosynthesis, choline => betaine	M00136 GABA biosynthesis, prokaryotes, putrescine => GABA	M00045 Histidine degradation, histidine => N-formiminoglutamate => glutamate	M00027 GABA (gamma-Aminobutyrate) shunt
Sand >25%	Complete	Partial	Partial	Partial
Gull >25%	Partial	Partial	Partial	Partial
Sewage >25%	Complete	Partial	NA	Partial
Sand >50%	Complete	NA	Partial	NA
Gull >50%	Partial	Partial	Partial	Partial
Sewage >50%	NA	NA	NA	NA

APPENDIX C.
Preliminary MLST Analysis

Summary

This data is from the preliminary genotypic analysis of *E. coli* used in the microcosm study and was conducted in an effort to gain insight into patterns of allelic distribution in *E. coli* populations from multiple sources including sand, sewage, and gull waste (n=188 each). This was achieved through Multi Locus Sequence Typing (MLST) assessment of portions of the house-keeping genes uidA and rpoS, which were generated through Sanger sequencing. The uidA gene was targeted due to its ubiquitous association with *E. coli*, as it codes for the Beta-glucuronidase enzyme, which is a trademark function of *E. coli* detected in 97.7% of previous environmental *E. coli* samples (Martins et al., 1993). The allelic distribution of the rpoS gene was also assessed, as it codes for the stress response sigma factor rpoS, which is the central regulator of stationary phase. Mutations in this gene have been associated with potential enhanced environmental survival, as some mutant variations have been observed to be selected for in the open environment (Chiang, Dong, Edge, & Schellhorn, 2011; Farrell & Finkel, 2003; Takahashi, Takayanagi, Fujita, Tanaka, & Ishihama, 2006). Together these targets allowed for a robust phylogenetic comparison of populations, which was achieved through building phylogenetic trees using maximum-likelihood analysis of the DNA sequences of these targets. The distribution of these amplicons was compared through assessing the phylogenetic trees generated through maximum likelihood of each target separately and concatenated. Overall, there was more clustering observed between sand and gull amplicons than those from sewage. Phylogenetic association revealed that the lesser common phylogroups of the cryptic clades, D & F, clustered independently, while the members of A and B1, the more common phylogroup were interspersed throughout the tree.

Methods

Samples were isolated from each source with the use of the selective media mTec and stored in 96-well plates at -80°C in a 25% glycerol 75% MUG solution. To prepare for this analysis, freezer samples were cultured overnight (~18hr) at 45.5 °C in MUG media then diluted 1:10 with molecular grade water to form the template for amplification PCR. For each target, uidA & rpoS, a separate amplification PCR and sequencing preparation step was conducted. Briefly the targets in each sample were amplified by combining 12.5 µl 2X Promega master mix, 10.0 µl of molecular grade water, 1.5 µl each forward and reverse primers (300 nM DNA concentration), and 1 µl of template (400 ng DNA concentration). PCR cycling was performed as the following: 5 min at 94 °C, followed by 35 cycles of 92 °C for 1 min, 60 °C for 1 minute, and 72°C for 1 minute, completed with a final extension step of 72°C for 5 min. The amplicons from each sample from were then prepared for sequencing through an additional PCR, in which 0.5 µl of BigDye Terminator, 0.4 µl of primer, 0.4 µl of 5X buffer, 0.1 µl of DMSO, 1.8 µl of molecular grade water, and 1 µl of template (400 ng DNA concentration) were combined and processed in a thermocycler that was programed for 2 min at 96°C followed by 35 cycles of 96°C for 10 sec, 5 sec at 56°C, and 4 minutes at 60°C, finished with a final extension step of 4 min at 60°C. Amplicons from each sample were analyzed for quality control with the program 4 peaks and were analyzed through the use of DNA star (SeqMan NGen®. Version 12.0. DNASTAR. Madison, WI.), Mega 8(Stecher, Tamura, & Kumar, 2020), and iTOL (Tree of Life. Version 5.7). Phylogenetic trees were formed through a maximum-likelihood method with Kimura 2-parameter (K2) model with gamma-distribute rates and invariant sites (G+I), bootstrapped for 1000 replicates.

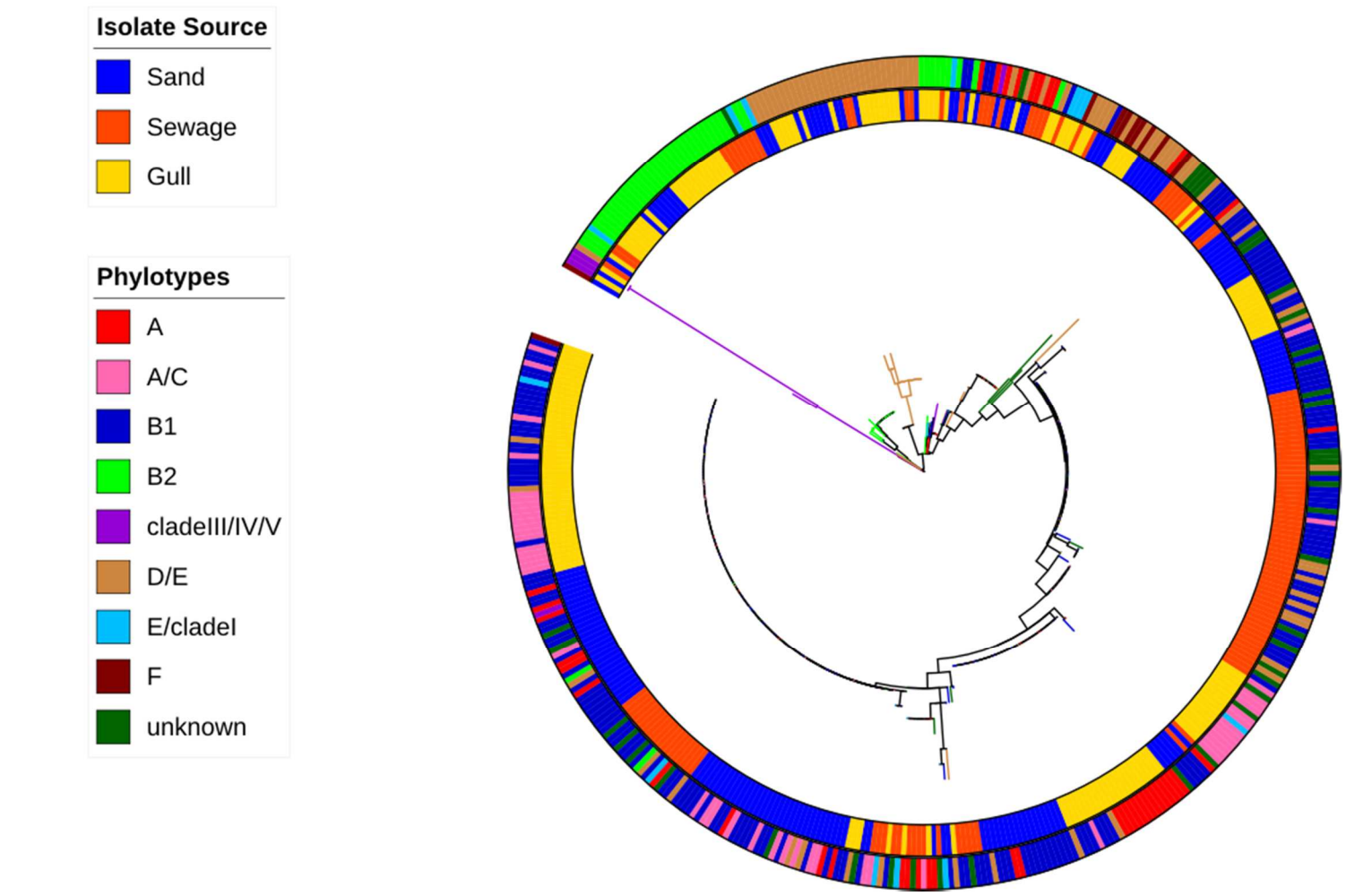


Figure 1. Maximum likelihood of uidA genes in populations of *E. coli* isolated from freshwater beach sand, gull waste, and sewage (Inner ring) in relation to their phylotypic association (outer-ring) as determined through Clermont Phylotyping (Clermont et al., 2013). Note the inter-dispersed clustering of members belonging to A & B1, with other groups such as B2 and D/E clustering much tighter.

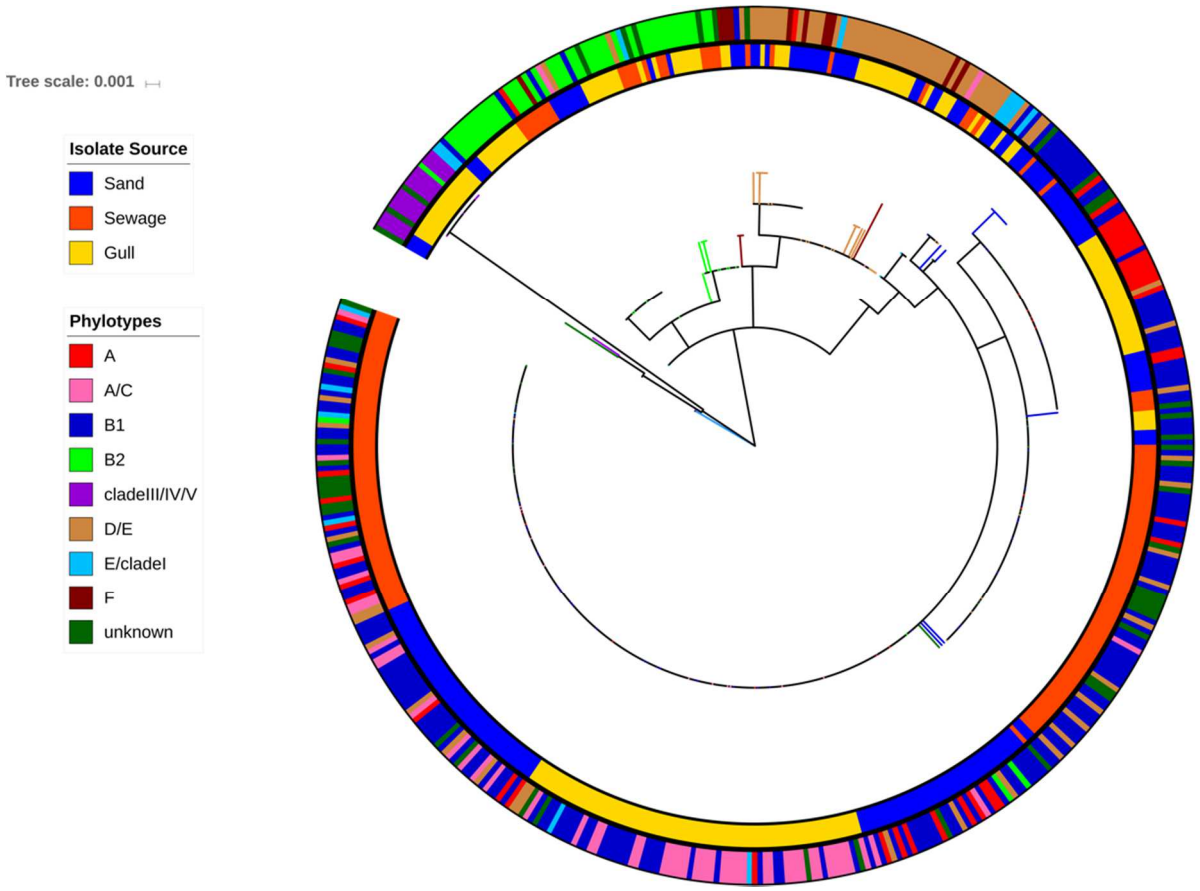


Figure 2. Maximum likelihood of rpoS genes in populations of *E. coli* isolated from freshwater beach sand, gull waste, and sewage (Inner ring) in relation to their phylotypic association (outer-ring) as determined through Clermont Phylotyping (Clermont et al., 2013). Note the inter-dispersed clustering of amplicons was not as stringent as those from uidA. However, it can be observed that the samples the originated from sewage clustered closer with the rpoS amplicon compared to the uidA target.

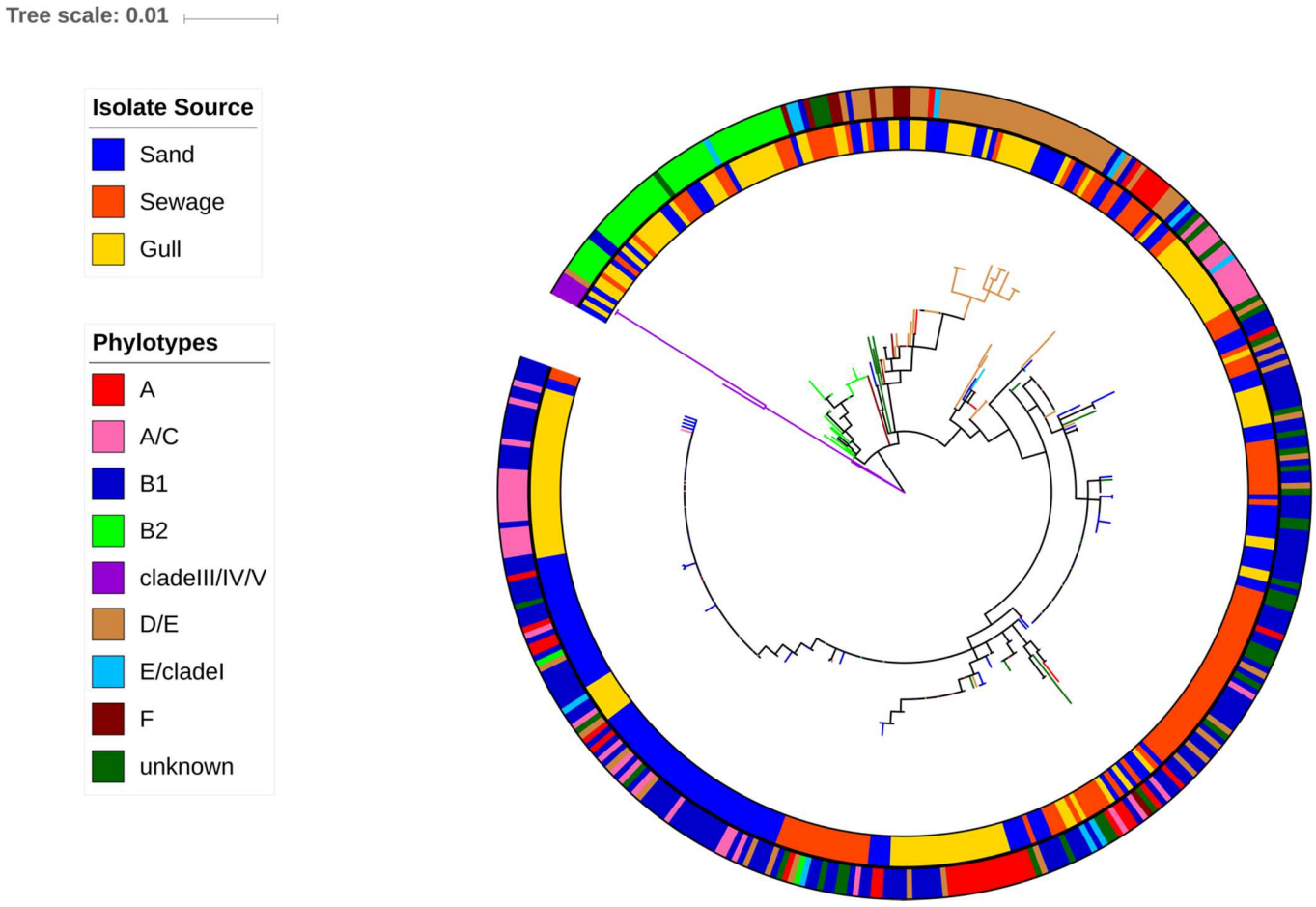
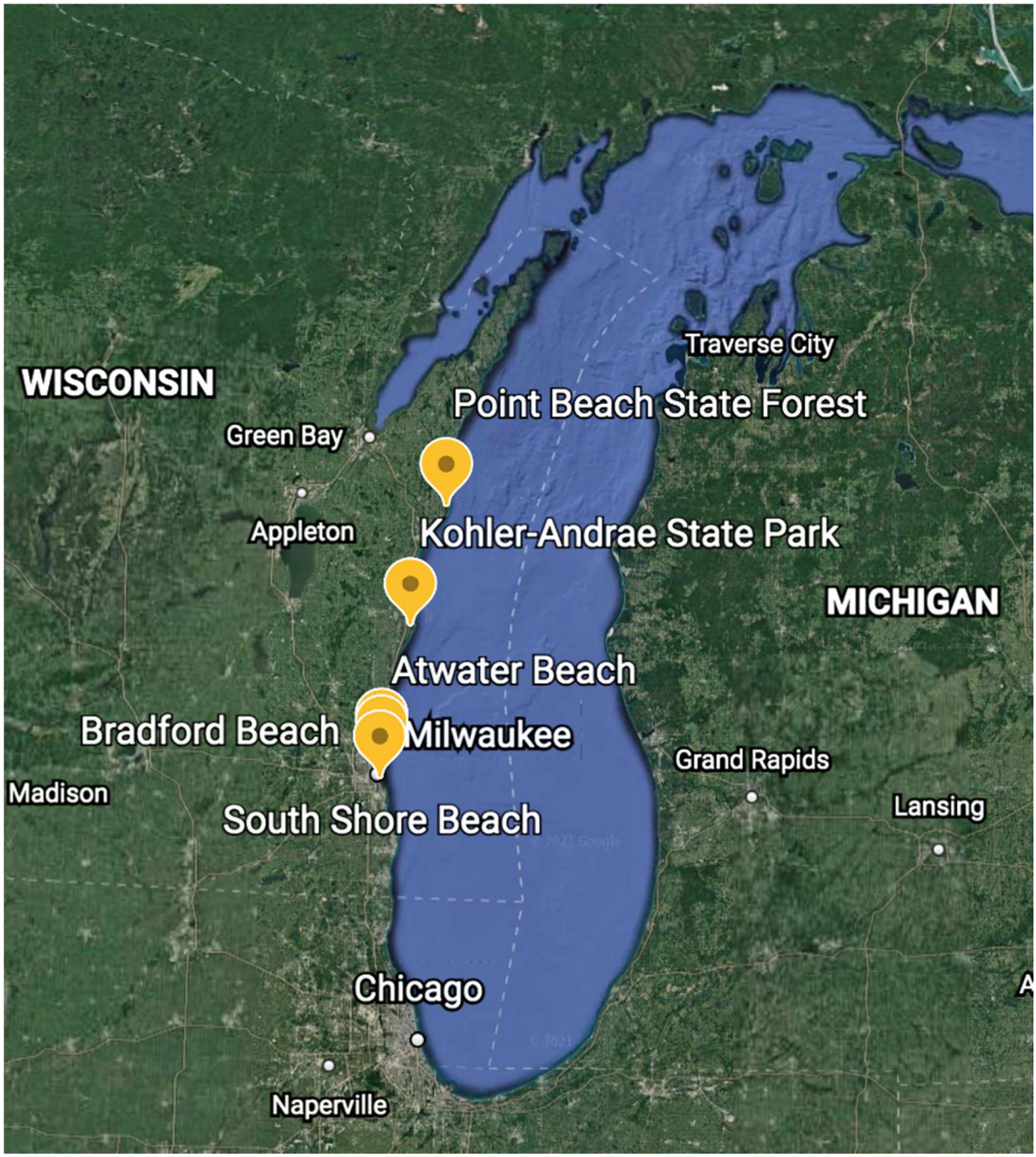
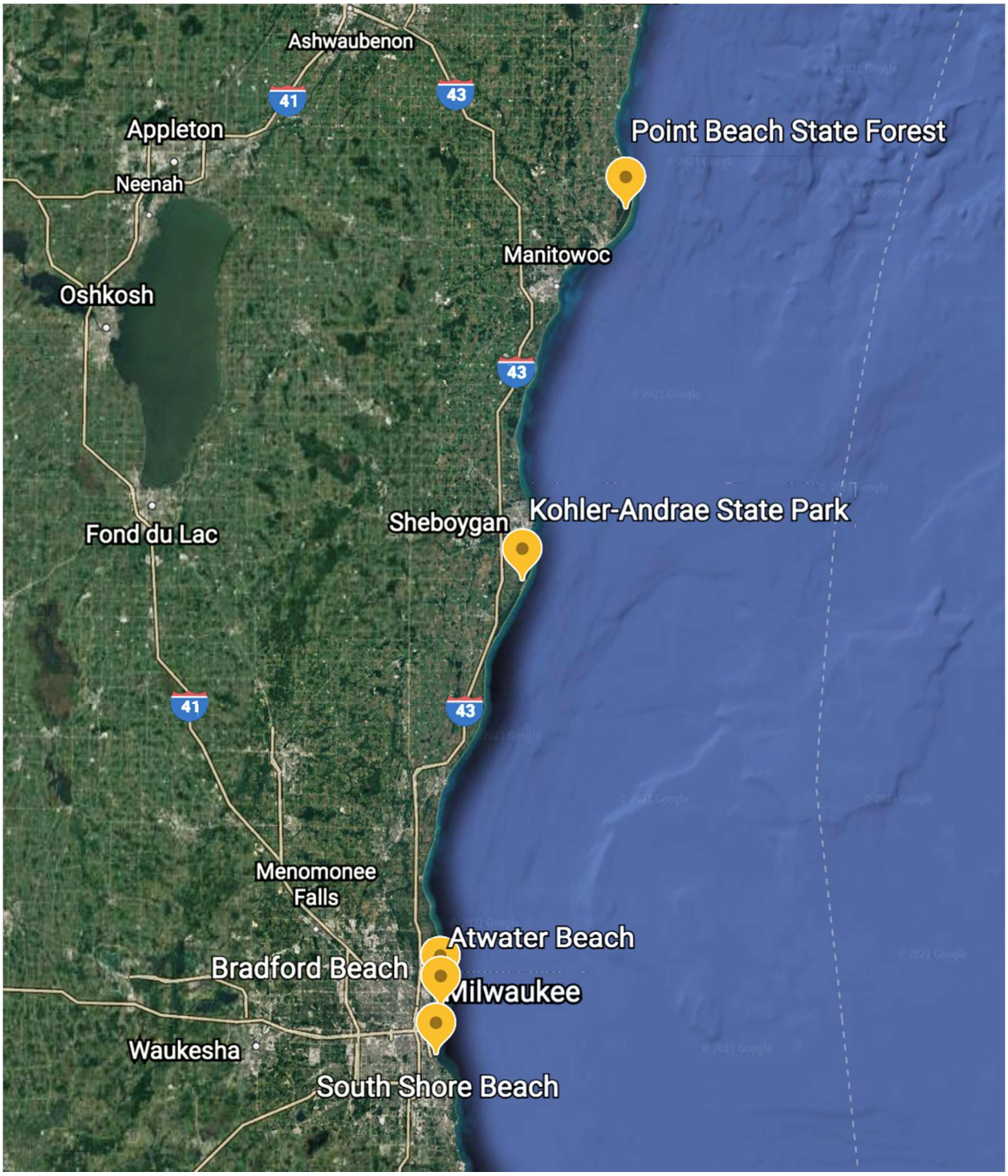


Figure 3. Maximum likelihood of concatenated uidA & rpoS genes in populations of *E. coli* isolated from freshwater beach sand, gull waste, and sewage (Inner ring) in relation to their phylotypic association (outer-ring) as determined through Clermont Phylotyping (Clermont et al., 2013). It can be observed that there was more clustering of sand and gull amplicons together than with sewage, and that overall the B2, clade, and D/E groups clustered tightly independently, with A & B1 members being more widely distributed.

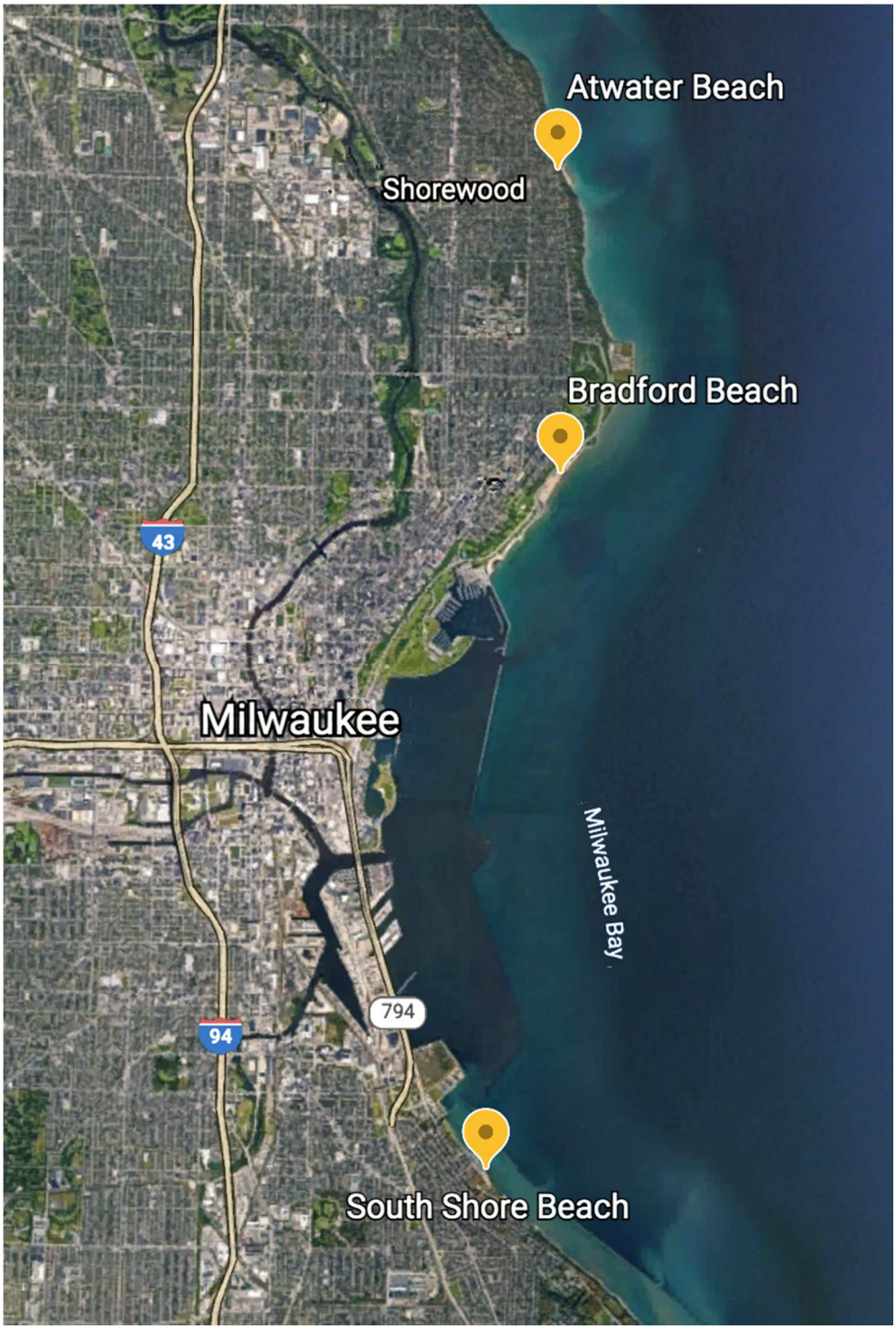
APPENDIX D.
Study site map



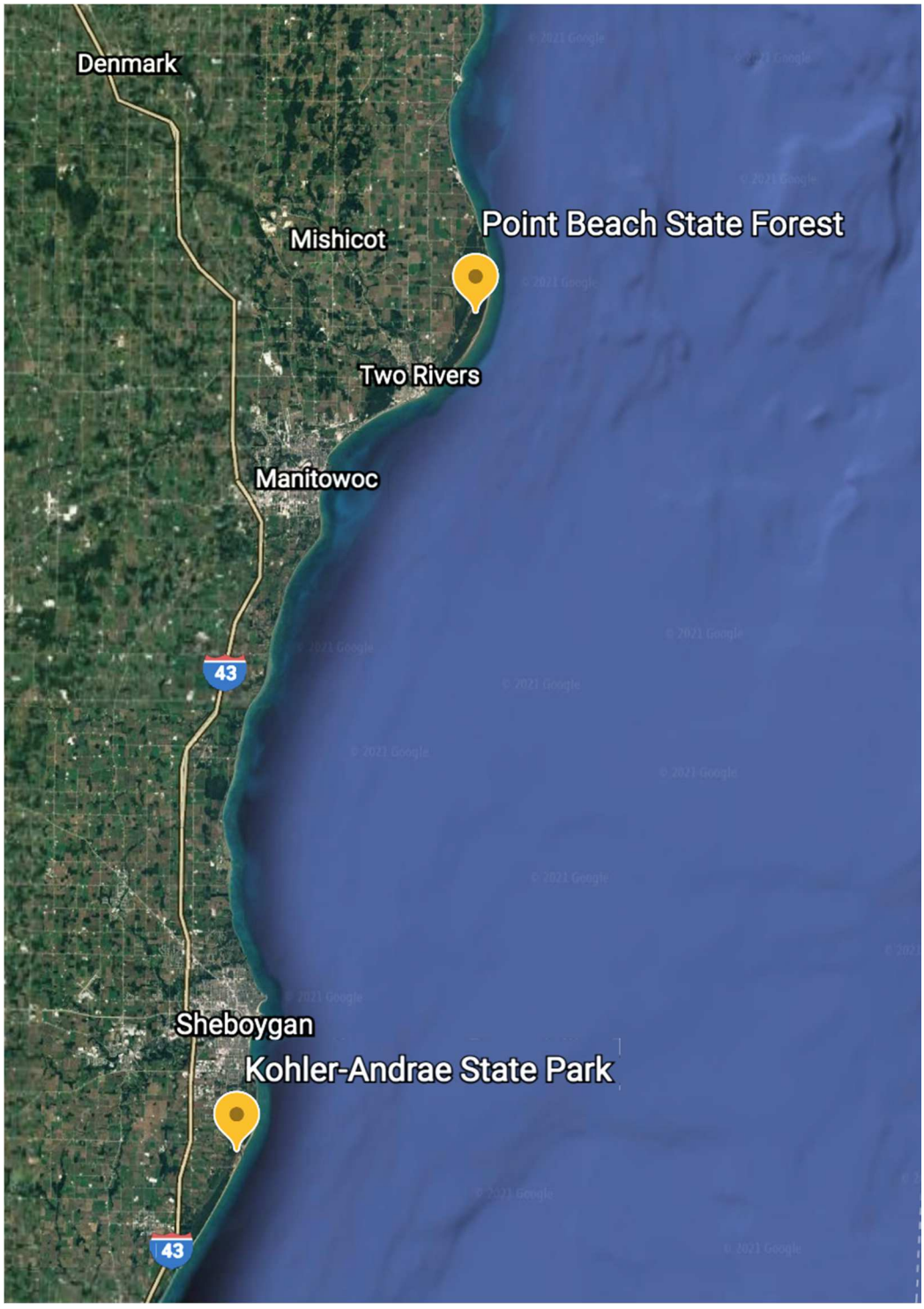
Map 1A. Overview of beach sites used in this study from a lake wide viewpoint. Google Earth, earth.google.com/web/.



Map 1B. Overview of Beach Sites used in this study from a Wisconsin coast viewpoint. Google Earth, earth.google.com/web/.



Map 2. Milwaukee metro sampling sites. Google Earth, earth.google.com/web/.



Map 3: Northern rural sampling sites. Google Earth, earth.google.com/web/.

CURRICULUM VITAE

Natalie Rumball

CURRICULUM VITAE

EDUCATION

University of Wisconsin-Milwaukee	Freshwater Sciences, PhD	2016- 2021
Gannon University	Biology *Magna Cum Laude*, B.S.,	2011-2016

APPOINTMENTS

2016-2021	Graduate Research Assistant of Dr. Sandra McLellan PI, UW-Milwaukee
2014-2016	Laboratory Assistant of Dr. Steven Mauro and Dr. Sarah Ewing, Gannon University
2015-2016	Teacher's Assistant of Melanie Ropski, Molecular Cellular Biology & Animal Form and Function, Gannon University
2014	Laboratory Technician Shipboard Science Program, EPA Lake Guardian Research Vessel

HONORS & AWARDS

2019	Three Minute Thesis Finalist
2018	The Loescher Great Lakes Scholarship
2018	Outstanding Abstract Award American Society for Microbiology 2018
2017	First place poster presentation, American Society for Microbiology North Central Branch Regional Meeting
2016	Magna Cum Laude, Gannon University
2011-2016	Dean's list, Gannon University
2014-2016	Newcombe Scholarship for Mature Women Students
2014	First place undergraduate oral presentation, Regional Science Consortium, Erie PA

PUBLICATIONS

2021	Selective Survival of Escherichia coli Phylotypes in Freshwater Beach Sand; Applied and Environmental Microbiology
2021	The genetic determinates of *E. coli* survival in beach sand; Publisher TBD

LECTURES & CONFERENCE PRESENTATIONS

2019	American Society for Microbiology Annual Meeting 2019, San Francisco, CA, The Phylotypic Characterization of Native *E.coli* Populations in Freshwater Beach Sand

2018	American Society for Microbiology Annual Meeting 2018, Atlanta GA, Phenotypic and Genotypic Characterization of *E coli* in Beach Sand
2017	Navigating the Future of Water, Milwaukee WI, Phenotypic and Genotypic Characterization of *E. coli* in Beach Sand, poster presentation
2017	American Society for Microbiology Regional Meeting, North Central Branch, De Pere WI, Phenotypic and Genotypic Characterization of *E. coli* in Beach Sand, poster presentation
2017	American Society for Microbiology General Meeting, New Orleans, Phenotypic and Genotypic Characterization of *E. coli* in Beach Sand, poster presentation
2015	Regional Science Consortium, Erie PA, The Development of Molecular Methods to Detect Genomic Sequence Variants in Skin Tumors/Lesions on Brown Bullhead, *Ameiurus nebulosus*, from Presque Isle Bay, oral presentation
2015	American Society for Microbiology General Meeting, New Orleans, The Pervasiveness of Cyanobacteria and Microcystin in Lake Erie Water, poster presentation
2014	Regional Science Consortium, Erie PA, The Pervasiveness of Cyanobacteria and Microcystin in Lake Erie Water, oral presentation

VOLUNTEER & ACADEMIC INVOLVEMENT

February 2019 – November 2019	High school student mentor, entered Regeneron science talent search
March 2019	Demonstrator at Women and Girls in Science Event
January 2018	Science Demonstrator – 2018 Sturgeon Bowl
February 2018 - 2021	WI Coastal Beaches Steering Committee
January 2018	Science Demonstrator – 2018 Sturgeon Bowl
September 2017	Volunteer Coordinator/Science Demonstrator - Open Doors MKE
May 2017	Science Demonstrator - Discovery World Career Day
April 2017	Science Demonstrator/Guide - Girl Scout Troupe 722

WORKSHOPS & OTHER SPECIAL TRAINING

February 2018	Introduction to Research Computing
March 10, 2018	Introduction to parallel computing
July 2017	Michigan Sea Grant Extension Community-Engaged Research Institute
February 2017	Nature Master Class – Writing workshop
July 2018	Explorations in Data Analysis for Metagenomics Advances in Microbial Ecology (EDAMAME)
Summer 2015	Niagara Field Ecology Program, US Brig Niagara

PROFESSIONAL ACTIVITES

2016-2021	President of American Society for Microbiology Student Chapter at UW-Milwaukee
2014-present	Member of the International Torch Club

MEDIA COVERAGE

August 2020	Interview on NPR Lake Effect
February 2018	UMW 2018 research magazine- graduate feature story
September 2017	Milwaukee Radio 88.9 – in the wings, interview