# A Polyphasic Perspective on *Escherichia coli* Niche-Specificity: The Characterization of Naturalized, Wastewater-Specific *E. coli* and the Emergence of Wastewater Treatment Resistance in the Microbial World

by

Daniel Yu

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Public Health

University of Alberta

School of Public Health

# Abstract

Although typically considered to be capable of colonizing and transiting between several different niches, evidence suggests that the *Escherichia coli* species exhibits a significant degree of host- and niche-specificity. This has led to the hypothesis that *E. coli* taxonomy may be more appropriately described using a 'species-complex' model, consisting of several distinct niche-specific groups known as 'ecotypes'. While various microbial taxonomic schemes have been developed to date, current conventional microbial diagnostic and classification methods appear to be unsuitable for reliably identifying and differentiating between microbial ecotypic groups.

Herein, we describe the application of a logic regression-based workflow for the discovery of putative *E. coli* ecotypes, and the subsequent polyphasic characterization of the specific genotypic, phenotypic, and ecotypic adaptations underlying their niche-specificity. Building on previous studies in our laboratory, logic regression was used to analyse the sequence variation contained within intergenic regions (ITGRs) across the *E. coli* genome. Reinforcing previous findings, logic regression modelling was found to generate highly niche-specific single nucleotide polymorphism (SNP) biomarkers across a wide range of host and environmental niches that not only appeared to demarcate putative *E. coli* ecotypes, but also could be used for the source (i.e., ecotype) attribution of unknown environmental strains. Interestingly, in addition to being effective for differentiating between host-associated *E. coli* ecotypes, logic regression also consistently distinguished wastewater-derived strains from their host-associated counterparts, suggesting they may represent a unique *E. coli* ecotype adapted to engineered environments as primary niches.

Through a polyphasic approach, various characteristics underlying this 'naturalized-engineered' ecotype were identified. Wastewater- and meat plant-derived strains were found to group into naturalized-engineered-associated sequence types (ST635 and ST399) and several

serotypes, representing distinct lineages that may have independently emerged across food- and water-associated engineered environments. Strains belonging to these naturalized-engineered ecotypic groups were characterized by unique genetic traits, as they were enriched in genes and genetic markers associated with biofilm formation, microbial defense, and stress resistance (i.e., against DNA-damaging stimuli, oxidizing agents, heat, heavy metals), but lacked host-adaptive genes related to colonization and virulence. Recapitulating these genotypic findings, wastewater-specific (WWS) *E. coli* strains were also characterized by various phenotypic adaptations that could promote their survival across the wastewater treatment train, including enhanced resistance to heat, slower growth kinetics, and enhanced biofilm formation under nutrient-limiting and low temperature conditions.

As such, with the WWS strains as a model system, we were able to demonstrate the utility of our logic regression-based, polyphasic workflow for the exploration of putative ecotypes within the *E. coli* species. While these findings have important conceptual implications for prokaryotic taxonomy, especially in how bacterial species may be defined, the characterization of a wastewater treatment resistant ecotype implies that other *E. coli* populations could also be evolving resistance to wastewater disinfection. Reflecting this, comprehensive comparative genomic approaches revealed that other, non-naturalized wastewater-derived *E. coli* strains surviving wastewater treatment were virtually identical to clinically relevant extraintestinal pathogenic *E. coli* (ExPEC) strains. Our findings, therefore, also point to a concerning public health prospect – that pathogenic microbes could similarly be evolving resistance to wastewater treatment and disinfection.

# Preface

Chapter One of this thesis has been published as D. Yu, G. Banting, and N.F. Neumann, "A review of the taxonomy, genetics, and biology of the genus *Escherichia* and the type species *Escherichia coli*" in the *Canadian Journal of Microbiology*, vol. 67, issue 8 (2021): 553–571. D. Yu and N.F. Neumann drafted the initial manuscript. All authors contributed to the revisions of the manuscript.

Chapter Two of this thesis has been published as D. Yu, M. Andersson-Li, S. Maes, L. Andersson-Li, N.F. Neumann, M. Odlare, and A. Jonsson, "Development of a logic regression-based approach for the discovery of host- and niche-informative biomarkers in *Escherichia coli* and their application for microbial source tracking" in *Applied and Environmental Microbiology*. D. Yu, S. Maes, M. Odlare, N.F. Neumann, and A. Jonsson conceived the study. S. Maes and A. Jonsson collected the water and faecal samples, while S. Maes isolated the *E. coli* strains from the collected samples. M. Andersson-Li and L. Andersson-Li provided the logic regression code. D. Yu performed the bioinformatic and logic regression analyses. D. Yu drafted the manuscript, and all authors helped revise the manuscript drafts.

Chapters Three and Four of this thesis have been jointly published as D. Yu, P. Stothard, and N.F. Neumann, "Emergence of potentially disinfection-resistant, naturalized *Escherichia coli* populations across food- and water-associated engineered environments" in *Scientific Reports*, vol. 14, issue 13478 (2024): 1–14. D. Yu and N.F. Neumann conceived the study. P. Stothard provided the remote server for all *in silico* analyses. D. Yu performed all phylogenetic and bioinformatic analyses. D. Yu and N.F. Neumann drafted the manuscript. All authors helped revise the manuscript drafts.

Chapter Six of this thesis has been published as D. Yu, K. Ryu, S.J.G. Otto, P. Stothard, G. Banting, N.F. Neumann, and S. Zhi, "Differential survival of septicemia- and meningitis-causing pathogenic *E. coli* across the wastewater treatment train – a potentially emerging public health problem?" in *npj Clean Water*, vol. 5, issue 1 (2022): 1–12. N.F. Neumann and S. Zhi conceptualized and planned the experiments. G. Banting conducted the screening tests to identify the putative wastewater ExPEC strains. D. Yu, S. Zhi, and K. Ryu performed the bioinformatic analyses. D. Yu, S. Zhi, and N.F. Neumann drafted the initial manuscript, and all authors helped revise the manuscript drafts.

## Acknowledgements

As I reflect upon the end of this PhD journey and all of its highs and lows, I would be remiss without properly thanking all the wonderful people who have helped me along the way. First and foremost, I would like to express my sincerest gratitude to my supervisor, Dr. Norman Neumann. As someone wise once said, "research is 90% perspiration and 10% inspiration" – and I am incredibly grateful for your support and encouragement during the 90% and that you continued to challenge and push me during the other 10%. Thank you for being an amazing supervisor, mentor, and friend (and for all the rigorous debates and board games matches!) over the last five years, your advice has helped me to not only become a better researcher but also a better person. I would also like to express my thanks to my supervisory committee for their insight and help with the research project. To Dr. Paul Stothard, thank you for sharing all of your expertise in bioinformatics, especially when my analyses (inevitably!) went wrong. To Dr. Irina Dinu, thank you for your help in keeping my statistical analyses in check. To Dr. Ben Willing, thank you asking the tough questions, and getting me to always think critically about my research.

I am also incredibly grateful for all the wonderful collaborators who helped make a lot of the work in this thesis possible. To Dr. Shuai Zhi, thank you for your guidance in continuing your work on the niche-specificity and evolution of wastewater treatment resistance in *E. coli*, and for all of the fruitful collaborations that we were able to work on together. Similarly, I would like to express my thanks to Dr. Anders Jonsson, Dr. Sharon Maes, and Dr. Martin Andersson-Li for your help with the isolate collection, sequencing, and coding for the logic regression analyses, and for your enthusiasm which allowed us to extend its use for source tracking for the first time.

To everyone in the Hanington and Otto labs in Environmental Health Sciences, thank you creating a friendly learning environment over the years. Additionally, thank you to everyone in

Neumann lab for your endless support. To Candis Scott, thank you for teaching me everything I know in the lab, and for always having the time and patience to answer my (many!) questions about my experiments and even just to chat. To Kanghee Ryu, thank you for your friendship over the last five years, and for your generosity, mentorship, and willingness to help, especially early on in my graduate studies when I was still learning the ropes. To Dr. Graham Banting, Liam Carson and Markus Gaenzle, thank you for all of your support and encouragement.

I would also like acknowledge all the funding and scholarships that made this thesis possible, including from: the School of Public Health, the Faculty of Graduate Studies and Research, the Natural Sciences and Engineering Research Council of Canada, and Alberta Innovates.

Last but not least, I would like to wholeheartedly thank all of my friends and family that have helped me along this journey. Thank you to all my friends who have not only provided me unending support and encouragement during tough times, but also introduced me to my many passions and current obsessions (including snowboarding, board games, and music). Most importantly, I am eternally grateful for the unconditional love, support, and patience from my parents, Ken Yu and Quan Le Chung, and siblings, Vincent Yu and Kelly Yu – you all inspire me to do and be better every day.

# Table of Contents

# List of Tables

# List of Figures

## List of Abbreviations

| | |
|---|---|
| aEPEC | Atypical enteropathogenic *E. coli* |
| AFLP | Amplified fragment length polymorphism |
| AIEC | Adherent-invasive *E. coli* |
| AME | Aminoglycoside modification enzyme |
| ANI | Average nucleotide identity |
| APEC | Avian pathogenic *E. coli* |
| ARCC | Average rate of correct classification |
| ARG | Antibiotic resistance gene |
| BBEC | Bloodborne *E. coli* |
| bp | Base-pairs |
| CARD | Comprehensive Antibiotic Resistance Database |
| C-ExPEC | Clinical ExPEC |
| CGGC | Compare Groups of Growth Curves |
| CSI | Conserved signature indels |
| DAEC | Diffuse-adherent *E. coli* |
| DNA | Deoxyribonucleic acid |
| *E. coli* | *Escherichia coli* |
| EAEC | Enteroaggregative *E. coli* |
| EAF | *E. coli* adherence factor |
| ECOR | *E. coli* reference collection |
| EHEC | Enterohaemorrhagic *E. coli* |
| EIEC | Enteroinvasive *E. coli* |
| EPEC | Enteropathogenic *E. coli* |
| ERIC | Enterobacterial repetitive intergenic consensus |
| ESBL | Extended-spectrum beta-lactamase |
| ETEC | Enterotoxigenic *E. coli* |
| ExPEC | Extra-intestinal pathogenic *E. coli* |
| gDNA | Genomic DNA |
| HGT | Horizontal gene transfer |
| IMViC | Indole, methyl-red, Voges-Proskauer and citrate reaction |
| InPEC | Intestinal pathogenic *E. coli* |
| ITGR | Intergenic region |
| kb | Kilobases |
| KCN | Potassium cyanide |
| LHR | Locus of heat resistance |
| MAR | Multiple antibiotic resistance |
| Mb | Megabases |
| MGE | Mobile genetic element |
| ML | Megaliter |
| µL | Microliter |
| mL | Milliliter |
| MLEE | Multilocus enzyme electrophoresis |
| MLSA | Multilocus sequence analysis |
| MLST | Multilocus sequence typing |

| | |
|---|---|
| MM | Minimal media |
| MPN | Most probable number |
| MPS–*E. coli* | Meat plant-specific *E. coli* |
| MST | Microbial source tracking |
| NMEC | Neonatal meningitic *E. coli* |
| OD600 | Optical density, measured via absorbance at 600nm |
| Pan-GWAS | Pan-genome-wide association study |
| PBS | Phosphate-buffered saline |
| PCR | Polymerase chain reaction |
| PFGE | Pulsed-field gel electrophoresis |
| ppm | Parts per million |
| qPCR | Quantitative PCR |
| RAPD | Random amplified polymorphic DNA |
| rDNA | Ribosomal DNA |
| rep-PCR | Repetitive extragenic palindromic sequence PCR |
| RFLP | Restriction fragment length polymorphism |
| RNA | Ribonucleic acid |
| rRNA | Ribosomal RNA |
| SNP | Single nucleotide polymorphism |
| SRA | Sequence Read Archive |
| ST | Sequence type |
| STEC | Shiga toxin-producing *E. coli* |
| tLST | Transmissible locus of stress tolerance |
| TraDIS | Transposon-directed insertion sequencing |
| TSA | Tryptic soy agar |
| TSB | Tryptic soy broth |
| UPEC | Uropathogenic *E. coli* |
| UV | Ultraviolet |
| VFDB | Virulence Factor Database |
| VG | Virulence gene |
| W-ExPEC | Wastewater ExPEC |
| WWS-*E. coli* | Wastewater-specific *E. coli* |
| WWTP | Wastewater treatment plant |

# Chapter One: Introduction[1]

## 1.1 Introduction

In 1885, Dr. Theodor Escherich identified a common commensal of the gastrointestinal tract, which he termed *Bacterium coli commune,* from the fecal material of neonates and early infants (Escherich 1988 [English translation of his original work]). Originally designated as *Bacillus coli* in 1895, this bacterium was renamed as *Escherichia coli* after its founder in 1919 (Henry 2015). The revised nomenclature was then officially recognized in 1958, thereby establishing *Escherichia* as a genus with *E. coli* as the first species. Since its discovery, *E. coli* has become one of the most-studied and best-characterized microorganisms, serving as one of the fundamental model systems in microbiology. While *E. coli* has been subject to extensive genotypic and phenotypic (i.e., biochemical) examination, and as such remains the best-characterized and the representative type species within the genus, the other *Escherichia* species have been characterized only relatively recently. As such, the taxonomic status of the rest of the genus has experienced considerable flux as conventional classification schemes progressed from being predominantly biochemically-based to including genotypic- and genomic-based techniques.

Interestingly, the same advances made to the diagnostic and classification techniques underlying bacterial systematics have also revealed an extensive degree of diversity within the *E. coli* species. Following an overview of the history of, and changes to, the *Escherichia* genus, this

[1] A version of this chapter has been published as: Yu, D., Banting, G., and Neumann, N.F. 2021. A review of the taxonomy, genetics, and biology of the genus *Escherichia* and the type species *Escherichia coli*. Can. J. Microbiol. 67(8): 553–571. doi:10.1139/cjm-2020-0508. This manuscript received the Editor's Choice Award for the August 2021 issue of the Canadian Journal of Microbiology.

literature review explores how the extensive genetic and biological diversity of *E. coli* appears to be reflected in its vast host- and niche-specific lifestyles. Importantly, despite being the prototypical species within the *Escherichia* genus, we argue that current conventions of *E. coli* as a cohesive species may be misleading. Indeed, the great genotypic, phenotypic and ecotypic diversity that can exist between strains suggests that *E. coli* may be more appropriately described as a species-complex.

### 1.1.1 The Genus *Escherichia*

*E. coli* remained the sole member of the *Escherichia* genus until 1962 when Leclerc (1962) first characterised the biochemical profile of a group of *Escherichia* strains using the indole, methyl-red, Voges-Proskauer and citrate (IMViC) reaction test, prompting their classification as a new species known as *Escherichia adecarboxylata*. Following this, the biochemical characterization of a group of bacterial strains isolated from the hindgut of the Oriental cockroach (*Blatta orientalis*) led to their designation as *E. blattae* (Burgess *et al.* 1973). Further examination of the biochemical boundaries defining the *E. coli* species revealed several additional atypical groups that appeared to be distinct species in the family *Enterobactericeae*. This included the atypical Enteric Group 1 and Enteric Group 11, which were both commonly associated with clinical specimens. Both atypical groups were found to consist of yellow-pigmented enteric strains capable of growing on potassium cyanide (KCN) media and metabolizing cellobiose, distinguishing them as distinct species from *E. coli*. These two atypical groups were themselves differentiated based on their decarboxylase reaction profiles and other metabolic differences, leading to the designation of Enteric Group 1 and Enteric Group 11 as separate species, named *E. vulneris* and *E. hermannii* respectively (Brenner *et al.* 1982a; Brenner *et al.* 1982b). Similarly, the

distinct biochemical profile of the atypical Enteric Group 10, also associated with human clinical samples, prompted its reclassification as *E. fergusonii* (Farmer *et al.* 1985).

While the early identification of novel *Escherichia* species relied on the biochemical differentiation of closely related groups of *Enterobacteriaceae*, the development of genotypic and genomic diagnostic tools provided additional means for the discrimination of *Escherichia* species. For instance, a set of strains initially named as *Hafnia alvei* were later reclassified as a separate *Escherichia* species*, E. albertii* (Huys *et al.* 2003), following biochemical profiling (Abbott *et al.* 2003; Janda *et al.* 1999) supplemented with genotypic approaches including DNA-DNA hybridization, 16S rDNA sequencing analyses and virulence gene screening (i.e., for the enteropathogenic *E. coli*-specific *eaeA* gene and the *Shigella*-specific *phoE* gene). This taxonomic classification has since supported with whole-genome sequencing and average nucleotide identity (ANI) analyses (Ooka *et al.* 2015). More recently, phylogenetic analyses utilizing the 16S rRNA gene and core genome sequences classified a group of enterobacterial strains isolated from fecal samples of the Himalayan marmot (*Marmota himalayana*) as a novel *Escherichia* species, *E. marmotae* (Liu *et al.* 2015; Liu *et al.* 2019).

The growth of DNA sequence-based diagnostic methods also proved essential for the characterization of the 'cryptic' *Escherichia* lineages. The cryptic *Escherichia* consist of distinct strains commonly associated with non-host environments that appear to be phenotypically indistinguishable, but genotypically divergent, from *E. coli.* Traditional biochemical tests fail to distinguish these cryptic lineages from *E. coli* (Walk *et al.* 2009), but multilocus sequence typing (MLST)-based phylogenetic analyses have identified 5 divergent, monophyletic groups of cryptic strains, termed the cryptic *Escherichia* clades I–V (Walk *et al.* 2009; Clermont *et al.* 2011a). A growing body of evidence suggests that these cryptic clades, especially clades III and V, consist

of strains that are native residents of natural environments. Reflecting this, comparative genomic and physiological studies have demonstrated cryptic *Escherichia* strains to be particularly adapted to non-host environments. For instance, the cryptic strains possess several genes, such as those for diol utilization and lysozyme production, commonly associated with environmental survival while simultaneously lacking genes, including those related to antibiotic resistance, adhesins, and sugar transporters, commonly associated with gastrointestinal colonization (Luo *et al.* 2011). Physiologically, the cryptic lineages also appear capable of growing at lower temperatures and forming robust biofilms compared to other *Escherichia* species, both of which are suggested to enhance survival in external environments (Ingle *et al.* 2011). Given these adaptations, there is relatively strong support for the designation of the cryptic clades as ecologically distinct subpopulations within the genus *Escherichia*. Indeed, while clade I may represent a divergent *E. coli* subgroup, MLST and ANI data propose that the other cryptic clades comprise novel *Escherichia* species, one consisting of clades III and IV, and another consisting of clade V (Walk 2015; Beghain *et al.* 2018). Interestingly, while the cryptic clades appear to be predominantly associated with natural environments, they have also been sparingly identified in a variety of animal hosts, particularly in birds (Blyton et al. 2015; Clermont et al. 2011a).

As the type species of the genus, the classification of *E. coli* has remained stable since its discovery; however, the taxonomic status of the rest of the genus has experienced considerable flux. While early *Escherichia* 'species' were identified based on the characterization of biochemical profiles similar to that of *E. coli*, they were later determined to be phylogenetically distant from the rest of the genus through genotypic and genomic analyses. For instance, DNA hybridization analyses have since demonstrated the divergence of *E. adecarboxylata* from other *Enterobacteriaceae* species, leading to its reclassification as *Leclercia adecarboxylata* (Tamura *et*

*al.* 1986). Similarly, 16S rRNA sequencing analyses revealed that *E. blattae* was incorrectly classified within the *Escherichia* genus which, alongside corroborating MLST analyses, prompted its transfer to another genus as *Shimwellia blattae* (Priest and Barker 2010). More recently, the designation of *E. vulneris* and *E. hermannii* as *Escherichia* species have also come into question. *E. hermannii*, for instance, has been found to be both biochemically and phylogenetically distant from the other *Escherichia* species, leading to its proposed reclassification into a novel genus as *Atlantibacter hermannii* (Hata *et al.* 2016). Similarly, *E. vulneris* was found to exhibit distinct metabolic reaction profiles compared to other *Escherichia* species, which was later corroborated by separate core genome phylogenomic analyses. In line with this evidence, using an approach that identified characteristic molecular markers referred to as conserved signature indels (CSI) in key *Enterobacteriaceae* proteins, Aljanar and Gupta (2017) demonstrated that *E. vulneris* did not harbor any of the CSI markers specific to the *Escherichia* genus, justifying its transfer to a novel genus as *Pseudescherichia vulneris*.

The development of genotypic and genomic-based diagnostic tools has also helped clarify the taxonomic relationship between *E. coli* and *Shigella,* a previously unique genus in the family *Enterobacteriaceae*. Upon its discovery in 1898 by Dr. Kiyoshi Shiga, the dysentery-causing bacillus was originally classified in the same tentative genus as *Bacillus coli* (now *E. coli*), where it was named *Bacillus dysenteriae* (Trofa et al. 1999). Despite their apparent relatedness, the distinct antigenic properties of these strains prompted their re-classification into the *Shigella* genus (Ewing 1949). This distinction has since been maintained based on the non-motility, distinct biochemical profile, and patterns of pathogenesis characterizing the *Shigella* species (Van Den Beld and Reubsaet 2012). Although the morphological and biochemical evidence support the designation of *Shigella* as a unique genus, sequence-based diagnostic methods reveal a

considerably high degree of relatedness between *Shigella* and *E. coli*. Indeed, a variety of molecular methods, including 16S rRNA sequencing and MLST, fail to distinguish *Shigella* strains from *E. coli* (Devanga Ragupathi et al. 2018), while core-genome based phylogenetic approaches cluster *Shigella* and *E. coli* strains together (Zhou et al. 2010), suggesting that they collectively represent a single species. As such, although it remains a unique genus largely for clinical reasons, by various metrics *Shigella* appears to be phylogenetically *E. coli* and essentially a specific pathovar within the *E. coli* species (Van Den Beld and Reubsaet 2012).

Provided the reclassification of many early-identified *Escherichia* species and the cryptic *Escherichia* clades, there are now three recognized species within the genus. This includes: *E. coli* (including *Shigella,* as the type species), *E. fergusonii*, and *E. albertii*, with *E. marmotae* as a potentially novel member species within *Escherichia* (Figure 1-1). Despite the ever-changing status of the other members of the genus, *E. coli* remains steadfast as the representative *Escherichia* species*,* supported by both biochemical and phylogenetic evidence. As demonstrated by the reclassification of *L. adecarboxylata, S. blattae*, *A. hermannii*, and *P. vulneris*, however, the diagnostic tools underlying current taxonomic schemes may fail to consider all criteria pertinent for classification. Indeed, it has been suggested that the current emphasis on genotype for the basis of bacterial taxonomy is too restrictive, and it fails to consider the importance of phenotype (Kämpfer 2014), the processes of speciation (Georgiades and Raoult 2011), and ecology. Indeed, we argue that the intragenotypic and genomic variability (i.e., diversity in gene content and core- and pan-genome dynamics), diversity in lifestyles (i.e., pathogenic vs commensal vs naturalized), and wide array of host- and niche-specific strains (i.e., ecotypes) indicates that *E. coli* may be better described as a species complex.

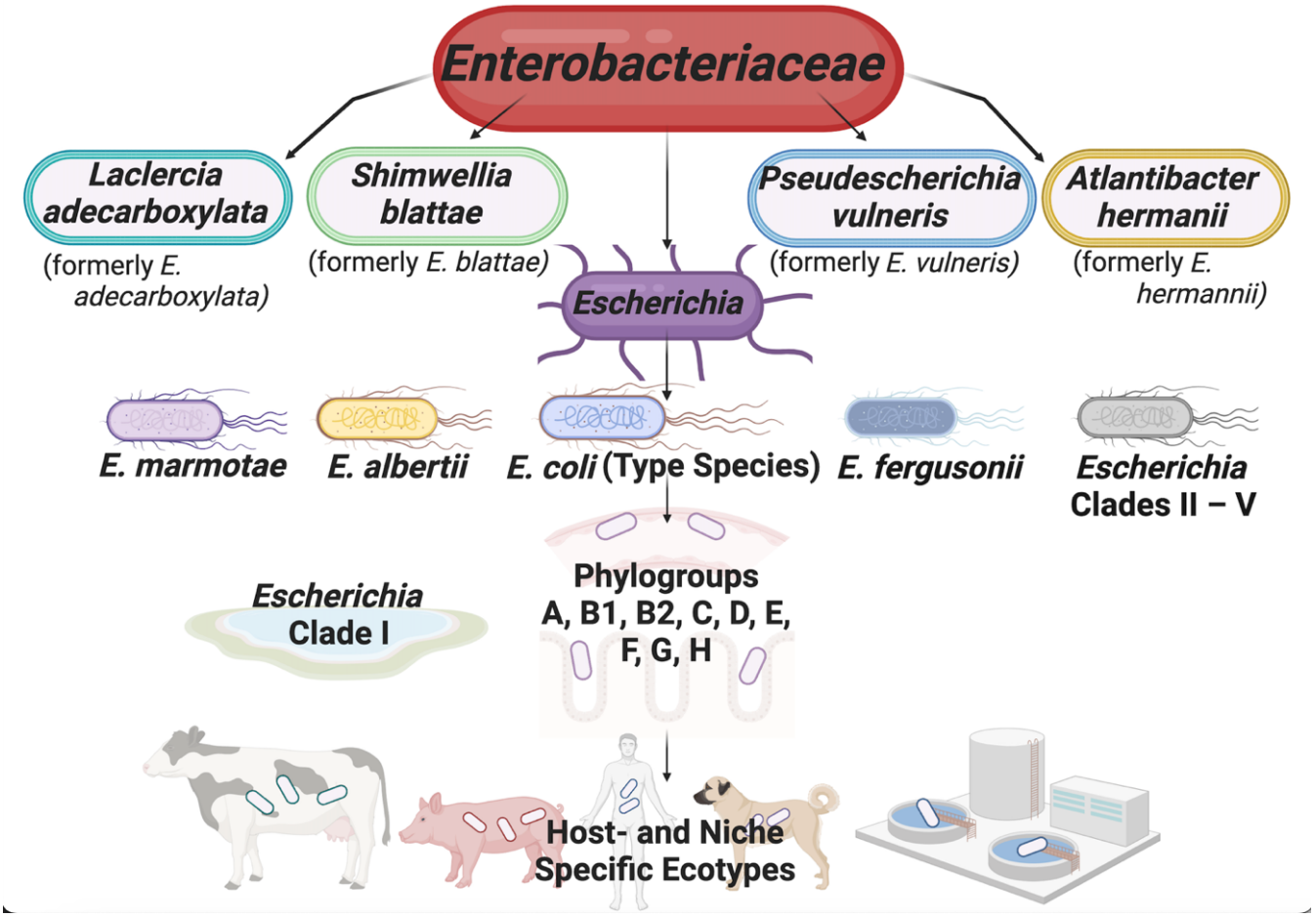



**Figure 1-1. The taxonomic structure of the *Escherichia* genus and the type species *E. coli.*** The growth and development of genotypic diagnostic methods from biochemical methods has greatly informed the current taxonomic status of the *Escherichia* genus, as early-identified species such as *L. adecarboxylata, S. blattae, P. vulneris,* and *A. hermannii* have been reassigned to other genera, while new lineages including the cryptic *Escherichia* clades have been discovered. While biochemical and phylogenetic evidence continue to place *E. coli* as the representative species within the genus, the great genetic and biological diversity coupled with the extensive host- and niche-specificity of the model bacterium suggests that a reconsideration of current taxonomy is warranted, as *E. coli* may consist of a complex of several distinct host- and niche-specific ecotypes.

### 1.1.2 Basic *Escherichia coli* Biology

*Escherichia coli* is one of the most well-characterized model organisms in microbiology. The reference lab strain *E. coli* K-12 substrain MG1655 and its derivatives have played key roles in the advancement of several scientific fields, including microbiology, genetics, and molecular biology. Outside of the laboratory environment, the Gram-negative, non-sporulating facultative anaerobe comprises part of the faecal microbiome of various vertebrate host species including mammals, birds, and reptiles (Gordon and Cowling 2003). There is great variation in the body size, gastrointestinal physiological and biochemical conditions, diets, and microbiome composition between the various host species that *E. coli* can colonize. Thus, it follows that there is also substantial variation in the prevalence of *E. coli* among these host species, which can range from more than 90% in humans (Penders *et al.* 2006), to 56% in wild mammals, 23% in birds, and only 10% in reptiles (Gordon and Cowling 2003).

In the vertebrate gastrointestinal environment, commensal *E. coli* strains reside in the mucus layer covering the epithelial cells along the tract, especially in the caecum and colon of the large intestine (Tenaillon *et al.* 2010). This overlying mucus layer is rich in mucin, which primarily consists of glycoproteins with various O-linked glycans (Marcobal *et al.* 2013), representing a nutrient-rich niche enriched with adhesion sites for *E. coli* colonization. While *E. coli* lacks the appropriate enzymatic complement to directly degrade the complex mucin-associated polysaccharides, the mucolytic activity of other commensal gastrointestinal anaerobes facilitates the release of key mono- and disaccharides for *E. coli* metabolism (Conway and Cohen 2015). Beyond acting as a carbon source, the mucin-associated O-linked glycans can also serve as ligands for *E. coli* pili/fimbriae and flagella to initiate adhesion and biofilm formation, thereby subsequently promoting *E. coli* colonization within the gastrointestinal tract (Sicard *et al.* 2017).

### 1.1.3 Commensal and Pathogenic *Escherichia coli*

Under normal circumstances, *E. coli* exists as a commensal in the vertebrate gastrointestinal tract; however, various strains have acquired virulence factors enabling them to become pathogenic in both humans and animals. The pathogenic *E. coli* are responsible for causing a wide range of intestinal infections, which includes various diarrheal diseases, as well as extra-intestinal infections such as urinary tract infections, sepsis, and meningitis (Bekal *et al.* 2003). A wide variety of *E. coli* pathotypes have been described, including the:

a) enterotoxigenic *E. coli* (ETEC), associated with traveler's diarrhea as well as diarrhea in porcine and bovine hosts (Kai *et al.* 2010; Luppi 2017);
b) enteropathogenic *E. coli* (EPEC), associated with diarrhea in children and various animals (Kai *et al.* 2010);
c) enterohaemorrhagic *E. coli* (EHEC), including strains that produce Shiga toxin (STEC), associated with hemorrhagic colitis and hemolytic-uremic syndrome in humans (Kai *et al.* 2010);
d) enteroaggregative *E. coli* (EAEC), associated with persistent diarrhea in humans (Kai *et al.* 2010);
e) diffuse-adherent *E. coli* (DAEC), associated with acute diarrhea, particularly in young children (Scaletsky *et al.* 2002);
f) enteroinvasive *E. coli* (EIEC) and *Shigella*, associated with invasive intestinal infections and dysentery in humans and various animals (Kai *et al.* 2010; Van Den Beld and Reubsaet 2012);
g) adherent-invasive *E. coli* (AIEC), associated with the chronic inflammatory bowel condition Crohn's Disease (Barnich and Darfeuille-Michaud 2007);

h) extra-intestinal pathogenic *E.* coli (ExPEC), which is comprised of the uropathogenic *E. coli* (UPEC) associated with urinary tract infections (Terlizzi *et al.* 2017), neonatal meningitic *E. coli* (NMEC) causing meningitis in newborn infants (Wijetunge *et al.* 2015), and bloodborne *E. coli* (BBEC) causing septicemia in humans and animals (Mokady *et al.* 2005).

Several virulence genes appear to be associated with each of the various *E. coli* pathotypes. Despite their supposed role in virulence, however, evidence suggests that these virulence genes have evolved and are maintained by natural selection due to the other roles they play in the ecology of commensal strains (Le Gall *et al.* 2007). For example, beyond their role in extra-intestinal virulence, genes encoding for adhesins, iron capture systems, toxins, and protectins have been implicated in the ability of commensal strains in colonizing the gastrointestinal tracts of humans (Wold *et al.* 1992; Nowrouzian *et al.* 2006), dogs (Johnson *et al.* 2008), and piglets (Schierack *et al.* 2008). Similarly, the adhesin intimin, which is associated with enteric pathogenicity, has also been shown to be essential for the colonization of the bovine rectal mucosa by bovine commensal strains (Sheng *et al.* 2006). More recently, Zhi et al. (2019) demonstrated that naturalized *E. coli* and cryptic *Escherichia* strains also possess many of these virulence determinants, suggesting they may also be functionally relevant for survival outside the host environment.

The prevalence of these virulence genes among commensal *E. coli* populations has been found to vary widely across host species. Globally, human commensal isolates are characterized by a higher prevalence of virulence genes than commensal strains derived from animal hosts. Furthermore, the prevalence of virulence genes among commensal populations also appears to increase with host body mass, likely due to an increase in host gut complexity (Escobar-Páramo *et al.* 2006). The presence of virulence genes in commensals, therefore, and their range in prevalence

across different host species could reflect the broader role of these genes in enhancing the survival and adaptation of commensal strains to their local environment. Indeed, the selective pressures in the host gastrointestinal tract may promote the emergence and maintenance of virulence factors, indicating that commensal strains may serve as reservoirs for the evolution of pathogenic *E. coli* (Tenaillon *et al.* 2010). In this context, virulence genes may be considered as adaptive genes required for the survival of commensals or pathogens in a specific host or even for the survival of strains in certain ecological niches outside of the host environment.

## 1.2 *Escherichia coli* Population Genetics

The genetic structure of a bacterial species largely depends on the balance between mutation and recombination, where a clonal structure is observed when recombination is low and a panmictic structure is observed when recombination is high (Smith *et al.* 1993). Although it is now known to occupy a broad range of niches and adopt a variety of lifestyles, early studies primarily declared *E. coli* to be a clonal species.

### 1.2.1 Early Conceptions of *E. coli* as a Clonal Species

Early genotyping analyses first suggested that the *E. coli* species consisted of several distinct clonal lineages between which little recombination occurred. Conducting the first study using multilocus enzyme electrophoresis (MLEE), Milkman (1973) observed relatively little variation in the electrophoretic mobility of 5 enzymes across different *E. coli* strains isolated from various natural sources, offering the first evidence of the species' clonal structure. In a follow-up study, Selander and Levin (1980) also utilized MLEE to examine the electrophoretic mobility of 20 enzymes from 109 strains of *E. coli*. Although twice as much diversity was observed as in the

Milkman study, it was still considerably less than expected if recombination was frequent. Subsequent studies using more extensive collections, including both environmental and host-associated strains, produced similar results, in which similar or even identical electrophoretic profiles were obtained from geographically and temporally distinct *E. coli* isolates (Ochman *et al.* 1983; Ochman and Selander 1984a). Conclusions regarding the clonal population structure of *E. coli* in turn corroborated early serotyping studies that identified the global, and non-random, distribution of certain *E. coli* serotypes (Ørskov *et al.* 1976).

Based on the results of these early MLEE studies, a standard reference collection of *E. coli* strains, referred to as the ECOR (*E. coli* reference) collection, was established to represent the full diversity of the species. Strains were selected to maximize electrophoretic diversity, geographical distribution and host range, including both pathogenic (mainly UPEC) and commensal isolates from various animal hosts (Ochman and Selander 1984b). From this MLEE data, initial phylogenetic relationships were estimated, revealing that the isolates grouped into several major clusters consisting of the major phylogenetic groups (Chaudhuri and Henderson 2012). These phylogenetic groups (A, B1, B2, D) have since been supported across studies using a variety of methods, including restriction fragment length polymorphism (RFLP) (Desjardins *et al.* 1995) and the Clermont phylogrouping method (Clermont *et al.* 2013).

### 1.2.2 Recombination in the *E. coli* Genome

Although several properties of the *E. coli* genome correlate well with the phylogenetic structure derived from the ECOR collection (Chaudhuri and Henderson 2012), enzyme electrophoresis data is ultimately limited in utility for phylogenetic analyses as enzymes can exhibit similar electrophoretic mobility without significant sequence similarity (Bisercic *et al.*

1991). In contrast, nucleotide sequences allow for more refined phylogenetic analyses since they provide detailed information for several loci at a time and are less likely to be influenced by convergence. Early studies attempting to construct phylogenetic trees using nucleotide sequence data focused on individual genes, producing widely disparate phylogenies that not only varied according to the gene sequences used, but were also discordant with the pre-existing MLEE-based ECOR phylogenetic tree. Reflecting this, phylogenetic trees produced from the sequences of the *trp*, *gnd*, and *phoA* genes were found to be widely discordant (Dykhuizen and Green 1991); however, this data also indicated that specific loci in the *E. coli* genome appeared to be under the influence of homologous recombination. Subsequent studies implicating the importance of homologous recombination led to the proposal that the *E. coli* genome largely consisted of a vertically inherited 'clonal frame' interrupted by short recombinational segments, resulting in a mosaic genomic structure (Milkman and Bridges 1990). This model suggested that only certain regions within the clonal frame were susceptible to recombination, although the level of recombination occurring in these regions would not be sufficient to blur the phylogenetic signal.

As nucleotide sequencing became routine, MLST was proposed as an alternative method to discern bacterial phylogeny using sequences from multiple housekeeping genes that were dispersed across the bacterial chromosome. As with MLEE, housekeeping genes are used as they are likely to be under strong purifying selection, and any observed variations are likely to be selectively neutral (Maiden *et al.* 1998). Early analyses with MLST data corroborated the phylogenetic relationships, with the inclusion of an additional phylogroup (group C), identified in earlier MLEE studies (Escobar-Páramo *et al.* 2004). Despite this, other studies utilizing different gene sets (i.e., MLST schemas) produced phylogenetic trees that did not provide strong support for any of the phylogenetic group arrangements previously described (Turner *et al.* 2006; Wirth *et*

*al.* 2006). This provided further evidence that recombination could be frequent in the *E. coli* genome, further challenging the proposed clonal population structure of the species. The notion that *E. coli* may not be clonal was also supported by the identification of strains that appeared to be hybrids between distinct phylogenetic groups (i.e., A•B1), or strains that had multiple sources of ancestry (Turner *et al.* 2006). Additionally, the conflicting evidence from various MLST studies illustrated an important limitation of MLST, as the resulting evolutionary relationships produced were highly dependent on the ultimately arbitrary choice of genes in the MLST schema used for phylogenetic reconstruction.

### 1.2.3 Horizontal Gene Transfer and the *E. coli* Genome

Continuing advancements in sequencing technology permitted the sequencing and comparison of whole genomes. The ability to generate complete genomic sequences provided additional insight into the processes of *E. coli* genome evolution – particularly, with regards to the importance of horizontal gene transfer. Indeed, at the genome level, the most striking source of variation between genomes appears to be in gene content, largely due to the acquisition and loss of genes. Genes acquired through horizontal gene transfer are characterized by an atypical base composition, such as a different GC content, relative to the rest of a particular genome's 'native' genes. These variations are due to the distinct selective pressures experienced by different genomes, leading to differing patterns of evolution that are then reflected in the base composition of each genome's respective genes (Chaudhuri and Henderson 2012). Importantly, these variations are best observed with recently acquired genes, as they retain the base composition of the original genome and can thus still be identified as foreign within the new genome. Otherwise, horizontally acquired genes gradually undergo the process of amelioration, where they eventually adopt the

characteristics of the recipient genome as they become subject to the same evolutionary pressures over time (Lawrence and Ochman 1997).

Initial assessments of the genomic sequence of *E. coli* K-12 MG1655 led to the estimation that 24.5% of its genes were acquired through horizontal transfer (Lawrence *et al.* 2002). The sequencing of additional genomes from other *E. coli* strains provided further evidence that a significant proportion of the genes in a given *E. coli* genome were laterally acquired. For instance, a comparison of the 5.5 Mb genome of *E. coli* O157:H7 EDL933 with the genome of *E. coli* K-12 MG1655 revealed a highly similar, shared core genomic backbone that was only 4.1 Mb in size (Hayashi *et al.* 2001; Perna *et al.* 2001). The addition of a third genomic sequence, that of the UPEC strain CFT073, revealed that only about 10% of CFT073-associated genes were represented in the O157:H7 EDL933 genome. Furthermore, a comparison of these strains with K-12 MG1655 demonstrated that only 39.2% of the combined gene set was represented across the three genomes (Welch *et al.* 2002). Indeed, these early genomic analyses provided additional support for the mosaic model of the *E. coli* genome, consisting of a shared core genome interspersed with 'islands' specific to different strains that are obtained through horizontal gene transfer (HGT).

### 1.2.4 *E. coli* Core- and Pan-Genomic Structure

As additional *E. coli* genomes were sequenced, the existence of a 'core-' and 'pan-' genome became increasingly apparent. The core genome consists of the genes and gene families that are conserved among all *E. coli* strains, whereas the collection of all genes found across all *E. coli* strains sampled comprises the pan-genome (Hendrickson 2009; Tenaillon *et al.* 2010). Early comparative genomic studies revealed the underlying core- and pan-genomic structure of the *E. coli* gene pool. For example, a comparison of the genomes of twenty different *E. coli* strains

revealed that, on average, the *E. coli* genome contains 4721 genes, with only ~2000 of these genes conserved across all strains, while the pan-genome consisted of 17838 genes (Touchon *et al.* 2009). In another study comparing 53 sequenced *E. coli* genomes, Lukjancenko *et al*. (2010) demonstrated that the core genome consisted of 1472 conserved gene families, with the pan-genome comprising 13296 gene families in total. As such, these studies indicate that roughly only 20% of a given *E. coli* genome will consist of core genes, while the remaining 80% is derived from the variable pangenome.

As additional studies continued to compare the genomes of an increasing number of *E. coli* isolates, a clear trend emerged: as the number of genomes compared grew, the number of genes in the core genome decreased, while the number of genes in the pan-genome increased. The genome of a single *E. coli* strain, therefore, can be thought to consist of a core genome that is shared between all *E. coli* strains, supplemented by an 'accessory' genome that may include subsets of genes that are specific only to select strains. Analyses of *E. coli* core and accessory genes suggest that the core and accessory-genomes play distinct roles in *E. coli* evolution. The conserved core genes are thought to encode for core metabolic processes and essential housekeeping functions in all *E. coli* strains (Rasko *et al.* 2008), thus rationalizing the early, biochemical-based classification schemes for *Enterobacteriaciae*. In contrast, the components of the accessory-genome are believed to serve a broad set of functions and new capabilities that afford the various strains of *E. coli* the ability to exploit and adapt to a wide range of niches (Lukjancenko *et al.* 2010). As such, the ever-growing pan-genome affords *E. coli* a high level of plasticity, resulting in a large diversity of adaptive paths facilitating the rapid evolution and expansion of *E. coli* into diverse host and non-host environments.

The existence of an *E. coli* core genome raises its potential utility for whole genome phylogenetic analyses (phylogenomics). As the core genome theoretically comprises all the conserved genes across the *E. coli* species, using it to assess phylogeny circumvents the issue of schema choice, which can otherwise lead to inconsistencies between phylogenies derived from MLST. Importantly, however, the validity of phylogenomics depends on the extent to which the core genome is disrupted by homologous recombination (Chaudhuri and Henderson 2012). In an early genomic study, Mau *et al.* (2006) used the genome sequences of six *E. coli* strains and one *Shigella flexneri* strain to estimate the intra-specific recombination rate of the *E. coli* species. In contrast to the larger, more easily detectable horizontal transfer events that lead to gene gain or loss, intra-specific recombination involves more subtle horizontal transfer events resulting in the substitution of alleles within conserved regions of the *E. coli* genome. The comparison of the *E. coli* and *S. flexneri* genomes suggested that this intra-specific recombination affects about 10% of the core genome, usually involving small clusters of substitutions in certain types of genes, including those typically involved in motility and various cellular processes such as DNA replication, repair, recombination, and small molecule biosynthesis (Mau *et al.* 2006).

Through examining the observed pattern of non-random allelic associations at multiple loci (i.e., linkage disequilibrium), Touchon *et al* (2009) suggested that the rate of recombination was more than twice as high as the mutation rate for short sequences of 50 bp. Given this, a single base has a 100-fold higher chance of being involved in recombination than mutation. This finding provided strong evidence for the important role of recombination in the evolution of the *E. coli* genome. While it was expected that this level of recombination would be incompatible with the phylogenetic relationships established previously through traditional phylogenetic methods (i.e., MLEE and MLST), simulation experiments demonstrated that the phylogenetic signal was robust

and remained unaffected. Indeed, whole-genome-based phylogenies were found to cluster genomes into the same phylogenetic groups identified earlier using MLEE and MLST (Chaudhuri and Henderson 2012).

Thus, while recombination does appear to occur at a greater rate than mutation, evidence suggests that the mode and scope of recombination is compatible with an apparent clonal population ancestry in *E. coli*. This allows for the construction of a phylogeny that appropriately reflects the relationships between different *E. coli* strains, especially as whole genome sequence-based methods, such as multilocus sequence analysis (MLSA), became the gold standard for phylogenetic analyses (Glaeser and Kämpfer 2015).

### 1.2.5 The Correlation Between *E. coli* Phylogenetics and Phenotype

In line with its clonal population structure, and reflected in its whole genome-based phylogeny, there is extensive substructuring within the *E. coli* species, allowing for the assignment of individual strains to one of several major phylogenetic groups (Figure ***1-1.*** **The taxonomic structure of the Escherichia genus and the type species E. coli.** The growth and development of genotypic diagnostic methods from biochemical methods has greatly informed the current taxonomic status of the Escherichia genus, as early-identified species such as L. adecarboxylata, S. blattae, P. vulneris, and A. hermannii have been reassigned to other genera, while new lineages including the cryptic Escherichia clades have been discovered. While biochemical and phylogenetic evidence continue to place E. coli as the representative species within the genus, the great genetic and biological diversity coupled with the extensive host- and niche-specificity of the model bacterium suggests that a reconsideration of current taxonomy is warranted, as E. coli may consist of a complex of several distinct host- and niche-specific ecotypes. 1-1). Currently, *E. coli*

strains can be classified according to eight phylogroups; seven (A, B1, B2, C, D, E, F) belong to *E. coli sensu stricto*, whereas the eighth consists of *Escherichia* cryptic clade I (Clermont *et al.* 2013; Luo *et al.* 2011). Although *E. coli* strains themselves are not assigned to cryptic clade I, the extensive recombination observed between cryptic clade I *Escherichia* strains and other *E. coli* strains assigned to the major phylogenetic groups supports its inclusion as a divergent *E. coli* phylogroup (Walk *et al.* 2009). Recently, additional phylogroups, G and H, have also been identified from whole genome sequence analyses, appearing to have diverged earlier than the other, more recent phylogroups (Clermont *et al.* 2019; Lu *et al.* 2016).

The diversification of *E. coli* strains into different phylogroups appears to reflect the diversity in lifestyle patterns within the species. The different phylogroups, for example, appear to differ with regards to their patterns of pathogenesis. Out of the main phylogroups, including A, B1, B2, and D, *E. coli* strains belonging to phylogroups B2 and D are more likely to cause extraintestinal infections, and possess the corresponding virulence genes, than groups A and B1 (Johnson and Stell 2000; Picard *et al.* 1999). Furthermore, between the extraintestinal-associated phylogroups B2 and D, phylogroup B2 was positively associated with the UPEC pathotype, whereas phylogroup D was more closely correlated with the other ExPEC pathotypes (Hutton *et al.* 2018). In contrast, the more severe enteric pathotypes, including EHEC, ETEC, and EIEC, were restricted to phylogroups A, B1, and E, whereas milder enteric pathotypes, including EAEC and DAEC, were dispersed across all *E. coli* phylogroups (Escobar-Páramo *et al.* 2004). This suggests that while all *E. coli* strains possess the capacity to express virulence factors associated with milder intestinal manifestations, the genetic background of strains in phylogroups A, B1, and E appear to be necessary for the expression of virulence factors related to the more severe enteric pathologies.

The different phylogroups also appear to vary in their capacity for antibiotic resistance. For instance, strains in phylogroup A (Mammeri *et al.* 2009) and phylogroup D (Deschamps *et al.* 2009) appear to be particularly amenable to the development of resistance to third-generation cephalosporins, particularly though a single nucleotide polymorphism (SNP) in the *ampC* $\beta$-lactamase gene and the acquisition of CTX-M plasmids encoding class A extended-spectrum beta lactamase (ESBL) enzymes that confer resistance to a wide range of beta-lactam antibiotics. In contrast, B2 strains appear to be less resistant to antibiotics, regardless of the resistance mechanism involved (Johnson *et al.* 1994; Ochman and Selander 1984a). Interestingly, the recently identified phylogroup G, which consists of a poultry-associated *E. coli* lineage that can cause extra-intestinal disease in humans, has been found to contain a sequence type complex, STc117, that also exhibits multi-drug resistance through a tendency to acquire and produce CTX-M class beta-lactamases (Clermont *et al.* 2019).

## 1.3 Host and Environmental Niche Specialization in *Escherichia coli*

The subdivision of strains into separate phylogenetic groups thus reflects the high plasticity of the *E. coli* genome, allowing for a large diversity in the adaptive paths throughout *E. coli* evolution. The concept of the pan-genome suggests that certain genes or genomic islands will only be found in a subset of strains, and select combinations of these genetic elements will favour the survival and adaptation of these strains in specific environments. *E. coli* strains have been isolated from a variety of host animal species, including humans, domestic and wild mammals, birds, and reptiles. Presumably, each host species will constitute a distinct environment that an *E. coli* strain must adapt to such that it can competitively establish a niche amongst the competing microbiome. Considering the myriad of selective pressures that can be exerted by a given host, including pH,

temperature, available nutrients, and the competing microbiome, different strains of *E. coli* will likely evolve to adapt to distinct gut environments (i.e., and toward host-specificity). Thus, if *E. coli* can be clustered according to specific patterns of genotypic and phenotypic characteristics (i.e., capacity for antibiotic resistance and patterns of pathogenesis), there may also be a specific range in the host species that different *E. coli* strains (i.e., ecotypes) can associate with.

### 1.3.1 Specialization and Generalization as Strategies for Microbial-Host Association

Host-specificity is defined as the partiality to colonize one, or a defined group, of hosts (Kirzinger and Stavrinides 2012). In host-microbe interactions, generalists are characterized by a comparatively wide host range as they are capable of interacting with and colonizing multiple host species. In contrast, specialists are characterized by a restricted host range, in which they can establish an intimate relationship with only a single host-species at a time (Bäumler and Fang 2013; Leggett *et al.* 2013; Pfennig 2001).

Generalization and specialization represent two opposing strategies a microbial species may utilize to maximize their evolutionary success. As different host gastrointestinal tracts effectively represent different environmental niches, the fitness of a microbe can vary across host-species. Microbes can deal with this problem of differential fitness in one of two ways. Generalists will express alternative survival and colonization strategies as needed depending on the present host-species, maximizing their evolutionary success by exploiting as many host niches as possible. In contrast, specialists evolve highly refined survival strategies to adapt to the gastrointestinal environment of a specific host, thus maximizing their ability to exploit the limited resources and their replication potential over less-adapted conspecific strains within a host gastrointestinal system (Bäumler and Fang 2013; Pfennig 2001). Although both host-generalization and host-

specialization can enhance a microbial species' evolutionary success, each strategy has trade-offs with associated fitness costs. Generalists, for instance, can interact with a wider range of hosts, but will be less able to exploit any one of their given hosts efficiently compared to resident specialists. In contrast, specialists will be better adapted to their particular host-species; however, they will be less likely to modify their phenotypic expression in changing contexts, such that their fate becomes inextricably linked with the survival of their host (Brown *et al.* 2013; Leggett *et al.* 2013).

Classically, the evolutionary advantages of host-specialization have been argued to be outweighed by the advantages of host-generalization (Woolhouse *et al.* 2001), largely due to the wide prevalence of pathogens that are capable of infecting and transmitting between multiple host-species. Sixty percent of described human pathogen species, for instance, are zoonotic, while 80% of pathogens of domestic animals adopt multi-host lifestyles (Taylor *et al.* 2001; Cleaveland *et al.* 2001). Despite this, many zoonotic pathogens have been shown to genetically sub-structure into specialist groups. For instance, while *E. coli* O157 is commonly considered a zoonotic pathogen, isolates cluster into two genetic lineages: strains that only colonize animals (i.e., cattle-specialists) and strains that can colonize both cattle and humans as generalists (Eppinger *et al.* 2011; Kim *et al.* 1999; Yang *et al.* 2004). As such, arguments have shifted to suggest that evolution instead favors host-specialization, either due to the functional trade-offs that limit the fitness of generalists in any one of their potential niches (i.e., antagonistic pleiotropy) or because host-specialization is more readily maintained due to the strong diversifying selection pressures existing across contrasting host gastrointestinal environments that lead to the more rapid evolution of specialists compared to generalists (Kassen 2002; Whitlock 1996; Woolhouse *et al.* 2001). Indeed, although it is conventionally considered to be a host-generalist, *E. coli* has been found to exhibit a high degree of host-specificity.

### 1.3.2 *Escherichia coli* Phylogroups Reflect Host-Association

The prevalence of host-specificity in the *E. coli* species is reflected in the characteristic distribution of phylogenetic groups across various host-species. In a survey of *E. coli* strains isolated from various Australian vertebrates, Gordon and Cowling (2003) demonstrated that *E. coli* strains belonging to phylogroups A and B1 were recovered from any vertebrate, and thus potentially represented 'generalist' strains, whereas isolates belonging to phylogroups B2 and D were largely restricted to endothermic vertebrates as potential 'specialists'. The distribution of *E. coli* phylogenetic groups also differs depending on the particular host-species. For humans, the main phylogroups represented appear to be group A (40.5%) and group B2 (25.5%), whereas groups B1 (41%) and group A (22%) are predominant in animals. More specifically, phylogroup B2 appears to be the predominant phylogenetic group in herbivorous and omnivorous mammals, whereas group B1 is the most prevalent in carnivorous mammals and birds (Gordon and Cowling 2003). Certain phylogenetic groups also contain *E. coli* strains that are restricted to particular host-species. For instance, haemolysin-producing B1 strains, exhibiting distinct O-antigen types and MLST profiles, have been found exclusively in animals (Escobar-Páramo *et al.* 2006), while human-specific *E. coli* strains have also been reported, including an avirulent clone in phylogroup B2 subgroup VIII ($B2_8$), and a clone in phylogroup B2 subgroup III ($B2_3$) (Carlos *et al.* 2010; Clermont *et al.* 2008).

### 1.3.3 Genetic Markers of *Escherichia coli* Host-Specificity

On the smallest scale, the host-specificity of *E. coli* may be influenced by distinct SNPs in key genes. For instance, in an analysis of *E. coli* O157:H7 isolates obtained from bovine and

human hosts, Franz *et al.* (2012) identified an informative SNP that distinguished the human clinical isolates. The SNP of interest was found in the *tir* gene, which encodes for the translocated intimin receptor, a bacterial protein which mediates the adhesion of *E. coli* O157:H7 to mammalian cells. In particular, the distribution of two alleles, *tir*(255A) and *tir*(255T), among isolates was closely associated with the host from which they were obtained. Indeed, amongst the bovine isolates *tir*(255A) and *tir*(255T) were approximately equally distributed (54.8% and 45.2%, respectively), whereas the *tir*(255T) allele was over-represented among the human clinical isolates (92.9%). Thus, the *tir*(A255T) polymorphism appeared to be a strong differentiating genetic feature for human *E. coli* O157:H7 isolates (Franz *et al.* 2012).

While individual SNPs can be predictive of host-specificity in *E. coli*, collections of SNPs across the genome appear to better reflect the host-specific lifestyles of *E. coli* strains. For instance, using a highly discriminatory SNP-based approach, Sheludchenko *et al*. (2010) identified host-informative SNP sets in an analysis of 782 *E. coli* strains isolated from various host sources. Through examining the sequences of housekeeping genes, a set of 8 highly discriminatory SNPs were identified amongst *E. coli* isolates obtained from various animal hosts including cattle, dogs, kangaroos, pigs, rabbits, and horses, as well as clinical isolates obtained from humans. Overall, 74 distinct SNP profiles were obtained, of which 8 were determined to be host-specific including 7 that were unique to human isolates and 1 that was unique to animal isolates (Sheludchenko *et al.* 2010). Taking a more novel approach, Zhi *et al.* (2015; 2016b) utilized a logic regression-based approach to identify highly specific host-informative SNP-biomarker patterns for *E. coli* isolated from human and animals hosts within select intergenic regions (ITGRs) across the genome (*csgBAC–csgDEFG*, *uspC–flhDC, asnS–ompF*, and *ydeR–yedS*). Interestingly, while host-informative SNP-biomarker patterns were identified among several ITGRs, different ITGRs

appear to be more host-informative than others depending on the host-species of interest. Specifically, ITGRs informative for human strains were predominantly flanked by genes involved in antibiotic resistance, whereas the most informative ITGRs for bovine isolates were flanked by genes involved in environmental stress responses – thus potentially reflecting the functional diversity between *E. coli* strains colonizing different host environments (Zhi *et al.* 2016b).

The host-specificity exhibited by *E. coli* may also be influenced at the gene level. In the same study that identified the discriminating *tir*(A255T) polymorphism for bovine and human O157:H7 isolates, Franz *et al.* (2012) also identified gene variants that could differentiate the two host-derived groups. In particular, variants in the *Q* gene (*q21* and *Q933*) and *stx*$_2$ gene (*stx*$_{2a}$ and *stx*$_{2c}$) were found to be highly associated with the host of origin. Indeed, bovine isolates were found to primarily harbor the *stx*$_{2c}$ gene variant (86.4%), whereas the distribution of the *stx*$_{2a}$ and *stx*$_{2c}$ were less skewed in human clinical isolates, with 36.5% possessing the *stx*$_{2c}$ variant, 22.4% possessing the *stx*$_{2a}$ variant, and 22.4% possessing both variants. Furthermore, the distribution of the *Q* gene variants correlated with the distribution of the *stx*$_2$ subtypes, with the presence of *q21* strongly correlating with *stx*$_{2c}$ and *Q933* strongly correlating with *stx*$_{2a}$. Thus, bovine isolates were similarly found to be dominated by the *q21* variant alone, whereas human isolates were typically characterized by the *Q933* variant, either alone or in conjunction with the *q21* variant (Franz *et al.* 2012).

A variety of other gene targets have been suggested to mediate host-specificity in *E. coli*, especially for select *E. coli* pathotypes. The host-specificity of ETEC, for instance, is thought to be due to the presence/absence of particular host-specific colonization factors that can interact with the gut receptors expressed by each host. Reflecting this, different host-associated ETEC strains are characterized by distinct assemblages of fimbrial adhesins and enterotoxin variants.

Concerning fimbriae, pig-specific ETEC strains are characterized by F4, F6, and F18 fimbriae, whereas cattle-specific ETEC strains possess F17 fimbriae (Dubreuil *et al.* 2016). Similarly, different host-associated ETEC strains are characterized by distinct enterotoxin variants, as human-specific ETEC strains are characterized by the LTIh variant of the heat-labile enterotoxin and STh variant of heat-stable enterotoxin A, whereas pig-specific ETEC strains are characterized by variants of heat-stable enterotoxin B (Wang *et al.* 2019). The host-specificity of ExPEC strains similarly appears to be mediated in part by the presence of key genes. Several adhesins of ExPEC strains, for instance, have been determined to be highly host-specific, including the AC/I pili of O78 avian pathogenic *E. coli* (APEC) strains and the K99 fimbriae of septicemic strains in lambs (Ron 2006). Indeed, using a transposon-directed insertion sequencing (TraDIS) approach to evaluate the importance of various genes for the pathogenesis of the potentially zoonotic ExPEC XM strain, distinct assemblages of genes were identified to be required for virulence in mammalian and avian models (Zhang *et al.* 2019a). While 151 genes were found to be essential for virulence in both models and could thus be thought of as 'general' virulence genes, 97 and 280 genes were specifically required for virulence in the mammalian and avian model, respectively. This suggests that, in addition to the genes required for virulence in general, host-restricted ExPEC strains also exclusively harbor genes that are specifically required for virulence in their particular host-species.

The strong correlation between particular gene targets and host origin have led to their use as genetic markers for source tracking. Indeed, some gene targets that have been implemented for source tracking have been found to be specific for strains derived from a particular host species. For instance, swine-derived *E. coli* isolates are characterized by a region of the STII variant of the ETEC heat-stable enterotoxin gene not found in strains isolated from other host-species (Khatib *et al.* 2003). Similarly, in a comparative genome analysis of 22 *E. coli* isolates, Gomi *et al.* (2014)

identified several genomic regions from the accessory genome that were host-informative for humans, cows, pigs, and chickens. Of these, 2 cow- and 3 pig-specific gene markers were identified to have no cross-reactivity with strains derived from the other host-species examined.

### 1.3.4 Host-Specificity of *Escherichia coli* Pathotypes

The host-specific lifestyles of various pathogenic strains of *E. coli* provide further evidence of *E. coli* host-specialization. For instance, EHEC/STEC isolates from bovine and human hosts appear to be genetically and ecotypically distinct, as they substructure into two lineages, including: the bovine-specific lineage II, which is not readily capable of transmitting to and causing disease in humans, and lineage I, which consists of strains that are transmissible between cattle and humans as cattle-human generalists (Eppinger *et al.* 2011; Kim *et al.* 1999; Yang *et al.* 2004).

The population genetics of EPEC also provides evidence of host-specialization in *E. coli*. For instance, after genetically characterizing 159 atypical EPEC (aEPEC) strains lacking the *E. coli* adherence factor (EAF) plasmid, Wang *et al*. (2013) demonstrated that aEPEC strains from bovine, swine, and human isolates were disproportionately represented in different phylogenetic groups. Phylogenetic analyses revealed that bovine aEPEC isolates (79%) predominantly clustered in phylogenetic group B1, swine aEPEC isolates (54%) predominantly clustered in group A, and isolates obtained from healthy humans (54%) mainly clustered in group B2. Furthermore, the pathogenesis of EPEC appears to be highly host-species specific. Localized adherence assays performed by Tobe and Sasakawa (2002), using the human pathogenic EPEC strain B171-8 on mouse intestine-derived epithelial cells (CMT-93 and Colon-26 cell lines) and human intestine-derived cells (Caco-2, T-84, and Intestine-407 cell lines), revealed that the human EPEC strain adhered to these cell types with differing efficiencies. The efficiency of B171-8 adherence to, and

subsequent infection in, the mouse-derived cell types was significantly lower (less than 5%) than in the human-derived intestinal epithelial cells, where over 49% of the human-derived cell lines were infected. The binding properties of the human-associated B171-8 was also shown to be biologically relevant for its *in vivo* behavior in non-human hosts, as this strain could not establish an infection in mice, even when introduced into a germ-free model (Tobe and Sasakawa 2002).

Another important *E. coli* pathotype, ETEC, is known to exhibit a high degree of host-specificity. Despite its prevalence in many host-species, ETEC infections are reportedly extremely rare or non-existent in a variety of other animals, such as poultry, rabbits, and horses (Dubreuil *et al.* 2016). This restriction in host-range for ETEC in some animals and not others is puzzling, as non-permissive hosts still express the appropriate ETEC fimbrial and enterotoxin receptors for ETEC colonization and pathogenesis. The high degree of host-specificity of ETEC is also suggested in the lack of suitable animal models of ETEC infection. For instance, while both mouse and rabbit models have been used to study ETEC infection, neither are naturally susceptible to ETEC, leading to limited standalone infection models that either resist colonization after ETEC inoculation (Allen *et al.* 2006) or fail to exhibit diarrheal symptoms in response to produced enterotoxins (Wenzel *et al.* 2017). For instance, the oral introduction of a human diarrheagenic strain of ETEC into a mouse model by Allen *et al.* (2006) produced very low levels of colonization without diarrheal manifestations typical of ETEC infection, suggesting that human-adapted ETEC strains lack the ability to sufficiently adhere to, and colonize, the mouse intestinal epithelium. This specificity is also observed between typical ETEC hosts, as pig models are not effectively colonized by human-adapted strains of ETEC (Wenzel *et al.* 2017). Interestingly, while ETEC exhibits a high degree of host-specificity in line with other enteric pathotypes, this specificity does not appear to be linked to the evolution of distinct host-specific ETEC lineages. Indeed, MLST-

based phylogenetic analyses of human-, bovine-, and porcine-derived ETEC strains demonstrated that both human- and animal-associated ETEC are distributed across several phylogroups and sequence types (Clermont *et al.* 2011b; Steinsland *et al.* 2010; Turner *et al.* 2006). Rather, the host-specificity of ETEC seems to be mediated by the acquisition of host-specific colonization factors through HGT.

Similar to the host-specificity exhibited by the enteric pathotypes, extraintestinal strains also display extensive host-specificity despite their genetic similarity. For instance, early genomic comparisons of a variety of *E. coli* serogroup O78 isolates obtained from various host-species revealed that human septicemic O78 strains clustered into the same clonal grouping that was assigned to *E. coli* isolates derived from poultry hosts (Chérifi *et al.* 1994). Furthermore, APEC and ExPEC strains have been shown to share a common repertoire of virulence genes, such as for iron acquisition systems, P and S fimbriae, and a K1 capsule (Mokady *et al.* 2005; Rodriguez-Siek *et al.* 2005). Reflecting this genetic similarity, APEC and human ExPEC strains display considerable overlap in the predominant serogroups represented, and they phylogenetically subcluster into the same phylogroups, including B2 (subgroup 1), B1 and D (Moulin-Schouleur *et al.* 2007). While the close similarity of human and avian ExPEC strains has led to the suggestion that they have zoonotic potential, these extraintestinal pathogenic populations still appear to exhibit considerable host-specificity. Indeed, APEC strains isolated from cases of avian septicemia have been found to be more virulent in chicks than strains isolated from cases of newborn meningitis, and are also more virulent in birds than in mice (Ron 2006). Similarly, although Zhang *et al.* (2019a) characterized ExPEC XM as a potential zoonotic strain based on its virulence in both mammalian and avian models, other ExPEC strains examined were found to exhibit some degree of host-preference. Indeed, upon inoculation into mammalian and avian models, the human ExPEC

strain RS218 was significantly more virulent in the mammalian model while the APEC strain DE471 was more virulent in the avian model.

### 1.3.5 Beyond the Host: Niche-Adaptation of *Escherichia coli* in Non-Host Environments

Although the vertebrate gastrointestinal tract is considered the primary niche of *E. coli*, up to half of the total *E. coli* population has been estimated to survive, and in some cases even grow, in a variety of natural environments (Tenaillon *et al.* 2010). While the presence of *E. coli* in natural sources (i.e., water) was originally thought to be caused by faecal contamination within the environment, growing evidence points to the existence of distinct 'naturalized' populations of *E. coli*, separate from the cryptic *Escherichia* clades, that appear to have diverged from their host-associated counterparts to preferentially survive and persist outside of the host environment (Jang *et al.* 2017). Reflecting this, distinct naturalized *E. coli* populations have been described in a variety of natural environments, including subtropical and temperate soils (Ishii *et al.* 2006; Ishii *et al.* 2010; Byappanahalli *et al.* 2012), surface water and sediments (Jang *et al.* 2011; Jang *et al.* 2015), and estuary water environments (Berthe *et al.* 2013).

While it is possible that environmental *E. coli* strains could be host-derived, several studies have been able to readily characterize these naturalized subpopulations as being genotypically, phenotypically, and ultimately ecotypically distinct from their enteric counterparts. For instance, when compared to host-derived strains, naturalized *E. coli* isolated from soil (Ishii *et al.* 2006; Ishii *et al.* 2010; Byappanahalli *et al.* 2012) and river water (Jang *et al.* 2011; Jang *et al.* 2015) were repeatedly found to harbor unique DNA fingerprint patterns generated with horizontal fluorophore-enhanced rep-PCR. Similarly, using an accessory gene fingerprinting approach,

Tymensen *et al.* (2015) were able to differentiate naturalized surface water isolates from enteric strains based on the presence of 'accessory gene fingerprints' composed of characteristic combinations of iron acquisition, complement resistance, and biofilm formation genes that would presumably enhance survival outside of the host environment. Interestingly, these naturalized populations have also been found to possess various phenotypic adaptations that have likely facilitated their evolution towards external, non-host environments as a primary niche. For instance, naturalized *E. coli* strains isolated from temperate and subtropical soils in the United States were found to exhibit longer survival times at lower temperatures ranging between 4–25°C compared to the typical 'optimal' temperature range of 30–37°C, suggesting these strains have adapted to the lower, ambient temperatures of non-host environments. (Ishii *et al.* 2006; Ishii *et al.* 2010). Furthermore, naturalized *E. coli* strains linked to coliform blooms across several geographically separated lakes in Australia were characterized by the production of a group 1 capsule, which likely enhanced their survival against environmental stressors such as ultraviolet (UV) radiation and desiccation (Power *et al.* 2005)

### 1.3.6 Into New Niches: Naturalization of *Escherichia coli* in Non-Host, Engineered Environments

While the proliferation of *E. coli* across various host-species and natural environments has been well documented (Jang *et al.* 2017; Tenaillon *et al.* 2010), relatively recent evidence suggests that the niche range of *E. coli* may extend to include various man-made, built environments. For instance, in a survey of 28 *E. coli* isolates collected from samples of treated (i.e., chlorinated) drinking water in Australia, Blyton and Gordon (2017) identified 9 that appeared to be water-associated, free-living isolates. These strains were characterized by a dearth of antibiotic resistance

and virulence genes, and thus did not appear to be host-associating. Similarly, Yang *et al.* (2021) characterized a set of *E. coli* isolates collected from meat processing plants that appeared to have adapted to a host-independent lifestyle. Indeed, compared to their enteric counterparts, these naturalized meat plant strains were found to be enriched in stress resistance genes, including a genetic element known as the locus of heat resistance (LHR) (also known as the transmissible locus of stress tolerance [tLST]), while simultaneously lacking various genes related to epithelial attachment and virulence that would be required for host colonization.

The prospect that distinct *E. coli* populations could evolve to exploit non-host, non-natural environments may be best exemplified by a series of landmark studies conducted by Zhi *et al.* (2016a; 2017; 2019) that describe distinct strains of *E. coli* that appear to have evolved to exploit wastewater as a primary niche. Interestingly, these wastewater strains were found to be genotypically, phenotypically, and ecotypically distinct, potentially representing a novel, wastewater-specific *E. coli* ecotype (Wang *et al.* 2020; Zhi *et al.* 2016a; Zhi *et al.* 2017; Zhi *et al.* 2019). Using a supervised, logic regression-based approach to analyze the SNP variation in various *E. coli* intergenic regions, Zhi *et al.* (2016a) first identified novel strains of *E. coli* that appeared to be highly adapted to wastewater/sewage environments as a primary niche. Genetically, the naturalized wastewater strains carried distinct SNP ITGR biomarkers, identified using logic regression, that were not found in any enteric or cryptic strains evaluated (Zhi *et al.* 2016a; Zhi *et al.* 2019). Interestingly, a proportion of these wastewater strains were also found to harbor a unique insertion element, IS30, located specifically within the *uspC–flhDC* intergenic locus, that was also found to be absent from their enteric and cryptic counterparts. Confirming the distinct genetic nature of the wastewater strains, phylogenomic analyses grouped the wastewater strains (i.e., those that carried the unique SNP biomarker and the *uspC*–IS30–*flhDC* marker) into a separate cluster

distinct from the rest of the *E. coli* species. Subsequent phenotypic characterization revealed several remarkable adaptations in these strains to survive the challenging conditions of a wastewater treatment plant. For example, compared to their enteric counterparts, these wastewater strains were found to display enhanced resistance to chlorine and other oxidizing agents (Wang *et al.* 2020). The strains were also shown to be highly adept at forming biofilms compared to human enteric strains (Zhi *et al.* 2017) and displayed an extreme heat-resistance phenotype (Zhi et al. 2019; Wang *et al.* 2020). Interestingly, these wastewater strains were found to harbor the tLST, encoding various stress resistance genes including chaperones, DNA repair proteins, and heat shock proteins (Mercer *et al.* 2015; Mercer *et al.* 2017) that confer cross-resistance between chlorine, advanced oxidants, and heat – stressors that could be encountered during the wastewater treatment train (Wang *et al.* 2020). In addition to the tLST, the genomes of these naturalized wastewater strains were further characterized by an overabundance of stress-resistance genes compared to their enteric counterparts, likely intended to enhance their survival against the stressors encountered during wastewater treatment (i.e., chlorination, UV irradiation, oxygenation, high temperatures [biosolids composting], microbial competition, predation, etc.), that were seemingly acquired through HGT (Zhi *et al.* 2019).

These naturalized wastewater strains have been isolated from wastewater treatment plants across Canada, the U.S., and Switzerland (Zhi *et al.* 2019), suggesting that this naturalized, wastewater-specific ecotypic group has become globally disseminated. Interestingly, despite their global distribution, comparative genomics revealed that the naturalized wastewater strains exhibited significantly higher within-group genomic similarity (~96%) compared to their enteric (~82%) and cryptic (~62%) counterparts (Zhi *et al.* 2019). Indeed, the collective genetic and phenotypic evidence point towards a distinct *E. coli* ecotype that appears to have evolved to

abandon the host gastrointestinal tract in favor of specifically inhabiting sewage and treated wastewater matrices as its primary niche (Figure 1-2).

## 1.4 A Polyphasic Perspective on *E. coli* Diversity

The prospect that there exists a wide array of distinct *E. coli* ecotypes that each occupy a distinct ecological niche, either through colonizing different host-species or exploiting various natural or man-made environments, has important implications for understanding *E. coli* as a species. This is especially important when taking a polyphasic taxonomic approach, which incorporates multiple perspectives, including genotypic, phenotypic, and ecotypic considerations, to obtain a more holistic conception of prokaryotic diversity (Kämpfer 2014; Kämpfer and Glaeser 2012). Currently, the designation of *E. coli* as a single cohesive species is largely supported by biochemical and phylogenetic evidence; however, *E. coli* is incredibly diverse, with strains exhibiting a wide range of genetic diversity and phenotypic plasticity, which in turn appears to be reflected in a high degree of host- and niche-specificity within the species. Indeed, considering that only roughly 20% of any given genome is core to the *E. coli* species, the incorporation of additional criteria beyond just genotypic information may provide a more accurate description of *E. coli* taxonomy.

Current taxonomic schemes place a heavy emphasis on genotypic, DNA sequence-based criteria; however, this appears to have come at the expense of phenotype and ecotype, even though genotypic information is only made biologically and evolutionarily relevant through its phenotypic and ecotypic expression (Kämpfer 2014). This emphasis on genetic-based criteria has thus been criticized to be too restrictive, as it incorporates little phenotypic or ecotypic information, presents a limited and arbitrary perspective on prokaryotic taxonomy, and fails to consider the underlying

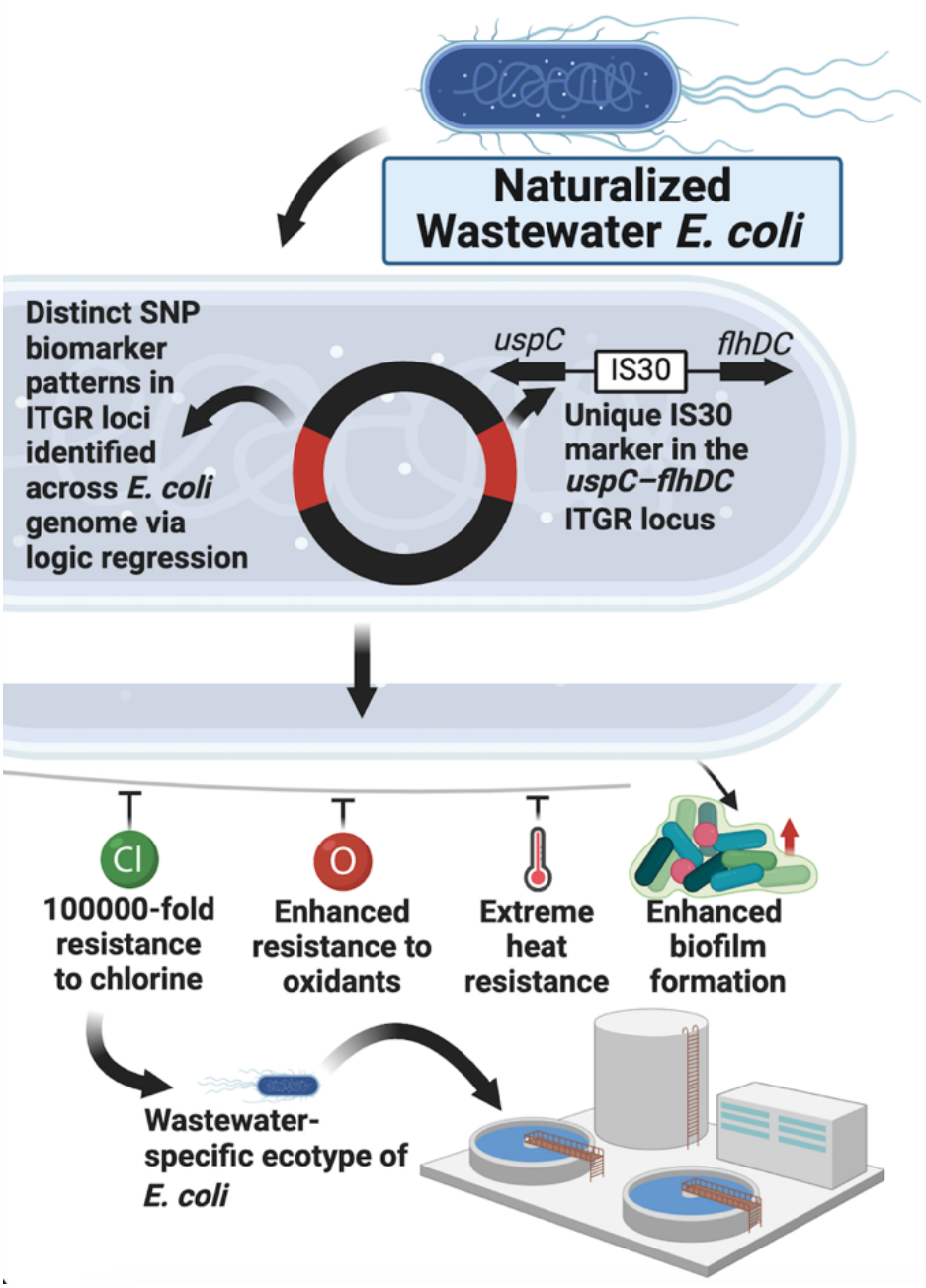



**Figure 1-2. Genotypic, phenotypic, and ecotypic insights into *E. coli* niche- and host-specificity.** Strains of *E. coli* isolated from wastewater matrices harbored unique SNP biomarkers in the *csgBAC–csgDEFG* and *asnS–ompF* ITGR loci, as well as the insertion element IS30 within the *uspC–flhDC* ITGR locus, not found in enteric and cryptic strains. These genotypic markers were demonstrated to correlate with a distinct wastewater-specific ecotype of *E. coli* that exhibited a variety of phenotypic adaptations (including extreme heat, chlorine and oxidant resistance as well as enhanced biofilm formation capacity) that presumably allowed these wastewater *E. coli* to survive the harsh conditions (chlorine treatment, high temperature, predation, etc.) of the wastewater treatment process.

processes of microbial speciation (Georgiades and Raoult 2011; Van Ingen *et al.* 2018). These limitations become especially apparent in *E. coli* taxonomy, particularly when considering the significant genetic variability that can exist between two different strains. While *E. coli* is designated as a single cohesive species, it is characterized by a large, open pan-genome – a characteristic observed in species complexes (Dillon *et al.* 2019; Zhou *et al.* 2020). Furthermore, the *E. coli* core genome has been estimated to constitute only 20% of the genetic background of a single strain, which is comparable to the *B. cepacia* complex [15.33–25.20%] (Zhou *et al.* 2020) and more diverse than the *P. syringae* complex [37.69%–50.70%] (Dillon *et al.* 2019).

Based on genetic metrics alone, therefore, *E. coli* may be more appropriately described as a species-complex comprised of a wide array of host- and niche-specific ecotypes. A significant degree of genetic diversity exists within the *E. coli* species, which could reflect the divergence between different ecotypic groups and their adaptation and subsequent specialization to their respective hosts or environmental niches. As the genetic discrepancy between different ecotypic groups increases, they become less likely to co-exist within the same niche and the level of genetic trafficking that can occur will be effectively reduced. This will result in the emergence of several isolated host- and niche-specific *E. coli* ecotypes that retain genetic coherence (Dillon *et al.* 2019) due to the *E. coli* 'generalists' that can still transit across multiple environments and mediate a residual level of genetic exchange between the specialized ecotypes that inhabit them.

The incorporation of ecotype in *E. coli* taxonomy, therefore, may better reflect the processes underlying the evolution of *E. coli* towards niche-specificity. Ecotypic information is already incorporated in other taxonomic schemes, such as for parasite classification (Šlapeta 2013), suggesting that criteria such as niche-range and niche-specificity may lead to a more holistic understanding of prokaryotic diversity. Interestingly, certain bacterial classification schemes also

utilise ecotypic information for taxonomic purposes. The *Borrelia burgdorferi* sensu lato complex (Rudenko *et al.* 2011), for instance, incorporates information including vector and host range when distinguishing the sub-groups that comprise the complex. Indeed, the vast ecological range and significant host- and niche-specificity exhibited by *E. coli*, which appears to be comparable to that observed for the *B. burgdorferi* complex (Rudenko *et al.* 2011), provides additional support for the designation of *E. coli* a 'species-complex'.

## 1.5 Research Rationale and Hypotheses

Although there is a growing recognition that the niche-specificity of *E. coli* may be more appropriately represented using a 'species-complex' model (Georgiades and Raoult 2011; Guertsen *et al.* 2022), there is currently no systematic workflow to discover and characterize putative *E. coli* ecotypes. Indeed, a wide variety of genotypic and DNA fingerprinting approaches have already been developed for source attribution and classification purposes; however, these methods are often labour intensive and dependent on the construction of representative reference libraries, making them unsuitable for the discovery of novel ecotypes – especially through high-throughput analyses using large volumes of readily-available genomic data. Similarly, while various machine learning algorithms have been developed and even applied for source attribution purposes (Lupolova *et al.* 2019), these methods are often considered 'black-box' methods as they generate complex results that can be difficult to interpret.

One potential approach to address these issues is to build a workflow around logic regression (Ruczinski *et al.* 2003) as an ecotype discovery tool. As a supervised machine learning method, logic regression can be scaled up to analyse large volumes of genomic sequence data to identify putative *E. coli* ecotypes for subsequent analysis. Furthermore, unlike other computational

approaches, logic regression also has the advantage of generating interpretable, biologically plausible results. Indeed, logic regression has already been used extensively for studying *E. coli* host- and niche-specificity, particularly through the discovery of source-specific SNP biomarkers within intergenic regions (i.e., the 'regulome') that appear to reflect the ability of a given strain to sense, respond, and adapt to the prevailing conditions of a given niche (Zhi *et al.* 2015; Zhi *et al.* 2016a; Zhi *et al.* 2016b).

Building from logic regression, any putative ecotypes identified must then be characterized and differentiated from other ecotypic groups within the *E. coli* species. Considering that ecotypes have been proposed to represent the fundamental units of bacterial diversity (Cohan and Perry 2007; Koeppel *et al.* 2008), various taxonomic approaches can be used to establish the characteristics defining the taxonomic boundaries of an ecotype. Instead of relying primarily on genotypic (i.e., sequence identity) or phenotypic (i.e., morphology, biochemical profiling) measures alone, which would reflect more conventional approaches to bacterial taxonomy, multiple lines of evidence can be integrated to better understand *E. coli* niche-specificity. Indeed, through a polyphasic approach, genotypic, phenotypic and ecotypic evidence can be integrated to gain a more comprehensive understanding of the taxonomic boundaries of different *E. coli* ecotypes.

## 1.6 Thesis Objectives and Overview

Although a growing body of evidence indicates that *E. coli* may exhibit a high degree of niche-specificity, the processes underlying the emergence of *E. coli* ecotypes and their evolution towards niche-specificity are unclear. Thus, the central aim of this thesis was to establish a workflow for evaluating *E. coli* niche-specificity, using distinct naturalized *E. coli* strains that have

emerged across food- and water-associated engineered environments, with a particular focus on wastewater-specific strains as a model system. Importantly, a polyphasic approach was used to characterize these strains, in order to gain a more holistic understanding of the specific adaptive mechanisms driving their evolution toward their respective niches. The overall objectives of this thesis include the following:

1) Explore the use of logic regression as a novel approach for the identification of putative *E. coli* ecotypes based on ITGR-encoded SNP-SNP biomarker discovery, and its application for source attribution efforts;
2) Demonstrate how genotype, phenotype, and ecotype recapitulate the evolution of niche-specificity in *E. coli*, through the polyphasic characterization of naturalized *E. coli* ecotypes that have emerged across food- and water-associated engineered environments, with a specific focus on wastewater-specific strains; and
3) Evaluate how the processes of niche-adaptation can lead to phenotypic convergence between different ecotypic groups, through the exploration of wastewater treatment resistance in naturalized, wastewater-specific *E. coli* and wastewater-borne ExPECs.

### 1.6.1 Objective 1: Development of Logic Regression as an Ecotype Discovery and Attribution Tool

Objective 1 is addressed in Chapter Two of this thesis, which aims to evaluate logic regression as a potential ecotype discovery and attribution tool. This chapter will build on previous work demonstrating the utility of logic regression for assessing host- and niche-specificity within the *E. coli* species (Zhi *et al.* 2015; Zhi *et al.* 2016a; Zhi *et al.* 2016b), with a focus on expanding the geographical and niche diversity of strains within our local *E. coli* genome sequence repository and the range of intergenic targets evaluated (i.e., beyond *uspC–flhDC*, *asnS–ompF*, *csgDEFG–*

*csgBAC,* and *yedS–yedR*). Extending its use for practical applications, we also evaluate the utility of logic regression for ecotype attribution purposes, through the host classification of environmental *E. coli* isolates collected from rivers in the Jämtland region of Northwestern Sweden.

### 1.6.2 Objective 2: Polyphasic Characterization of Naturalized *E. coli* Ecotypes Emerging Across Food- and Water-Associated Engineered Environments

Objective 2 will be jointly addressed through Chapters Three to Five, which aims to apply the polyphasic approach to characterize and describe novel naturalized *E. coli* ecotypes that appear to have recently emerged within food- and water-associated engineered environments as primary niches. Specifically, Chapter Three will focus on utilizing a comparative genomics approach to characterize a set of naturalized *E. coli* strains collected from meat processing facilities and wastewater treatment plants. Average nucleotide identity (ANI) analyses were used to first assess the genomic similarity of the naturalized-engineered strains compared to other ecotypic groups (i.e., enteric, ExPEC, environment, etc.) within the *E. coli* species. Following this, phylogenetic, typing, and ecotype prediction analyses were performed to assess whether the naturalized-engineered strains could represent distinct ecotypes adapted to 'engineered' niches. Additionally, the wastewater and meat plant strains were also screened for additional insertion element markers that could be ecotype-informative, similar to the *uspC*–IS30–*flhDC* locus.

Building on Chapter Three, Chapter Four examines the specific genetic adaptations that could underlie the evolutionary success of the wastewater and meat plant strains within their respective engineered niches. Specifically, pan-genome-wide association studies were performed to evaluate whether the wastewater and meat plant strains could be harboring unique genetic

repertoires that could reflect their naturalization within, and potential niche-specificity to, their respective engineered environments. Additionally, gene-gene interaction and gene-localization analyses were also conducted to determine the specific pan-genome dynamics associated with the genetic adaptations of the wastewater and meat plant strains.

Chapter Five supplements the pan-genomic evidence presented in Chapter Four by connecting our genotypic findings with discernable phenotypes, with a specific focus on the naturalized wastewater strains. Previous work explored some of the phenotypic properties of these naturalized *E. coli* populations, including biofilm formation capacity (Zhi *et al.* 2017), and resistance to sanitizers (Yang *et al.* 2021), heat (Yang *et al.* 2021; Wang *et al.* 2020), chlorine (Wang *et al.* 2020) and advanced oxidants (Wang *et al.* 2020); however, these studies primarily focused on characterizing the phenotypic responses of the naturalized strains to disinfection-related stressors. Building on this work, this chapter will continue to explore the phenotypic responses of the wastewater strains to other conditions reflective of the wastewater treatment train, including heat resistance (i.e., sludge digestion and biosolids composting) as well as growth and biofilm formation capacity in lower temperature, lower nutrient conditions.

### 1.6.3 Objective 3: Emergence of Wastewater Treatment Resistance in ExPECs

Objective 3 is addressed in Chapter Six, which aims to assess whether the conditions of the wastewater treatment plant could drive the evolution of other *E. coli* ecotypic groups to acquire wastewater treatment resistance. While the preceding four chapters of this thesis are focused on characterizing distinct wastewater-specific *E. coli* strains that appear to possess the genotypic and phenotypic adaptations necessary to survive wastewater treatment, their existence raises the prospect that pathogenic *E. coli* populations could also evolve resistance to wastewater treatment and sanitation. Using a comprehensive, comparative genomics approach, Chapter Six evaluates

whether clinically relevant ExPEC strains could also be differentially surviving wastewater treatment, raising the concerning public health prospect that wastewater treatment resistance may be an emerging phenotype within the microbial world.

# Chapter Two: Assessment of Logic Regression as an Exploratory Tool for the Identification of Putative Host- and Niche-Specific *Escherichia coli* Ecotypes and its Application as a Novel Microbial Source Tracking Approach[2]

## 2.1 Introduction

*Escherichia coli* is one of the most well-recognized organisms in microbiology, owing in part to its wide prevalence across different host species and environments. The sheer ubiquity of this microbe underscores its designation as a host- and niche-generalist capable of colonizing and transmitting between its various niches (Tenaillon *et al.* 2010; Jang *et al.* 2017); however, a growing body of evidence suggests that *E. coli* exhibits a significant degree of host- and niche-specificity (Nandakafle *et al.* 2021), such that the species may be more appropriately understood as a 'complex' of distinct niche-specific groups known as 'ecotypes' (Cohan and Kopac 2011). The prospect that the *E. coli* species has diversified into several groups of host- and niche-restricted strains has important implications for the field of environmental microbiology, especially given its use as a common indicator of faecal contamination within the environment (Devane *et al.* 2020). In particular, the concept of *E. coli* niche-specificity, and the existence of niche-specific *E. coli* ecotypes, has direct implications for the practice of microbial source tracking (MST).

[2] A version of this chapter has been accepted for publication as: Yu, D., Andersson-Li, M., Maes, S., Andersson-Li, L., Neumann, N.F., Odlare, M., and Jonsson, A. 2024. Development of a logic regression-based approach for the discovery of host- and niche-informative biomarkers in *Escherichia coli* and their application for microbial source tracking. Appl. Environ. Microbiol.. doi: 10.1128/aem.00227-24

Microbial source tracking hinges on the assumption that different subgroups of microorganisms will become better adapted to, and thus be dominant in, a particular host environment over time. This will eventually lead to the acquisition of distinct genotypic attributes (i.e., genes, DNA sequence polymorphisms, etc.) that can be used to differentiate between different host-adapted strains (Harwood *et al.* 2014; Scott *et al.* 2002). Reflecting the evolution of host- and niche-specificity within the species, various niche-specific *E. coli* genetic features have been discovered to date (Gomi *et al.* 2014; Khatib *et al.* 2002; Khatib *et al.* 2003; Tiwari *et al.* 2023) including several that appear suitable as potential source tracking targets (Paruch and Paruch 2022). Importantly, beyond their direct application as source tracking markers, the apparent association between these genetic features and different niche sources (i.e., host species, non-host environments, etc.) suggests that they may also represent key genetic markers that can be used to identify putative *E. coli* ecotypes. Unfortunately, there does not yet appear to be a standardized workflow for the identification and validation of putative *E. coli* ecotypes, and reliable approaches for the discovery of ecotype-informative genetic markers remain elusive.

One possible strategy could be to leverage the microbial source tracking 'toolbox' for a method that can be used to identify genetic signatures indicative of putative *E. coli* ecotypes. These 'source attribution' approaches generally aim to uncover ecotype-informative genetic markers in either a library-dependent (i.e., genotypic fingerprinting) or library-independent (i.e., niche-specific PCR markers) manner. While various library-dependent and library-independent approaches are available, many are considered to be labour- and time-intensive (Foley 2009; Fu and Li 2014). Additionally, conventional source attribution methods have been found to vary widely in performance (i.e., ability to correctly classify microorganisms) and reproducibility (Fry

*et al.* 2009; Fu and Li 2014). Most laboratory-based source attribution approaches, therefore, are ultimately limited in utility for large-scale, ecotype discovery analyses.

Alternatively, computational approaches may be more suitable for the identification of putative *E. coli* ecotypes. In particular, the use of machine learning could be leveraged for the discovery of ecotype-predictive markers from large volumes of readily available *E. coli* genome sequence data. Briefly, machine learning involves the use of algorithms for uncovering underlying patterns in large volumes of data (Goodswen *et al*. 2021). These algorithms typically fall under one of two categories: unsupervised learning and supervised learning. Unsupervised machine learning methods, such as k-means clustering, hierarchical clustering, and various dimensionality reduction procedures, are exploratory in nature and only seek to cluster data without consideration for any pre-existing labels (i.e., host or niche sources, ecotypes) that the data may have (Goodswen *et al*. 2021; Lupolova *et al*. 2019). In contrast, supervised machine learning approaches, which include clustering and regression algorithms such as logistic regression, support vector machines, random forests, and neural networks, are first 'trained' on an initial set of data such that future predictions or classifications can be made with new data (Goodswen *et al*. 2021). Supervised methods, therefore, attempt to uncover patterns in data correlating specifically with observed data labels (i.e., host or niche sources), making them more appropriate for identifying genetic markers that are source-informative and thus ecotype-predictive.

Several studies have explored the use of supervised machine learning methods for the analysis of bacterial genome sequence data for host source-attribution purposes. Zhang *et al.* (2019b), for example, used a random forest approach to identify 50 key genetic determinants that could reliably predict the original livestock source of outbreak-associated *Salmonella* Typhimurium isolates. From the host-informative markers identified, which included 10 core

genome single nucleotide polymorphisms (SNPs) and 40 accessory genes related to virulence, metal resistance and colonization, the authors were able to correctly identify the original source of 7 of 8 major *Salmonella* zoonotic outbreaks in the United States from 1998 to 2013. Opting for a different machine learning algorithm, Lupolova *et al.* (2017) developed support vector machine classifiers to predict the host-specificity of *Salmonella enterica* and *Escherichia coli* isolates. According to the distribution of predicted protein variants, the authors were able to correctly classify the host source of 67–90% of *S. enterica* isolates collected from avian, bovine, human, and swine hosts, as well as 83% of human and bovine-derived *E. coli* O157 strains. As a follow up, Lupolova *et al.* (2019) also developed random forest and neural network classifiers that could correctly classify roughly 80% of *S.* Typhimurium isolates collected from avian, bovine, human, and swine hosts.

Supervised learning methods thus appear to be particularly useful for the discovery of niche-informative genetic features that can be correlated with distinct microbial ecotypic groups. Despite this, approaches such as support vector machines, random forests, and neural networks are often considered 'black box' algorithms as they require specialist knowledge to perform and typically generate highly complex results that are difficult to interpret (Lupolova *et al*. 2019). Furthermore, when used for biomarker discovery purposes, these methods often generate large numbers of potential genetic features that could possibly, but not necessarily, reflect niche-specificity, which makes it difficult to pinpoint the specific determinants that are ecotype-informative. Instead, alternative supervised learning approaches may be better suited for the discovery of reliable, biologically-plausible biomarkers that are predictive of novel *E. coli* ecotypes. For instance, a series of studies conducted by Zhi *et al.* (2015; 2016b) have explored the use of a novel logic regression modelling approach (Ruczinski *et al.* 2003) to identify host-

informative SNP-SNP interactions encoded within intergenic regions (ITGRs) across the *E. coli* genome. Depending on the host species, logic regression modelling could cluster anywhere between 31.00–94.00% of *E. coli* strains collected from a specific host based on the presence of highly host-specific (i.e., exceeding 96.00% specificity) SNP-SNP biomarkers. The high specificity of the identified host-informative biomarkers was even observed between closely related host species, such as coyotes and dogs (i.e., within the family Canidae), demonstrating that logic regression can distinguish and classify *E. coli* isolates recovered from taxonomically-related host sources. Interestingly, logic regression analyses also revealed that the specific ITGRs that were particularly informative varied between different host groups, suggesting that strains belonging to different ecotypic groups may be defined by distinct regulatory adaptations (i.e., regulomes).

As such, logic regression appears to be capable of not only identifying niche-informative genetic regions in the *E. coli* genome, but also generating biologically plausible biomarkers that appear to reflect the adaptive processes driving the evolution of *E. coli* towards niche-specificity. Despite this, the ecotypic relevance of logic regression-generated biomarkers (i.e., whether the biomarkers can be correlated with a strain's original niche source) has yet to be validated through source attribution studies. Furthermore, while previous studies have evaluated logic regression for biomarker discovery purposes, its utility as an exploratory tool for ecotype discovery, especially when applied to an expanded collection of bacterial isolates, requires further evaluation. To address these gaps, we describe a logic regression-based biomarker discovery method, improving upon the workflows originally laid out by Zhi *et al.* (2015; 2016b), for the identification of niche-informative SNP-SNP biomarkers within ITGRs across the *E. coli* genome for strains recovered from a wide range of host and niche sources. Extending our methodology for source attribution

purposes, we then apply logic regression for the generation of human- and animal-specific biomarkers used for the source attribution of environmental *E. coli* isolates collected from the Indalsälven river in Northwestern Sweden (Maes *et al.* 2022) – thereby demonstrating the ecotypic relevance of logic regression as an ecotype discovery *and* attribution tool.

## 2.2 Material and Methods

### 2.2.1 Bacterial Strains for *In Silico* Whole Genome Sequence-Based Biomarker Discovery with Logic Regression

As previous studies have evaluated the use of logic regression for identifying ecotype-informative genetic markers indicative of *E. coli* niche-specificity (Zhi *et al.* 2015; Zhi *et al.* 2016b). we sought to assess its performance when applied to a broader data set. Specifically, in this analysis the performance of logic regression as an ecotype discovery tool was evaluated using larger collections of *E. coli* isolates derived from a greater number of potential niche sources that were also distributed across a wider geographical and temporal range. To assess this, a local repository of *E. coli* genome sequences was constructed, building on previous *E. coli* genome libraries (Zhi *et al.* 2016b), with an initial focus on expanding the range of host species represented. A total of 2925 *E. coli* genome sequences, collected from a range of host species (i.e., bovine, human, pig, sheep, chicken, turkey, mouse, rat, dog, cat, and other animals) and niches (i.e., wastewater) were downloaded from the NCBI GenBank database and then screened using a set of selection criteria designed to:

i) maximize the quality of the genome assemblies included in the final library;

ii) remove duplicate genomes and any genomes of strains with mislabeled isolation sources;

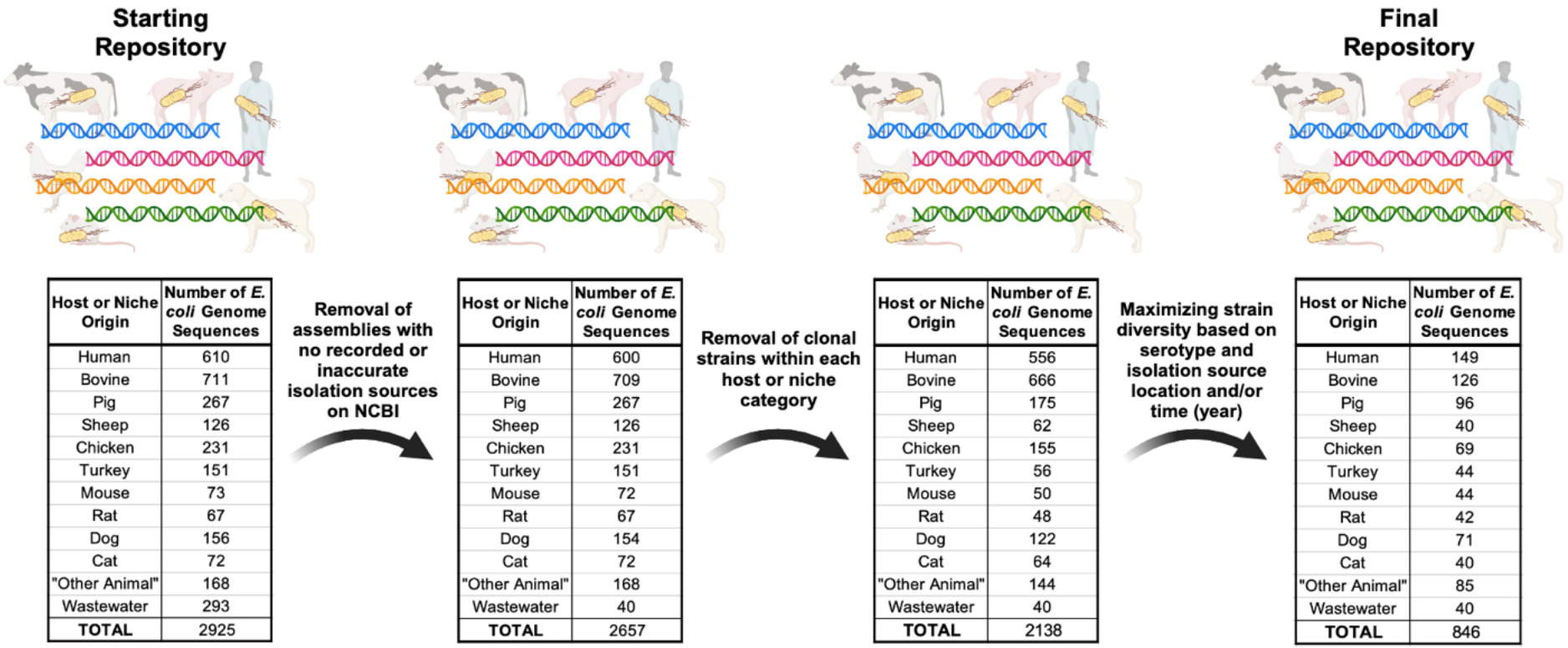
iii) minimize the degree of clonal representation (i.e., via sequence typing and serotyping) amongst the genome assemblies recovered from the same sequencing project; and

iv) maximize the temporal (i.e., year of isolation) and geographical (i.e., location of isolation) diversity of strains included in the final library (Figure 2-1).

The *E. coli* genome sequences that passed the screening criteria were then used to generate a local repository using BLAST+ v2.12.0 (Camacho *et al*. 2009). All *E. coli* strains used for the ecotype discovery analysis and their relevant metadata can be found in Supplementary Table 2–S1.

### 2.2.2 Selection of *E. coli* Intergenic Regions for Biomarker Discovery

Expanding on a set of *E. coli* ITGRs (i.e., *asnS–ompF*, *csgDEFG–csgBAC*, *uspC–flhDC*, *yedS–yedR*) that were previously found to be host-informative (Zhi *et al*., 2015; Zhi *et al*., 2016b), 58 additional ITGRs (i.e., 62 total) were selected as candidate loci for the discovery of ecotype-informative biomarkers via logic regression. The candidate ITGRs were selected based on the role of the flanking genes in mediating functions that could be associated with niche-adaptation, including nutrition, adhesion and biofilm formation, colonization factors, antibiotic resistance, and stress resistance (Supplementary Table 2–S2), as determined after reference against the UniProt (Bateman *et al*. 2017) and EcoCyc (Karp *et al*. 2018) databases. All candidate ITGR targets were extracted from the genome sequence of the laboratory reference strain *E. coli* K-12 MG1665 with bedtools v2.30.0 (https://github.com/arq5x/bedtools2), and screened against the local repository using BLAST+ v2.12.0 (Camacho *et al*. 2009). Only ITGR sequences that displayed ≥ 95% coverage with the queried sequence extracted from the reference *E. coli* K-12 MG1665 strain were retained for logic regression analysis. Additionally, ITGR loci that were found to be sparingly




| Host or Niche Origin | Number of *E. coli* Genome Sequences |
|---|---|
| Human | 610 |
| Bovine | 711 |
| Pig | 267 |
| Sheep | 126 |
| Chicken | 231 |
| Turkey | 151 |
| Mouse | 73 |
| Rat | 67 |
| Dog | 156 |
| Cat | 72 |
| "Other Animal" | 168 |
| Wastewater | 293 |
| **TOTAL** | 2925 |

| Host or Niche Origin | Number of *E. coli* Genome Sequences |
|---|---|
| Human | 600 |
| Bovine | 709 |
| Pig | 267 |
| Sheep | 126 |
| Chicken | 231 |
| Turkey | 151 |
| Mouse | 72 |
| Rat | 67 |
| Dog | 154 |
| Cat | 72 |
| "Other Animal" | 168 |
| Wastewater | 40 |
| **TOTAL** | 2657 |

| Host or Niche Origin | Number of *E. coli* Genome Sequences |
|---|---|
| Human | 556 |
| Bovine | 666 |
| Pig | 175 |
| Sheep | 62 |
| Chicken | 155 |
| Turkey | 56 |
| Mouse | 50 |
| Rat | 48 |
| Dog | 122 |
| Cat | 64 |
| "Other Animal" | 144 |
| Wastewater | 40 |
| **TOTAL** | 2138 |

| Host or Niche Origin | Number of *E. coli* Genome Sequences |
|---|---|
| Human | 149 |
| Bovine | 126 |
| Pig | 96 |
| Sheep | 40 |
| Chicken | 69 |
| Turkey | 44 |
| Mouse | 44 |
| Rat | 42 |
| Dog | 71 |
| Cat | 40 |
| "Other Animal" | 85 |
| Wastewater | 40 |
| **TOTAL** | 846 |

**Figure 2-1. Flowchart depicting the construction and refinement of local *E. coli* genome sequence repository for *in-silico* logic regression analyses.** The genome sequences of 2925 *E. coli* strains isolated from a wide range of host (i.e., human, bovine, pig, sheep, chicken, turkey, mouse, rat, dog, cat, and other animal hosts with lower representation in the repository) and environmental (i.e., wastewater) sources were downloaded from NCBI. The repository was then refined through: 1) the removal of assemblies with inaccurate or no available isolation source information; 2) the removal of clonal strains to reduce the degree of clonal representation in the repository; and 3) maximizing strain diversity through representation of sequence types, serotypes and isolation source location/time. The final 846 'representative' strains across the host species and niche source categories comprised the final repository, which was then used for the downstream *in-silico* logic regression analyses.

represented across the repository (i.e., in less than 750 strains), that were too short (i.e., less than 250bp in length), or that lacked sufficient sequence variation across the strains analysed (i.e., if over 50% the ITGR sequences extracted from the strains shared over 98% sequence identity) were removed and excluded from downstream analyses. The remaining ITGRs that passed the above screening criteria were then extracted from the strains in the repository with bedtools v.2.30.0 (https://github.com/arq5x/bedtools2), aligned with Clustal Omega v1.2.2 (Sievers *et al*. 2011), and then visualized and edited with the Jalview v2.11.0 platform (Waterhouse *et al.* 2009). The aligned sequences were then analysed with logic regression for the discovery of ecotype-informative, niche-specific biomarkers, as described below.

### 2.2.3 Identification of Ecotype-Informative SNP-SNP Biomarkers Within *E. coli* ITGRs via Logic Regression

Following previous workflows (Zhi *et al*. 2015; Zhi *et al*. 2016b) the sequence variation contained in the candidate ITGRs across each host and niche category in the repository was analysed using logic regression to identify ecotype-specific SNP-SNP biomarkers. Specifically, logic regression generates a binary classification for a given strain, corresponding to whether it originated from a specific host or niche or from some other source of origin. As it uses SNPs as predictive parameters, logic regression can thus generate logic models consisting of SNP-SNP interactions, represented with the Boolean logic terms 'AND', 'OR' and 'NOT', that can then serve as biomarkers of niche-specificity in *E. coli*, as follows:

$$\text{logit}(E[Y]) = \beta_0 + \beta_1 L_1 + \beta_2 L_2 + \ldots + \beta_p L_p$$

where:

- Y is a binary variable, corresponding to a strain's membership to one host or niche source group (Y = 1) or some other source of origin (Y = 0);
- $\beta_0$, $\beta_1$, $\beta_2$, …, $\beta_p$ are parameters indicating the degrees of association between the SNP patterns (L) and the prediction outcome (Y); and
- $L_1$, $L_2$, …, $L_p$ are the SNP-SNP interactions consisting of Boolean combinations (termed 'trees') of SNP genotypes (termed 'leaves') within the ITGRs.

All logic regression analyses were performed using a custom R script. As a massive number of potential models can be built with a varying number of trees of leaves, a simulated annealing algorithm was incorporated into the logic regression script to select the number of trees and leaves adaptively based on deviance to find the best fitting model for each source category (i.e., ecotype). While previous studies using logic regression restricted the model building parameters to 2 trees and 10 leaves to limit the computational burden of the analyses (Zhi *et al*. 2015; Zhi *et al.* 2016b), the size of the model is likely to impact the fit of the model produced. Furthermore, the 'optimal' model parameters may also vary depending on the specific source category and ITGR sequence being assessed. As such, an iterative model building approach was used in this study, in which logic models of sizes ranging from 2 to 3 trees and 20 to 30 leaves were generated and compared to determine the best performing model size for each source category and ITGR. As part of the model building process, the script runs a 10-fold cross-validation step and calculates a mean cross-validation test-score (CV-score) to assess the fit for each model. The model size with the lowest mean CV-test score was then selected for downstream logic regression analysis, thereby lowering the chances that the models selected will be 'overfitted'.

Following previous studies (Zhi *et al*. 2015; Zhi *et al*. 2016b), model performance was evaluated using measures of sensitivity and specificity, where sensitivity was defined as the

proportion of strains from a target source category that carried a specific SNP pattern, while specificity was defined as the proportion of strains from sources other than the target source category that did not carry the SNP biomarker of interest. Additionally, to evaluate the significance of the association between each biomarker and their corresponding source category, a permutation test was performed to assess the validity of each logic model (i.e., host- or niche-specific biomarker) that was generated by logic regression. To perform the permutation test, the source labels were randomly permuted, and the data was re-analysed with logic regression 1000 separate times. The number of instances where the permuted data sets produced logic models with higher performance values (as measured by the mean of the sensitivity and specificity of the models) than the models produced from the original data were counted, and this value was divided by 1000 to generate a p-value.

Two separate trials of logic regression analyses were run. The first trial included all host/niche source categories in the repository to evaluate the ability of logic regression in identifying ecotype-informative biomarkers across an expanded host/niche range using just single ITGR sequence data, with the 'other animal' and wastewater groups serving as negative controls (i.e., strains not associated with any of the target host groups). To improve on the generated models (i.e., by identifying more specific and sensitive biomarkers), a second trial was performed with two modifications. First, the ecotype range was reduced to only include human, bovine, chicken, and pig strains, alongside wastewater as a negative control group, to reduce the potential background 'noise' from less-represented source categories that could be detracting from the model building process. Furthermore, additional logic regression analyses were performed on concatenated ITGR sequences as biomarkers produced from concatenated sequences appear to be better performing overall than those generated from just single ITGR sequence data (Zhi *et al.*

2016b). Unlike previous analyses, however, which involved concatenating target ITGR sequences in an arbitrary order, all possible permutations of ITGR combinations were assessed with logic regression through an iterative approach, such that the best performing logic models for each source category could be reported for each group of ITGRs that were concatenated.

### 2.2.4 Bacterial Strains for *In Vitro* Biomarker Discovery and Application for Ecotype Attribution Purposes

To confirm the ecotypic relevance of logic regression-generated biomarkers, a source attribution analysis was performed. Logic regression analyses were performed on *E. coli* isolates collected from beaver, human, and reindeer hosts – reflecting some of the major sources of faecal contamination impacting environmental waters in the Jämtland region of Northwestern Sweden. A total of 32 faecal samples were collected from beavers and reindeer from 27 sampling sites within Jämtland county (Figure 2-2). One gram of each faecal sample was diluted in 100mL of peptone water (Oxoid, LP0037), plated on Membrane Faecal Coliform Agar (mFC, DifcoTM mFC agar, BD Biosciences, 267720) with 0.01% Rosolic acid (DifcoTM Rosolic acid, BD Biosciences, 232281), and then incubated at 44 ± 0.5°C for 22 ± 2 hours. Morphologically distinct blue colonies that grew on the mFC plates were then picked and grown in Lactose Tryptone Lauryl Sulphate Broth (LTLSB, Oxoid, CM0921) supplemented with 4-methylumbelliferyl-β-D-glucuronide (MUG supplement, Oxoid, BR0071E) for 21 ± 3 hours at 44 ± 0.5°C for the isolation of putative *E. coli* isolates. Confirmed *E. coli* isolates were then stored at -18°C in Brain Heart Infusion Broth (BHI, Oxoid, CM1135) supplemented with 20% glycerol (Apl, 33868). Additional Canadian isolates collected in previous analyses (Zhi *et al.* 2015) were also provided to supplement the library of *E. coli* strains collected from the animal faecal samples in Sweden. In total, 227 *E. coli*

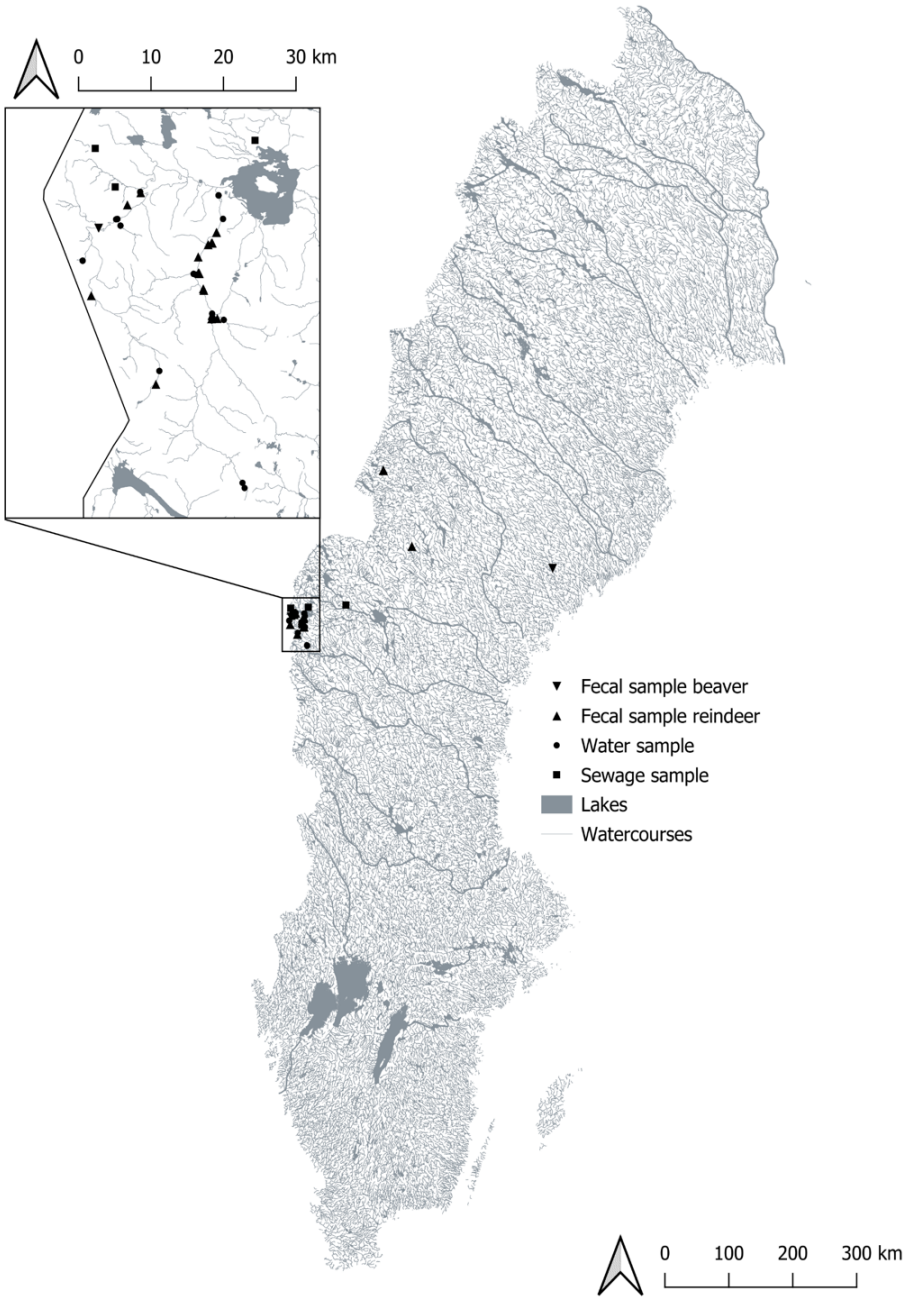



**Figure 2-2. Map of Sweden (© Lantmäteriet) depicting the sampling locations of faecal (n = 2 for beaver, n = 25 for reindeer), and water (n = 37) and sewage (n = 4) samples.** *E. coli* strains were isolated from faecal and water samples through collaborations with the Jonsson Lab. Most samples were taken in the main research area (depicted in the expanded view) located in the Jämtland county of Northwestern Sweden, while the remaining samples were taken at other locations (shown on the main map) in the surrounding area. These collected isolates were then supplemented with clinical human and beaver strains used in previous work conducted by Zhi *et al.* (2015), and *E. coli* genome sequences downloaded from NCBI to evaluate the practical application of logic regression for classifying environmental (i.e., water, sewage) *E. coli* isolates according to their original host source.

strains were used for this portion of the study, including: 51 reindeer (*Rangifer tarandus*) isolates collected from reindeer herds in Jämtland; 44 total beaver isolates, including 4 collected from local Eurasian beavers (*Castor fiber*) in Sweden and 40 isolates used in previous analyses (Zhi *et al.* 2015) collected from North American (*Castor canadensis*) beaver populations in Canada; and 133 total human isolates, including 115 isolates recovered from clinical faecal swabs collected at the Alberta Provincial Laboratory for Public Health (ProvLab) for routine microbiological testing (adhering to all ethics requirements; File #: Pro00005478_CLS3 at the University of Alberta) and 26 *E. coli* genome sequences with a global distribution screened from the NCBI GenBank database to bolster the logic regression model building process. In addition to these faecal isolates, 113 *E. coli* strains were also recovered from environmental water samples collected from mountain creeks feeding into Lake Ånnsjön in the Jämtland County region of Northwestern Sweden (Figure 2-2). As the host source of these water *E. coli* strains were unknown, these isolates were used to evaluate the applicability of the logic regression methodology for source tracking purposes based on its ability to classify the unknown isolates according to their potential original host-source. All information related to the environmental strains used for the targeted *in vitro* logic regression analysis can be found in Supplementary Table 2–S3.

### 2.3.5 *In Vitro* Validation of Logic Regression Analyses and Their Application for Microbial Source Attribution Purposes

For the *in vitro* logic regression analysis, the *asnS–ompF* and *csgDEFG–csgBAC* ITGRs were chosen as candidate targets as they have previously been shown to be ecotype-informative (Zhi *et al*. 2015; Zhi *et al*. 2016a; Zhi *et al*. 2016b) across a wide range of host- and niche-sources. The target ITGRs were amplified in each of the beaver, human, reindeer, and water isolates with

PCR using the primers listed in Table 2-1. The PCR conditions for the *asnS–ompF* and *csgBAC–csgDEFG* ITGRs were as follows: initial denaturation at 95°C for 4 min, 33 cycles of 95°C for 30 s, 58°C for 30 s, and 72°C for 1 min, followed by a 7 min extension at 72°C. The total volume of each PCR reaction was 50 µL and contained 10 µL of DNA template, 2U KAPA2G Robust Standard DNA Polymerase (Roche, KK5005) and each primer at a concentration of 500 nM. The PCR products were then sequenced bidirectionally with Sanger sequencing by Macrogen Europe (Amsterdam, The Netherlands), concatenated, and then aligned with Clustal Omega (Sievers *et al.* 2011). The 3' and 5' ends of the aligned sequences were then manually trimmed to remove any missing data.

Logic regression was then used to analyse the sequence variation within the *asnS–ompF* and *csgDEFG–csgBAC* intergenic sequences to identify host-informative biomarkers for the beaver, human, and reindeer isolates. Given that the results from the *in vitro* analysis will be used to classify the unknown water isolates, an additional step was taken to identify the 'optimal' model size parameters for building the logic models. Using the same custom R script, 5 random seed numbers were generated to run 5 separate iterations of model building for each of the beaver, human, and reindeer isolates, with models ranging in size from 1 to 5 trees and up to 30 leaves. The generated mean CV scores were then plotted against each model size for each host category to identify the 'optimal' model sizes for the logic model building process.

After training and optimization, the generated host models were used to attempt to classify the environmental water strains. As the original faecal-contributing source for each of the water strains was unknown, several rounds of classification were performed and the overall results were combined to determine the final classifications for the environmental isolates. Briefly, 1100 random seeds were generated with R, of which 1071 were retained after removing duplicate seed

**Table 2-1.** PCR primers used for in vitro, targeted ITGR logic regression analysis

| ITGR Target | Primer | Primer Sequence (5'–3') | Reference |
|---|---|---|---|
| *asnS–ompF* | *ompF-F* | TACGTGATGTGATTCCGTTC | Zaslaver *et al.* 2006 |
| | *ompF-R* | TGTTATAGATTTCTGCAGCG | |
| *csgDEFG–csgBAC* | *csgD-F* | GGACTTCATTAAACATGATG | Zaslaver *et al.* 2006 |
| | *csgD-R* | TGTTTTTCATGCTGTCAC | |

numbers. For each seed number, one iteration of logic regression was performed to produce host-specific logic models for the beaver, human, and reindeer strains. Only those logic building iterations (i.e., seed numbers) that generated logic models that were found to be at least 90% specific across all host categories were selected to be used for the source attribution of the water isolates.

Using another custom R script, the *asnS–ompF* and *csgDEFG–csgBAC* sequences extracted from the water isolates were then compared to the beaver-specific, human-specific, and reindeer-specific logic models generated for each 'iteration' (i.e., seed number) that passed the above criteria. Briefly, a maximum likelihood value was calculated for each water isolate corresponding to the likelihood that it could classified into the beaver, human, and/or reindeer groups. Isolates that received a positive classification (i.e., corresponding to a likelihood value of at least 0.5) against only one host model were tentatively classified as originating from that host source (i.e., water isolates with a positive classification when compared to the human model were tentatively called as being human in origin), whereas isolates that were not classified by any of the host models were left unclassified. In the case that a water isolate received positive predictions across multiple host models, a comparative evaluation step was used to resolve the final classification between the host categories. Specifically, an evaluation value was calculated for these indeterminately-classified isolates, which combined the model specificity value with the prediction/likelihood value assigned to the strain – thereby reflecting the ability of the model to predict the host source of a given isolate and the overall confidence in the model's prediction. The evaluation values corresponding to each host model assigned to the indeterminately-classified isolates were then compared. If the difference between these values was greater than 0.2, the isolate was classified according to the host model with the highest evaluation value; conversely, if the

difference was less than 0.2, the isolate was given a joint classification between the corresponding host models (i.e., as a potential host generalist *E. coli* isolate). To make a final classification for each water isolate, the 'tentative' classifications across all iterations (i.e., seed number) of logic regression that passed the above criteria were pooled. Classifications across the classification iterations remaining that were at least 80% consistent for each isolate were retained as final classifications, and these isolates were assigned to the corresponding host source. Conversely, water isolates with classifications that lacked this level of consistency across iterations or were consistently given an indeterminate (i.e., multi-host) classification were left unclassified with no known host source.

## 2.3 Results

### 2.3.1 Construction of Local *E. coli* Genome Repository and ITGR Candidate List for *In Silico* Ecotype-Specific Biomarker Discovery

A total of 2925 *E. coli* genome sequences were downloaded from NCBI to construct a local genome repository, including 610 human strains, 711 bovine strains, 267 pig strains, 126 sheep strains, 231 chicken strains, 151 turkey strains, 73 mouse strains, 67 rat strains, 156 dog strains, 72 cat strains, 168 strains that were grouped into an 'other animal' category due to the limited representation of their host source in the repository, and 294 wastewater strains. The initial repository was screened to maximize sequence quality and strain diversity (i.e., limiting clonal representation), resulting in a final repository of 'representative' *E. coli* genome sequences from each host and niche source (Figure 2-1). After screening, the final repository consisted of 846 *E. coli* genome sequences including 149 human strains, 126 bovine strains, 96 pig strains, 40 sheep strains, 69 chicken strains, 44 turkey strains, 44 mouse strains, 42 rat strains, 71 dog strains, 40 cat

strains, 85 strains from 'other animals' as a negative control group for the other host categories, and 40 wastewater strains as an additional non-host associated, negative control group (Supplementary Table 2–S1).

Expanding on a set of ITGRs that were previously found to be host-informative (Zhi *et al.* 2015; Zhi *et al.* 2016b), including the *flhDC–uspC*, *asnS–ompF*, *csgDEFG–csgBAC*, and *yedS–yedR* loci, 62 total candidate ITGRs were evaluated for biomarker discovery purposes (Supplementary Table 2–S2). The ITGRs that were selected were flanked by genes mediating functions relevant for niche-adaptation, including: adhesion and colonization factors, including fimbrial and pili systems; stress resistance, including genes associated with heat shock, acid stress, antibiotic stress, and environmental persistence; motility and flagellar systems; and nutrition, including metabolic pathways for various sugar substrates. Of the 62 candidate ITGRs identified, 29 were found to be sufficiently represented across the strains in the repository (i.e., in at least 750 strains), of satisfactory length (i.e., at least 250 bp), and displayed sufficient sequence diversity and were thus retained for subsequent analysis with logic regression.

### 2.3.2 Single-ITGR Logic Regression-Based Ecotype-Specific Biomarker Discovery Analysis with Expanded Source Range

The sequence variation contained within each of the 29 candidate ITGRs was analysed using logic regression to generate source-specific logic models for each host- and niche-category represented in the expanded repository. Although previous work using logic regression for biomarker discovery purposes restricted the model building parameters to just 2 trees and 10 leaves (Zhi *et al*. 2015; Zhi *et al*. 2016b), the iterative approach utilised in this study revealed that the 'best' model size differed based on the source category and ITGR sequence assessed. Depending

on the specific source-ITGR pairing, the generated logic models varied in size, ranging from as small as 3 trees and 16 leaves to as large as 3 trees and 25 leaves (Supplementary Table 2–S4). Furthermore, of the generated logic models that were source-informative, performance varied depending on the specific host- or niche-source and the specific ITGR locus analysed.

For the single-ITGR biomarkers, sensitivities ranged from as low as 3.13% to as high as 66.67% while specificities ranged between 84.25% to 100.00% (Supplementary Table 2–S4), though the most informative ITGRs differed across each ecotypic group (Table 2-2). While ecotype-specific biomarkers could be generated for each host or niche, the ecotype discovery approach appeared to be especially effective for certain source categories. Reflecting this, logic models of at least 30.00% sensitivity and over 97.00% specificity were produced for the pig (i.e., in the *csgDEFG–csgBAC* locus), sheep (i.e., in the *flhDC–uspC* and *yjjP–[yjjQ-bglJ]* loci), and mouse (i.e., in the *ompC–rcsDB,* and *nanCMS–fimB* loci) groups, with select mouse-informative biomarkers exceeding 50.00% sensitivity and 97.00% specificity (i.e., in the *yedS–yedR* and *csgDEFG–csgBAC* loci). Interestingly, although the wastewater group served as a non-host associated negative control for the other host-categories, the biomarkers generated for the wastewater strains were among the best performing, with sensitivities ranging from 37.50% to 50.00% and specificities exceeding 99.00% (Table 2-2).

### 2.3.3 Logic Regression-Based Ecotype-Specific Biomarker Discovery with Reduced Source Range

To improve the performance of the generated biomarkers, a second logic regression analysis was performed with two modifications: first, the host range was reduced (i.e., though it was still comparatively larger than previous analyses [Zhi *et al.* 2016b]) to include only bovine,

**Table 2-2.** Top 5 informative intergenic regions for each host/niche-category in the expanded repository, as determined via logic regression with ten-fold cross-validation

| Top 5 ITGRs | Bovine | | Top 5 ITGRs | Cat | | Top 5 ITGRs | Chicken | |
|---|---|---|---|---|---|---|---|---|
| | Sensitivity | Specificity | | Sensitivity | Specificity | | Sensitivity | Specificity |
| ***azuC*** | 13.64% | 100.00% | ***csgDEFG*** | 16.67% | 99.34% | ***araC*** | 12.50% | 100.00% |
| ***gadE*** | 9.09% | 100.00% | ***flhDC*** | 11.11% | 98.66% | ***ompC*** | 25.00% | 99.38% |
| ***ykgR*** | 9.09% | 100.00% | ***mdtABCD*** | 14.29% | 98.75% | ***bdm*** | 11.11% | 99.36% |
| ***fimE*** | 10.00% | 97.64% | ***acrEF*** | 16.67% | 96.13% | ***bssS*** | 9.09% | 100.00% |
| ***yedS*** | 31.25% | 84.25% | - | - | - | ***bluF*** | 11.11% | 98.68% |
| **Top 5 ITGRs** | **Dog** | | **Top 5 ITGRs** | **Human** | | **Top 5 ITGRs** | **Mouse** | |
| | **Sensitivity** | **Specificity** | | **Sensitivity** | **Specificity** | | **Sensitivity** | **Specificity** |
| ***ompF*** | 12.50% | 98.65% | ***ftnB*** | 17.65% | 99.33% | ***ompC*** | 37.50% | 100.00% |
| ***emrKY*** | 5.56% | 100.00% | ***gadE*** | 8.70% | 99.31% | ***rpoE*** | 14.29% | 100.00% |
| ***bdm*** | 5.56% | 100.00% | ***fucPIKUR*** | 12.50% | 97.79% | ***fimB*** | 33.33% | 100.00% |
| ***yobF*** | 5.56% | 99.33% | ***ecpRABCDE*** | 21.74% | 93.94% | ***yedS*** | 62.50% | 97.78% |
| ***mdtABCD*** | 9.09% | 100.00% | ***flhDC*** | 20.69% | 92.25% | ***csgDEFG*** | 66.67% | 97.26% |
| **Top 5 ITGRs** | **Pig** | | **Top 5 ITGRs** | **Rat** | | **Top 5 ITGRs** | **Sheep** | |
| | **Sensitivity** | **Specificity** | | **Sensitivity** | **Specificity** | | **Sensitivity** | **Specificity** |
| ***csgDEFG*** | 20.00% | 97.97% | ***ompF*** | 22.22% | 100.00% | ***flhDC*** | 33.33% | 100.00% |
| ***yobF*** | 12.50% | 98.68% | ***fimE*** | 18.18% | 100.00% | ***ftnB*** | 20.00% | 100.00% |
| ***rpoE*** | 30.00% | 97.99% | ***gadE*** | 16.67% | 100.00% | ***acrEF*** | 14.29% | 100.00% |
| ***rcsA*** | 19.05% | 99.31% | ***ypfM*** | 12.50% | 100.00% | ***yehABCD*** | 14.29% | 100.00% |
| ***agaR*** | 11.11% | 99.32% | ***csgDEFG*** | 12.50% | 100.00% | ***yjjQ*** | 33.33% | 98.08% |
| **Top 5 ITGRs** | **Turkey** | | **Top 5 ITGRs** | **Wastewater** | | | | |
| | **Sensitivity** | **Specificity** | | **Sensitivity** | **Specificity** | | | |
| ***ypfM*** | 16.67% | 99.38% | ***ftnB*** | 50.00% | 100.00% | | | |
| ***csgDEFG*** | 25.00% | 98.05% | ***araC*** | 37.50% | 100.00% | | | |
| ***fimB*** | 28.57% | 98.62% | ***fimB*** | 37.50% | 100.00% | | | |
| ***fucPIKUR*** | 14.29% | 99.35% | ***azuC*** | 44.44% | 100.00% | | | |
| ***ompC*** | 11.11% | 100.00% | ***ypfM*** | 44.44% | 99.36% | | | |

chicken, human, and pig strains, thereby reflecting the major zoonotic routes for *E. coli* transmission; and second, concatenated ITGR sequences were used as input for logic regression to improve the sensitivity and specificity of the generated biomarkers.

As in the first biomarker discovery analysis with the expanded source range, a significant degree of variability was observed across the single-ITGR-based logic models that were generated for each host category. The 'optimal' model size again differed depending on the host source of interest and the specific ITGR analysed, though to a lesser extent when compared to the logic models generated with the expanded source repository, ranging in size from 3 trees and 20 leaves to 3 trees and 24 leaves (Supplementary Table 2–S5). Similarly, the performance of the generated biomarkers also varied across host categories and ITGRs, with sensitivities ranging between 3.85% and 76.92% and specificities ranging from 65.00% to 100.00%. Interestingly, the performance of the logic models appeared to improve only for select host categories when the source range was reduced. The human-informative logic models, for instance, improved drastically during the biomarker discovery analysis with a reduced host range, with sensitivities reaching as high as 61.29% and specificities exceeding 94.00% (Table 2-3). Though to a lesser extent, the bovine-informative models also improved, with the bovine biomarkers displaying sensitivities as high as 26.92% and specificities of at least 91.00%. In contrast, when compared to the original biomarker discovery analysis with the expanded source repository, the generated logic models for the chicken and pig groups exhibited similar sensitivities and specificities, with no clear improvement in model performance between the two analyses.

With the significant variability observed across the single-ITGR logic models, no single ITGR seemed to be adequately informative across each of the host categories represented in the reduced repository. Additionally, compared to the first logic regression analysis with the expanded

**Table 2-3.** Top 5 performing intergenic regions for each host-category in the reduced repository, as determined via logic regression with ten-fold cross-validation

| Top 5 ITGRs | Bovine | | Top 5 ITGRs | Chicken | | Top 5 ITGRs | Human | | Top 5 ITGRs | Pig | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Sensitivity | Specificity | | Sensitivity | Specificity | | Sensitivity | Specificity | | Sensitivity | Specificity |
| ***ecpRABCDE*** | 18.18% | 98.33% | ***azuC*** | 16.67% | 100.00% | ***hns*** | 61.29% | 94.74% | ***rpoE*** | 28.57% | 95.95% |
| ***acrEF*** | 24.00% | 95.00% | ***ompC*** | 16.67% | 96.05% | ***yedS*** | 43.48% | 96.15% | ***ompC*** | 21.05% | 100.00% |
| ***emrKY*** | 26.92% | 91.94% | ***mdtABCD*** | 10.00% | 100.00% | ***emrKY*** | 38.71% | 94.74% | ***yobF*** | 25.00% | 91.55% |
| ***ftnB*** | 23.08% | 93.55% | ***fimE*** | 11.11% | 98.53% | ***mdtABCD*** | 33.33% | 94.44% | ***yjjQ*** | 22.22% | 92.75% |
| ***yobF*** | 23.08% | 91.80% | ***yedS*** | 40.00% | 87.69% | ***acrEF*** | 67.74% | 88.89% | ***fimB*** | 15.00% | 100.00% |

source range, the top performing ITGRs for the bovine, chicken, human, and pig categories differed when logic regression was performed using the reduced repository. As such, to identify target input sequences that could be informative across all host categories simultaneously, candidate ITGR sequences were concatenated and re-analysed using logic regression. Specifically, ITGRs that were found to produce informative biomarkers for more than one host group (i.e., *emrKY–evgAS*, *yedS–yedR*, *ompC–rcsDB*, and *ibsB–[mdtABCD-baeSR]*) were chosen for concatenation.

Upon re-analysis, 3 different collections of ITGRs, when concatenated, were found to be informative to varying degrees across the ecotypes interrogated, resulting in the generation of biomarkers with better sensitivities and specificities compared to the biomarkers that were produced using just single ITGR sequence data (Table 2-4). Concatenating the *emrKY–evgAS*, *yedS–yedR*, and *ibsB–(mdtABCD-baeSR)* intergenic loci, for instance, produced biomarkers with sensitivities exceeding 20.00% and specificities of at least 91.00% across all host categories. Interestingly, all biomarkers that were generated from these ITGRs were found to be statistically associated with each source category ($p < 0.05$), except for the biomarker that was produced for the pig strains ($p = 0.055$). Biomarker performance appeared to improve even further upon concatenating the *yedS–yedR*, *ompC–rcsDB*, and *emrKY–evgAS* ITGR targets, as all biomarkers were found to be significantly associated with each host category ($p < 0.05$), with sensitivities and specificities exceeding 30.00% and 90.00%, respectively. Biomarker performance continued to improve when all four ITGR targets were concatenated, resulting in the generation of biomarkers exceeding 36.00% sensitivity and 91.00% specificity that were also all found to be closely associated with their respective host groups ($p < 0.05$). Remarkably, although it was included in the logic regression analyses as a non-host-associated, negative control group, the biomarkers that

**Table 2-4.** Performance and strength of association of generated logic models with each host category in the reduced source repository, as determined with logic regression analysis on concatenated ITGR sequences and ten-fold cross validation

| Concatenated ITGR Targets | Host Category | Model Size | Model | Sensitivity | Specificity | P-Value (ANOVA) |
|---|---|---|---|---|---|---|
| *emrKY–evgAS; yedS–yedR; ibsB–(mdtABCD-baeSR)* | Bovine | 3 Trees, 23 Leaves | *=3.05 -3.29 * ((((not emrKY236_T) or (not mdtABCD54_A)) or (yedS198_T or (not yedS256_A))) and ((yedS336_T and (not yedS43_A)) and (yedS314_A and yedS353_A))) -4.74 * ((((not yedS237_T) or mdtABCD54_A) or (yedS375_T or mdtABCD187_A)) and (((not yedS211_G) or (not yedS429_T)) and (not emrKY59_C))) -3.2 * (((emrKY247_C and (not yedS233_C)) and (mdtABCD19_G and yedS397_G)) and (((not mdtABCD94_C) and (not emrKY305_G)) and (yedS110_T and yedS358_T)))* | 20.00% | 93.44% | 0.015 |
| | Chicken | 3 Trees, 23 Leaves | *=2.44 +2.66 * (((yedS237_C or (not emrKY210_G)) and (yedS316_T or (not emrKY2_T))) or ((emrKY196_C or (not yedS37_C)) or (yedS236_A and (not yedS169_-)))) -4.45 * (((yedS105_T and (not yedS351_A)) or emrKY214_T) and (((not yedS445_A) and yedS107_T) and ((not yedS368_C) and (not emrKY82_A)))) -6.62 * (((yedS135_C or yedS318_C) and (yedS171_G or mdtABCD252_T)) and (((not mdtABCD28_G) or yedS139_G) and ((not yedS374_G) or emrKY360_G)))* | 45.45% | 94.29% | < 0.001 |
| | Human | 3 Trees, 24 Leaves | *=-22 +21.8 * (((emrKY248_C and mdtABCD3_C) and ((not yedS37_A) and (not yedS37_A))) and ((yedS371_G and (not yedS319_A)) and ((not yedS345_T) and yedS185_G))) -4.57 * (((yedS208_G and (not yedS315_C)) and ((not yedS280_-) and (not yedS334_T))) or ((yedS241_T or mdtABCD54_A) or ((not mdtABCD29_C) or yedS292_G))) +3.51 * (((mdtABCD94_T or yedS37_A) or ((not yedS392_A) or yedS379_C)) or (((not emrKY218_T) and (not yedS167_C)) or ((not yedS37_A) or (not yedS80_A))))* | 50.00% | 91.80% | < 0.001 |
| | Pig | 3 Trees, 24 Leaves | *=-20.8 +22.7 * ((((not yedS100_A) and (not yedS84_A)) and (mdtABCD4_C and (not emrKY247_T))) and (((not emrKY325_T) and (not emrKY212_A)) and (mdtABCD46_G and yedS74_G))) -3.69 * (((mdtABCD130_A or yedS368_C) or ((not yedS290_A) or mdtABCD148_G)) and (((not emrKY218_T) or (not mdtABCD98_G)) or ((not yedS185_T) and yedS106_T))) +4.87 * ((((not emrKY4_A) or (not yedS256_A)) and ((not yedS375_T) or (not yedS372_T))) or ((emrKY4_A0 or yedS371_A) or ((not yedS287_C) or yedS367_G)))* | 23.53% | 95.31% | 0.055 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Wastewater | 3 Trees, 21 Leaves | *=-0.594 -18.3 * (((not yedS107_G) or ((not mdtABCD22_T) and (not yedS208_G))) and (((not mdtABCD92_C) or yedS316_C) or ((not yedS351_G) or yedS82_T))) -4.56 * (((not mdtABCD69_T) or ((not yedS211_G) or (not yedS379_T))) or ((mdtABCD77_C and (not yedS237_C)) or ((not yedS302_T) or yedS345_T))) +4.37 * ((((not emrKY131_A) or yedS315_A) or ((not emrKY220_G) or yedS87_C)) and (yedS24_C or ((not yedS242_G) and (not yedS224_T))))* | 53.85% | 100.00% | < 0.001 |
| *yedS–yedR; ompC–rcsDB; emrKY–evgAS* | Bovine | 3 Trees, 24 Leaves | *=-6.16 +4.91 * (((yedS197_G or (not yedS220_C)) and (emrKY359_C and (not yedS273_C))) or (((not ompC322_G) or rcsDB212_A) or (yedS314_G or yedS217_C))) +4.65 * ((((not yedS110_T) or (not ompC154_A)) or (rcsDB16_- or (not emrKY50_C))) and ((yedS189_C or emrKY248_T) and ((not yedS316_C) and emrKY236_T))) +5.23 * ((((not yedS358_T) or (not yedS60_C)) or ((not yedS242_A) and (not ompC271_A))) and (((not rcsDB193_C) and (not yedS375_T)) and ((not yedS492_A) and emrKY158_G)))* | 30.00% | 90.32% | 0.015 |
| | Chicken | 3 Trees, 24 Leaves | *=-3.5 -4.66 * ((((not ompC220_C) or yedS315_T) and ((not yedS37_A) and (not rcsDB210_A))) and (((not yedS237_C) or yedS429_T) or ((not emrKY242_C) or yedS37_A1))) +4.31 * (((yedS37_A08 and (not emrKY242_T)) or ((not yedS135_C) and (not rcsDB63_T))) or ((yedS429_T and (not rcsDB302_T)) and (yedS37_A88 or ompC293_A))) +3.21 * ((((not yedS294_A) or (not yedS345_C)) or (rcsDB107_T and yedS379_T)) and (((not yedS302_C) and yedS257_T) and ((not yedS316_C) and yedS212_C)))* | 37.50% | 93.24% | 0.034 |
| | Human | 3 Trees, 21 Leaves | *=-20.2 +20.5 * ((((not emrKY248_T) and yedS345_C) and (yedS171_G and (not rcsDB140_C))) or emrKY316_C) -2.86 * ((((not rcsDB123_A) or (not rcsDB56_T)) or (yedS230_C and (not yedS88_A))) and ((yedS74_G and yedS111_A) and (rcsDB63_C and (not ompC272_A)))) +2.04 * (((yedS80_T or yedS243_T) or ((not yedS383_C) or rcsDB188_-)) or ((yedS208_A and (not yedS92_A)) and (emrKY360_G and (not yedS37_A))))* | 41.94% | 92.16% | 0.002 |
| | Pig | 3 Trees, 21 Leaves | *=-25.7 +4.28 * (((not emrKY247_T) and (yedS303_C or rcsDB276_A)) and ((not yedS312_-) and yedS304_A)) +23.9 * ((((not yedS185_C) and yedS74_G) and ((not yedS255_A) and ompC231_C)) and (((not rcsDB210_A) and ompC258_T) and (yedS111_A and ompC271_C))) +4.48 * ((((not yedS111_A) or (not rcsDB74_A)) or (yedS111_A and (not emrKY217_T))) or (((not yedS375_C) and rcsDB107_C) or ((not yedS185_G) or emrKY184_A)))* | 41.17% | 92.31% | 0.002 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Wastewater | 3 Trees, 23 Leaves | *=-3.74 +3.52 * (((ompC184_A or yedS379_C) and ((not yedS351_A) and yedS129_T)) or ((yedS273_A or ompC168_A) and (yedS273_A or (not rcsDB63_C)))) +5.05 * ((((not yedS49_T) and emrKY203_A) and (ompC220_C and yedS240_C)) or (((not emrKY214_T) and (not yedS24_A)) or ((not emrKY95_C) or yedS429_T))) -7.82 * (((emrKY215_A or (not yedS24_A)) and ((not yedS24_A) or yedS270_A)) and (yedS24_A and ((not rcsDB134_G) and emrKY350_A)))* | 66.67% | 97.37% | < 0.001 |
| *emrKY–evgAS; yedS–yedR; ibsB–(mdtABCD-baeSR); ompC–rcsDB* | Bovine | 3 Trees, 24 Leaves | *=-22 -4.57 * ((((not emrKY217_T) or mdtABCD92_C) and ((not yedS318_C) and mdtABCD27_C)) and (((not yedS172_T) and (not yedS358_C)) and ((not ompC184_T) or (not emrKY92_A)))) +22.1 * (((yedS242_G and (not emrKY135_G)) and ((not yedS431_-) and (not yedS256_T))) or (((not mdtABCD54_G) and rcsDB183_C) and (rcsDB41_G and yedS198_C))) +4.16 * (((mdtABCD17_A or ompC220_C) and (yedS269_A and yedS318_T)) or ((ompC337_T or (not rcsDB16_T)) or (rcsDB30_T or (not yedS431_T))))* | 45.00% | 91.80% | < 0.001 |
| | Chicken | 3 Trees, 24 Leaves | *=0.587 +5.71 * ((((not yedS82_C) or yedS135_T) and (yedS270_C and (not rcsDB63_T))) or (((not yedS295_T) or yedS189_C) and (rcsDB134_C and rcsDB85_C))) -3.87 * (((yedS345_C and yedS107_T) and (yedS37_C or (not mdtABCD187_T))) and ((yedS16_A and (not emrKY196_C)) and ((not yedS380_G) and emrKY6_C))) -4.5 * ((((not rcsDB131_G) or (not ompC335_-)) and ((not ompC233_T) and (not yedS255_A))) and (((not rcsDB210_A) or (not mdtABCD94_A)) and (mdtABCD32_G and (not yedS371_T))))* | 36.36% | 97.14% | 0.002 |
| | Human | 3 Trees, 21 Leaves | *=-4.82 -21.9 * (((yedS255_T or rcsDB56_C) and ((not yedS374_A) or (not mdtABCD68_G))) or (((not mdtABCD29_C) or yedS378_A) or ((not yedS292_A) or yedS170_-))) +3.6 * (((yedS80_T or mdtABCD94_T) or ((not ompC80_A) or mdtABCD68_T)) or ((emrKY82_G and (not emrKY16_T)) or (not mdtABCD98_G))) +4.51 * (((rcsDB86_G or yedS208_A) and ((not emrKY407_C) and yedS345_C)) and ((not mdtABCD68_T) and yedS244_T))* | 52.17% | 93.10% | 0.004 |
| | Pig | 3 Trees, 20 Leaves | *=-0.241 +3.33 * (((yedS287_T or emrKY274_C) or ((not yedS256_A) or yedS185_T)) or (((not mdtABCD108_A) or yedS371_A) or yedS375_T)) -20.3 * ((((not mdtABCD112_C) or emrKY247_T) or (yedS273_C or yedS281_C)) and yedS251_A) -2.77 * (((mdtABCD187_T and (not yedS198_T)) and ((not yedS281_A) and (not rcsDB154_G))) or (((not yedS197_A) or (not rcsDB210_G)) or (rcsDB183_T or yedS375_T)))* | 47.06% | 93.75% | 0.004 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Wastewater | 3 Trees, 21 Leaves | *=-2.54 +3.14 * (((mdtABCD187_C or yedS244_C) and ((not yedS316_C) and emrKY131_A)) and (((not ompC338_-) or emrKY220_T) and (not mdtABCD184_A))) +5.95 * ((((not ompC174_A) and (not mdtABCD17_T)) or ompC271_A) and ((rcsDB179_A and mdtABCD163_G) and (yedS445_G and mdtABCD89_A))) -5.22 * ((((not yedS135_C) or (not ompC184_A)) or (yedS114_A or ompC174_A30)) or ((ompC174_A66 or emrKY92_G) or (not yedS351_G)))* | 70.00% | 97.18% | < 0.001 |

were produced for the wastewater strains were consistently the best-performing. Indeed, across all combinations of ITGRs that were found to be host-informative to some degree, the wastewater biomarkers consistently exhibited the highest performance parameters, with sensitivities reaching 70.00% and specificities exceeding 97.00%.

### 2.3.4 Validation of Logic Regression for the Ecotype Attribution of Environmental Water *E. coli* Isolates

While the previous analyses highlight the potential of logic regression for identifying ecotype-informative biomarkers using *E. coli* ITGR sequence data, *in silico* analyses alone do not necessarily demonstrate whether these biomarkers are ecotypically relevant. Thus, to validate its applicability for source (i.e., ecotype) attribution purposes, logic regression was used to generate additional human-, beaver-, and reindeer-informative biomarkers for the source (i.e., ecotype) attribution of unknown *E. coli* strains contaminating rivers in the Jämtland county of Northwestern Sweden. For the sequence selection, the *asnS–ompF* and *csgDEFG–csgBAC* intergenic loci were chosen for analysis as they have been previously shown to be particularly source-informative across a wide range of host- and niche-derived isolates (Zhi et al. 2015; Zhi et al. 2016a; Zhi *et al.* 2016b). To refine the model building process, the 'optimal' size for the models generated from the concatenated *asnS–ompF* and *csgDEFG–csgBAC* sequence was first determined for each of the human, beaver, and reindeer groups. Interestingly, regardless of the relative performance of the models for each host source, the mean CV-scores for each ecotype appeared to plateau from 18 leaves onward when the number of trees were set between 3 to 5 (Figure 2-3). As such, for all subsequent model building with the given ecotype range (i.e., beaver, human, and reindeer) and input sequences (i.e., *asnS–ompF* concatenated with *csgDEFG–csgBAC*), the size parameters were

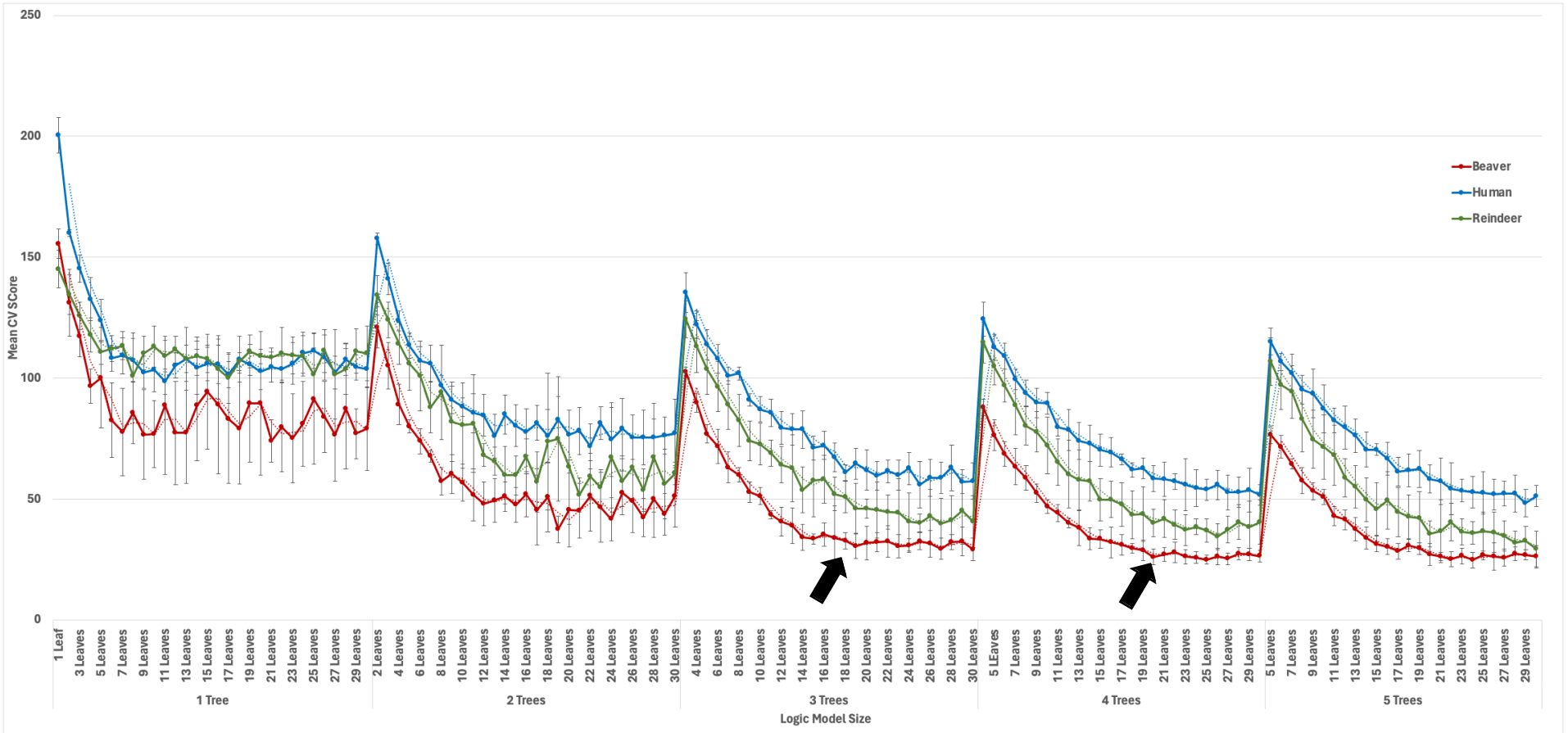



**Figure 2-3. Determination of the optimal model size parameters for beaver, human, and reindeer-specific logic models.** Host-informative logic models were generated for beaver (red lines), human (blue lines) and reindeer (green lines) *E. coli* strains collected from Sweden and Canada using the sequence variation contained within a concatenated sequence consisting of the *asnS–ompF* and *csgDEFG–csgBAC* intergenic regions. To determine the optimal model size for each host group, 5 independent iterations of logic regression model building was performed to determine the average performance, measured through cross-validation (CV) scores, for each model size between 1 to 5 trees and 1 to 30 leaves. The optimal size range (i.e., where the curves all began to plateau, indicated with black arrows) was then used to inform the model building portion of the classification analysis for the source attribution of unknown environmental water *E. coli* isolates.

restricted to 3 to 4 trees and 18 to 25 leaves.

Having determined the optimal model building parameters, a total of 1071 independent iterations of logic regression were performed to generate host-specific biomarkers for each of the beaver, human, and reindeer groups. Given that the generated logic models were used for the source attribution of environmental water *E. coli* isolates, only those iterations of logic regression (i.e., seed numbers) that produced host-informative biomarkers of at least 90% specificity across all host groups were retained for the classification analysis. Following screening, 273 logic regression iterations (i.e., 273 logic models per host source) passed the screening criteria and were used to classify the original host source for the unknown water *E. coli* strains (Supplementary Table 2–S6). Given that the classifications for each water isolate may vary across each iteration of logic regression analysis, only isolates with classifications that were at least 80% consistent across all 273 iterations were given a final classification. Overall, 63 water isolates were inconsistently classified across the 273 iterations and were left unclassified (Supplementary Table 2–S7). Interestingly, for 2 water isolates over 97% of classifications were multi-host (i.e., as 'Beaver | Human | Reindeer'), leaving these 2 isolates unclassified as potential host-generalist *E. coli* strains. The remaining 48 water isolates were classified with a sufficient level of consensus, of which 19 were designated as beaver in origin, 16 as human in origin, and 13 that were classified as reindeer in origin (Figure 2-4).

## 2.5 Discussion

Accruing evidence indicates that the *E. coli* species exhibits a significant degree of host- and niche-specificity, with several source-informative genetic determinants having been identified to date. While the presence/absence of these genetic markers might influence the ability of a strain

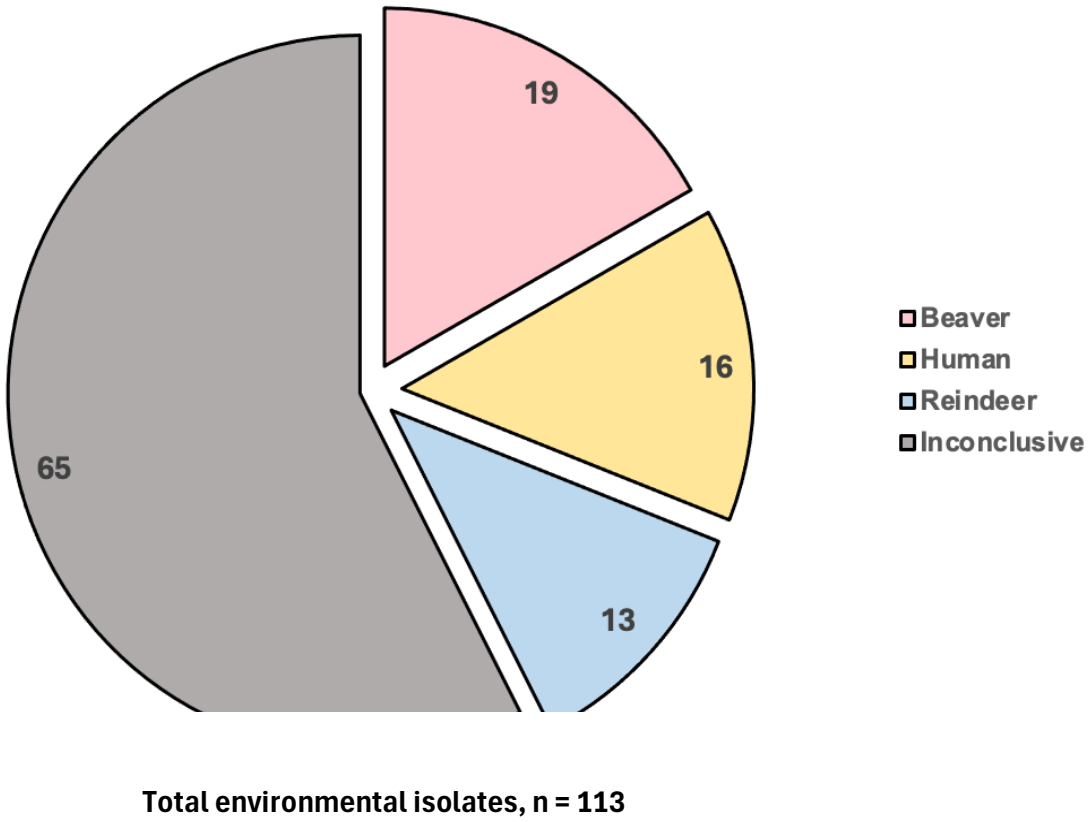



**Figure 2-4. Classification of unknown environmental water *E. coli* isolates according to presumptive original host source based on logic regression analyses.** A total of 113 *E. coli* isolates recovered from environmental water and sewage samples collected from the Jämtland County region of Northwestern Sweden were classified according to the application of beaver-, human-, and reindeer-specific biomarkers identified through logic regression. Each environmental isolate was individually classified as being beaver (red), human (yellow), or reindeer (blue) in origin across 273 independent iterations of logic regression analysis. Isolates with classifications reaching at least 80% consensus across the 273 iterations were given a final designation according to their predicted host source, whereas isolates lacking this level of consistency in their classifications or those that were consistently classified into multiple host groups were given an inconclusive final designation (grey).

to colonize a given host or niche, differences in the relative fitness of strains across different niches could alternatively be reflected in the 'regulome', or how well they are able to sense and respond to the specific stressors that are present within a given environment. Indeed, ITGRs contain various promoter and repressor sites that can influence the expression of the flanking genes, which can in turn influence the ability of a given strain to exploit and adapt to a specific niche. As the sequences within ITGRs appear to be under strong purifying selection (Thorpe *et al.* 2017), genetic markers of host- and/or niche-specificity may be better reflected through distinct SNP patterns within these regulatory regions. Reflecting this, previous studies by Zhi *et al.* (2015; 2016b) have demonstrated logic regression to be particularly effective at identifying these niche-specific SNP patterns within ITGRs across the *E. coli* genome. We have expanded on these findings, as we were able to similarly use logic regression for ecotype-informative biomarker discovery purposes, but with an expanded repository of *E. coli* strains collected from a wider range of geographical sites and niche sources.

Overall, considerable variability was observed in the performance of the biomarkers generated, depending on the specific ecotype and ITGR locus evaluated. While biomarkers of only limited sensitivity (i.e., less than 30.00%) were mostly generated for bovine-, cat-, chicken-, dog-, human-, rat- and turkey-derived *E. coli* strains, select single ITGRs appeared to encode for biomarkers of much higher sensitivity for the mouse, pig, and sheep groups. For instance, the *(rseD-rpoE-rseABC)–nadB* ITGR appeared to be particularly informative for the pig strains, as logic regression was able to identify a biomarker within this locus that was 30.00% sensitive and 97.99% specific to the pig group. Similarly, for the sheep strains, the *yjjP–(yjjQ-bglJ)* and *uspC–flhDC* loci were each found to encode biomarkers with sensitivities of 33.33% and specificities exceeding 98.00%. Amongst all the host categories, however, logic regression appeared to be most

effective for generating host-informative biomarkers for mouse-derived *E. coli* strains from single ITGR sequence data, as mouse-informative biomarkers exhibiting sensitivity values ranging between 33.33% to 66.67% and specificities from 97.26% to 100.00% were identified across multiple loci, including the *ompC–rcsDB*, *nanCMS–fimB*, *yedR–yedS,* and *csgDEFG–csgBAC* ITGR loci.

Although ecotype-informative biomarkers could be generated for each host and niche source using just single ITGR sequence data, additional steps were taken to improve the performance of the generated biomarkers. One strategy that was used to improve the efficacy of the biomarker discovery process was to utilise the sequence variation contained across multiple ITGRs at once. Indeed, previous studies have shown that the performance of produced logic models can be improved by appending multiple ITGR sequences together and generating biomarkers from the resulting concatenated sequences (Zhi *et al.* 2016b). Additionally, narrowing the source range interrogated, thereby reducing the extra 'noise' from extraneous host and niche sources in the sequence data analysed, was also hypothesized to improve biomarker performance.

Surprisingly, for single-ITGR logic regression analyses, reducing the source range alone did not necessarily improve the performance of generated biomarkers across all host categories. While the human and bovine single-ITGR biomarkers both exhibited significant improvements in performance, the performance parameters of the chicken and pig biomarkers were comparable across the biomarker discovery analyses. In contrast, concatenating multiple candidate ITGRs, including the *yedR–yedS, emrKY–evgAS*, *ibsB–(mdtABCD-baeSR),* and *ompC–rcsDB* intergenic loci, did appear to improve the performance of the generated biomarkers. Across all host categories, the biomarkers produced from the concatenated sequences were generally found to be more sensitive (i.e., from around 20.00–30.00% to up to 50.00% sensitivity), without any major

reductions in specificity.

Closer inspection of the genes flanking these intergenic regions revealed specific functions that could be biologically and ecologically relevant for strains colonizing human and livestock-associated animal hosts. Aside from the *rcsDB* locus, which acts as a master regulator for capsule biosynthesis (Wall *et al.* 2018), most of the flanking genes were primarily involved with antibiotic resistance, including: i) *yedS,* within the previously identified *yedR–yedS* locus, which appears to mediate resistance against carbapenems (Warner *et al.* 2013); ii) *emrKY,* within the *emrKY–evgAS* locus, which encodes for an efflux system that appears to be activated in response to tetracycline (Tanabe *et al.* 1997); iii) *evgAS*, within the *emrKY–evgAS* locus, which encodes for a two-component regulatory system that controls the expression of several antibiotic resistance genes in *E. coli* (Nishino *et al.* 2003); iv) *mdtABCD* and *baeSR*, within the *ibsB–(mdtABCD-baeSR)* locus, which appears to encode for a multidrug efflux pump and its corresponding two-component regulatory system respectively (Baranova and Nikaido, 2002); and v) *ompC*, within the *ompC–rcsDB* locus, which encodes for an outer membrane porin implicated in resistance to various antibiotics (Choi and Lee, 2019) as well as bile salts, and has been found to be required for colonizing the mammalian gut (Doranga and Conway 2023). Interestingly, these findings mirror previous analyses identifying antibiotic resistance-associated ITGRs as being particularly informative for human and bovine *E. coli* strains (Zhi *et al.* 2016b), suggesting that logic regression can identify key ITGRs that highlight the functional adaptations of strains belonging to different ecotypic groups. In this case, considering the elevated rates of antibiotic use in clinical (Llor and Bjerrum 2014) and agricultural (Manyi-Loh *et al.* 2018) settings, it appears that control over responses to antibiotic stress may be especially important for putative *E. coli* ecotypes adapted to human and livestock animal hosts.

Beyond *in silico* analyses, the ecotypic relevance of logic regression as an ecotype discovery approach was also demonstrated through its application for source attribution purposes. Indeed, applying logic regression to predict the host source of unknown *E. coli* isolates demonstrated that the biomarkers produced from logic regression can be correlated with the ecotypic group from which a given strain originated. Importantly, while other studies have used supervised learning algorithms for the source attribution of *E. coli* isolates recovered from various human and animal host species (Lupolova *et al.* 2017; Lupolova *et al.* 2019), our study is the first to extend their use for the source attribution of environmental isolates with no known host source. Remarkably, 48 out of the 113 environmental isolates could be successfully classified with a sufficient degree of consensus, and were determined to have originated from human, beaver, or reindeer. Although a significant proportion of the environmental isolates remained unclassified (i.e., representing potentially host-independent naturalized strains, host generalists, or specialists from other animal host sources), these findings indicate that the biomarkers produced from logic regression do have practical and ecotypic relevance.

Interestingly, the power of logic regression as an ecotype discovery and attribution tool was most strongly demonstrated through the characterization of wastewater-derived *E. coli* strains. From just single ITGR sequence data, the wastewater strains were found to possess highly informative biomarkers with sensitivities of up to 50.00% and specificities reaching 100.00%. Similarly, when assessed using concatenated ITGR sequences and against a reduced host range, the wastewater strains were found encode biomarkers with significantly higher sensitivities (i.e., up to 70.00%) and specificities (i.e., up to 100.00%) than those produced for the host groups. This is a remarkable finding considering that wastewater represents an effective ‘endpoint’ in the urban

water cycle and would thus be expected to contain a wide range of *E. coli* populations derived from human, animal, and environmental sources.

Indeed, these findings indicate that the selective pressures within the wastewater environment may be potent drivers of microbial evolution, such that the wastewater strains could serve as a particularly effective model system for understanding the evolution of *E. coli* towards niche-specificity. As previous studies have consistently distinguished these wastewater *E. coli* populations from their host-associated counterparts (Wang *et al.* 2020; Zhi *et al.* 2016a; Zhi *et al.* 2017), our findings provide additional evidence suggesting that these wastewater-derived strains may constitute a unique *E. coli* ecotype. To explore this hypothesis further, the following three chapters of this thesis will adopt a polyphasic perspective in describing the specific genotypic and phenotypic adaptations that may have allowed this wastewater-specific (WWS) *E. coli* ecotype to exploit wastewater as its primary niche.

# Chapter Three: Comparative Genomics and Ecotypic Characterization of Emerging Naturalized Lineages of *Escherichia coli* Across Food- and Water-Associated Engineered Environments[3]

## 3.1 Introduction

As demonstrated in the previous chapter, the evolution of niche-specificity in the *Escherichia coli* species appears to be recapitulated through the emergence of various ecotypes that have each evolved to become specifically adapted to a particular host or niche. While a variety of putative host-specific and environmentally-adapted *E. coli* ecotypes have been characterized to date (as reviewed in Chapter One of this thesis), the evolution of *E. coli* towards niche-specificity may be most strongly exemplified by distinct strains that appear to have evolved to exploit man-made environments as a primary niche. In the previous chapter, the application of logic regression for ITGR sequence analysis consistently distinguished wastewater-derived *E. coli* isolates from enteric strains derived from a wide range of human and animal host species. Remarkably, the wastewater strains were found to consistently harbor the best performing SNP biomarkers (i.e., achieving 70.00% sensitivity and 100.00% specificity) based on both single- and concatenated-ITGR sequence data. Considering this, the evidence suggests that wastewater-derived *E. coli* strains may constitute a distinct, non-host-associated *E. coli* ecotype.

Several studies by Zhi *et al*. (2016a; 2017; 2019) appear to support this hypothesis, as the

[3] A version of this chapter has been published as: Yu, D., Stothard, P., and Neumann, N.F. 2024. Emergence of potentially disinfection resistant, naturalized *Escherichia coli* populations across food- and water-associated engineered environments. Sci. Rep. 14(13478): 1–14. doi:10.1038/s41598-024-64241-y

authors were able to demonstrate that up to 82% of *E. coli* isolates collected from wastewater can be characterized by SNP-SNP biomarkers not found in any of their human-, animal-, or natural environment-derived counterparts. Interestingly, a subset of these wastewater strains were also found to harbor a unique *uspC*–IS30–*flhDC* biomarker that, at the time, was determined to be wastewater-specific (Zhi *et al.* 2022). Reflecting their distinct ecology, the WWS-*E. coli* strains harbored an abundance of stress resistance genes, including the transmissible locus of stress tolerance (tLST), that appear to confer resistance to the extreme stressors (i.e., disinfection) encountered during wastewater treatment (Zhi *et al.* 2019). Reflecting this, when compared to enteric and food-derived strains, the naturalized wastewater strains were also found to display an enhanced capacity for biofilm formation (Zhi *et al.* 2017), and were more resistant to heat, chlorine, and advanced oxidants (Wang *et al.* 2020).

The distinct genotypic, phenotypic, and ecotypic characteristics of these wastewater strains were originally thought to reflect specific adaptations acquired against the diverse stressors associated with wastewater treatment and sanitation practices (Zhi *et al.* 2019). Recently, however, Yang *et al*. (2021) isolated *E. coli* strains from a meat processing plant exhibiting several wastewater-associated genetic features, including the *uspC*–IS30–*flhDC* locus and the tLST, suggesting that additional *E. coli* populations may have become naturalized within other man-made environments. Thus, it is unclear whether these wastewater and meat plant *E. coli* populations represent distinct niche-specific ecotypes or reflect, more broadly, a standalone naturalized *E. coli* ecotype that has dispersed across various food- and water-associated engineered environments (i.e., a 'naturalized-engineered' ecotype). Herein, we demonstrate that naturalized wastewater and meat plant *E. coli* strains appear to be genomically, phylogenetically, and ecotypically distinct from their host- and natural environment-associated counterparts. Reflecting

their designation as distinct ecotypes, the wastewater and meat plant groups were also found to harbor several intragenic and intergenic insertion element markers that, alongside the *uspC*–IS30–*flhDC* biomarker (Zhi *et al.* 2022), appeared to be specific to the naturalized-engineered strains.

## 3.2 Materials and Methods

### 3.2.1 Screening of Presumptive Naturalized Wastewater and Meat Plant *E. coli* Strains from NCBI

All publicly available *E. coli* genomes representing presumptive naturalized wastewater-specific (WWS) and meat plant-specific (MPS) strains were identified in the NCBI GenBank database by screening for the *uspC*–IS30–*flhDC* target sequence (GenBank Accession Number: ON075843.1) using BLAST. *E. coli* strains bearing the full *uspC*–IS30–*flhDC* locus were then downloaded from NCBI GenBank (accessed: 01-26-2022) alongside additional previously-described naturalized wastewater and meat plant strains lacking the *uspC*–IS30–*flhDC* biomarker (Yang *et al.* 2021; Zhi *et al.* 2019). In addition to these strains, the assembled genome sequence of a lone naturalized wastewater strain isolated in China (SZ4) was also contributed separately through a collaboration with Dr. Shuai Zhi. For comparative purposes, the genome sequences of representative *E. coli* strains across phylogroups, lifestyles (i.e., commensal, intestinal pathogenic, extraintestinal pathogenic, environmental), and isolation source (i.e., host and environmental niches), as well as *Escherichia* strains across the cryptic clades (i.e., as an additional environmentally-adapted group of *Escherichia* strains for comparison) were also downloaded from NCBI GenBank. All information related to the bacterial strains evaluated in this chapter can be found in Supplementary Table 3–S1.

### 3.2.2 Comparative Genomics of Naturalized Wastewater and Meat Plant *E. coli* Strains Against Other Ecotypes

As an initial assessment of the genetic relationships between the naturalized-engineered (i.e., including both WWS- and MPS-*E. coli*) strains with strains belonging to other, predominantly host-associated, *E. coli* ecotypic groups and the cryptic *Escherichia* clades, an average nucleotide identity (ANI) analysis was performed. Specifically, the ANI shared between each strain across all putative ecotypes were calculated in a pairwise fashion (i.e., between two strains at a time) using fastANI v1.33 (Jain *et al.* 2018) to produce an ANI similarity matrix (Supplementary Table 3–S2). The pairwise ANI values shared between strains belonging to the same ecotypic group were then pooled and averaged to generate 'within-group' ANI values for each ecotype, which were then compared statistically using T-tests. Furthermore, to assess whether the calculated ANI values could be correlated with putative naturalized-associated lineages, the average ANI values associated with the naturalized-engineered strains were also compared to the average within-group ANI values calculated for additional groups of reference *E. coli* strains that were downloaded from NCBI (Supplementary Table 3–S1), including those belonging to pathogenic *E. coli* lineages characterized by a defined pathogenic presentation (i.e., O157:H7 EHEC strains, and ST131 and ST95 ExPEC strains).

### 3.2.3 Core Genome Phylogenetics and Typing of Naturalized *E. coli* Strains

To assess whether the genomic differences between the naturalized-engineered strains and other ecotypic groups could be reflected in phylogeny, a phylogenetic analysis was performed.

The genome sequences of a representative subset of *E. coli* strains from each phylogroup and ecotype (Supplementary Table 3–S1), alongside representative *Escherichia* strains across the cryptic clades, were annotated with Prokka v1.14.6 (Seemann 2014). Roary v3.13.0 (Page *et al.* 2015) was then used to produce a core genome alignment for the generation of a phylogenetic tree using the maximum likelihood algorithm with RAxML v8.2.12 (Stamatakis 2014), with *E. albertii* as the outgroup. The phylogroups of the strains included in the phylogenetic tree were determined through the ClermonTyping method, using the ClermonTyper v23.06 webserver (Beghain *et al.* 2018). Multilocus sequence typing (MLST) was performed using mlst v2.22.0 (https://github.com/tseemann/mlst) with the *Escherichia coli* #1 scheme, and then cross-referenced against the PubMLST database using the Achtman scheme (Jolley *et al.* 2018). The serotypes of each strain were determined with ABRicate v1.0.1 (https://github.com/tseemann/abricate) using the EcOH database. The phylogenetic tree was visualized and annotated with R software, using the R packages ggplot2 v3.4.2 (Wickham 2009), ape v5.4.1 (Paradis and Schliep 2019), and ggtree v2.2.4 (Yu *et al.* 2017).

### 3.2.4 Ecotype Prediction of Naturalized Wastewater and Meat Plant *E. coli* Strains

To determine whether the genomic and phylogenetic characteristics of the naturalized wastewater and meat plant strains could be reflected in their classification into distinct naturalized *E. coli* ecotypes, ecotype prediction analyses were performed. Putative ecotypes represented across the *E. coli* strains analyzed were predicted using two methods. First, a phylogenetics-based approach was employed using the Ecotype Simulation 2 algorithm (Wood *et al.* 2020). To maintain consistency with the analyses described above, the algorithm was run using the core-genome maximum-likelihood phylogenetic tree produced by RAxML and the core genome alignment

produced by Roary, such that all ecotype prediction analyses were performed on the same set of strains used for the phylogenetic analysis.

While Ecotype Simulation identifies putative ecotypes through partitioning phylogenies into ecologically-defined groups, machine learning methods have also shown promise for clustering bacterial isolates according to their ecological niche using readily-available genome sequence data (Lupolova *et al.* 2019). Specifically, supervised learning approaches, which attempt to uncover patterns in data correlating specifically with observed data labels (i.e., ecological source), appear to be particularly useful for clustering strains into putative ecotypes from a pooled sample. As such, logic regression (Ruczinski *et al.* 2003), which was found in the previous chapter to be effective at clustering *E. coli* isolates according to their original ecological source, was used as an alternative, phylogeny-independent method to assess whether the naturalized-engineered strains could be distinguished as distinct naturalized (i.e., host-independent) *E. coli* ecotypes. Following previous workflows (Zhi *et al.* 2015; Zhi *et al.* 2016b), the *asnS–ompF* and *csgDEFG–csgBAC* intergenic sequences were screened from all strains included in the analysis using BLAST (Camacho *et al.* 2009), extracted using bedtools v2.30.0 (https://github.com/arq5x/bedtools2), and then aligned with Clustal Omega (Sievers *et al.* 2011). Using a custom R script, the aligned intergenic sequences were analyzed using logic regression to identify key SNP-SNP patterns that could classify the strains as either naturalized or non-naturalized. Five random seed numbers were generated with R such that this classification step could be performed over five independent trials, after which the results from each classification iteration were pooled.

### 3.2.5 Identification and Evaluation of Naturalized-Specific Intergenic Sequence Element Biomarkers

From its original discovery by Zhi *et al.* (2016a), the *uspC*–IS30–*flhDC* locus was initially determined to be a highly specific genetic marker for identifying WWS-*E. coli* strains (Zhi *et al.* 2019; Zhi *et al.* 2022). Although this biomarker has since been found in presumptive MPS-*E. coli* strains (Yang *et al.* 2021), the *uspC*–IS30–*flhDC* locus still appears to be unique to naturalized strains derived from engineered niches. To assess whether the naturalized-engineered strains could be characterized by additional ecotype-informative insertion element-associated genetic markers, including those that could potentially differentiate between the WWS- and MPS-*E. coli* populations, a mobile genetic element (MGE) screening analysis was performed, with a focus on insertion element sequences.

The raw reads of the naturalized WWS- and MPS-*E. coli* strains were downloaded from the NCBI Sequence Read Archive (SRA) database (accessed: 10-20-2022). The FASTQ files were deduplicated using the SuperDeduper algorithm (Petersen *et al.* 2015) provided within the HTStream toolset (https://github.com/s4hts/HTStream), and then trimmed using TrimGalore v0.6.7 (https://github.com/FelixKrueger/TrimGalore). Two host-associated *E. coli* strains, *E. coli* HS and *E. coli* Fec6, were then selected as host-derived reference strains for the identification of any insertion element markers associated with the naturalized-engineered ecotype(s). The two reference genome sequences were first indexed using the Burrows-Wheeler Aligner (BWA) package (Li and Durbin 2009), following which the raw reads of the naturalized strains were aligned against the indexed reference sequences using the BWA-MEM algorithm. MGEfinder v1.0.6 (Durrant *et al.* 2020) was then used to identify any candidate insertion element markers, including the identity of the specific insertion sequence and their genomic insertion sites.

Putative insertion element markers that were identified using the two reference genome sequences were pooled and characterized as either 'intergenic', if the insertion element was integrated within the intergenic region between two genes, or 'intragenic', if the insertion element disrupted the coding sequence of a gene. To evaluate the ecological relevance of these insertion events, the relevant flanking (i.e., for intergenic insertions) or disrupted (i.e., for intragenic insertions) genes were functionally annotated after reference against the UniProt (Bateman *et al.* 2017) and EcoCyc (Karp *et al.* 2018) databases. Finally, to assess whether the identified insertion element markers could serve as naturalized-specific genetic markers, the sequences of each intergenic and intragenic insertion marker were extracted using bedtools v2.30.0 (https://github.com/arq5x/bedtools2) and then BLAST against all publicly-available *E. coli* genomes that had been sequenced to date on NCBI GenBank (accessed: 01-26-2024).

## 3.3 Results

### 3.3.1 Screening of Presumptive Naturalized Wastewater and Meat Plant *E. coli* Strains from NCBI

A total of 36 presumptive naturalized WWS- and MPS-*E. coli* strains were screened and downloaded from NCBI GenBank (Supplementary Table 3-S1), alongside an additional lone naturalized wastewater strain (SZ4) isolated from China that was sequenced and provided separately. Of the 37 total naturalized-engineered strains identified, 20 represented WWS-*E. coli* strains while 17 represented MPS-*E. coli* strains. Of these, 16 out of the 20 WWS strains and 11 out of the 17 MPS strains were found to harbor the *uspC*–IS30–*flhDC* biomarker. In line with previous findings (Zhi *et al.* 2019), the WWS-*E. coli* strains appeared to be globally distributed, with strains collected from wastewater treatment plants across Canada, the United States,

Switzerland, the United Kingdom, and China. In contrast, all MPS-*E. coli* strains were collected from a single meat processing facility within Canada, suggesting that the level of clonal representation may be higher for the meat plant strains than the wastewater strains.

### 3.3.2 Comparative Genomics of Naturalized Wastewater and Meat Plant *E. coli* Strains with Other Ecotypes

Building on previous work highlighting the distinct genotypic, phenotypic, and ecotypic characteristics of the wastewater (Wang *et al.* 2020; Zhi *et al.* 2016a; Zhi *et al.* 2017; Zhi *et al.* 2019) and meat plant (Yang *et al.* 2021) strains, comparative genomic approaches were used as an initial assessment into the genetic relationships between the naturalized-engineered strains and other, predominantly host-associated, ecotypes. Despite their distinct ecology, the naturalized-engineered strains expectedly shared ≥ 95% ANI with all other *E. coli* strains and the cryptic *Escherichia* clade I strain (Supplementary Table 3-S2), exceeding the cut-off for bacterial strains belonging to the same species (Jain *et al.* 2018; Konstantinidis and Tiedje 2005). In contrast, all *E. coli* strains shared less than 95% ANI with *Escherichia* strains belonging to cryptic clades II–V, providing support for previous proposals that these cryptic clades could represent novel, environmentally-adapted *Escherichia* species (Walk 2015; Beghain *et al.* 2018).

Amongst the other *E. coli* ecotypes, both the wastewater and meat plant strains were found to exhibit higher within-group ANI (i.e., amongst other wastewater and/or meat plant strains) than when compared to strains belonging to other ecotypes. Reflecting this, the wastewater strains were found to exhibit an average within-group ANI of 99.465% ± 0.551%, which was significantly higher ($p << 1E-5$) than the similarity shared with the other, non-naturalized-engineered strains, which ranged from as low as 96.702% ± 0.148% with NMEC strains to only as high as 98.314%

± 0.202% with ETEC strains (Figure 3-1A). Similarly, while the average ANI shared between meat plant strains and other ecotypic groups ranged from 96.625% ± 0.147% with NMEC to 98.265% ± 0.133% with ETEC, the within-group genomic similarity was significantly greater ($p << 1E-5$) at 99.707% ± 0.298% (Figure 3-1B). Interestingly, while the WWS- and MPS-*E. coli* groups were found to share 99.367% ± 0.343% whole genome similarity, it was still found to be significantly lower than the within-group similarity shared amongst the WWS ($p < 0.05$) and MPS ($p << 1E-5$) strains independently (Figure 3-1).

While the wastewater and meat plant strains shared a high degree of within-group genomic similarity, it is unclear whether these ANI values reflect the grouping of strains that can be defined by a common ecological niche or lifestyle (i.e., their ecotypic relevance). To better contextualize the degree of genomic similarity shared amongst the wastewater and meat plant strains, the within-group ANI of the naturalized-engineered strains were compared to the average ANI values shared by strains belonging to other prominent ecotypic groups or lineages in the *E. coli* species. Overall, both the wastewater and meat plant strains consistently exhibited higher within-ecotype genomic similarity than the other *E. coli* ecotypes, which ranged from 96.796%–98.322% for enteric strains, 97.689%–98.806% for ExPECs, 98.392% for laboratory reference strains, and 98.579% for environmental strains (Figure 3-2). In contrast, the within-group ANI values of the WWS- and MPS-*E. coli* strains were found to be comparable to that observed for several prominent *E. coli* 'lineages' with a defined ecology or lifestyle, including O157:H7 strains (99.534% ± 0.753%) associated with the EHEC pathotype, as well as ST131 (99.320% ± 1.010%) and ST95 (99.653% ± 0.128%) strains associated with the ExPEC pathotypes.

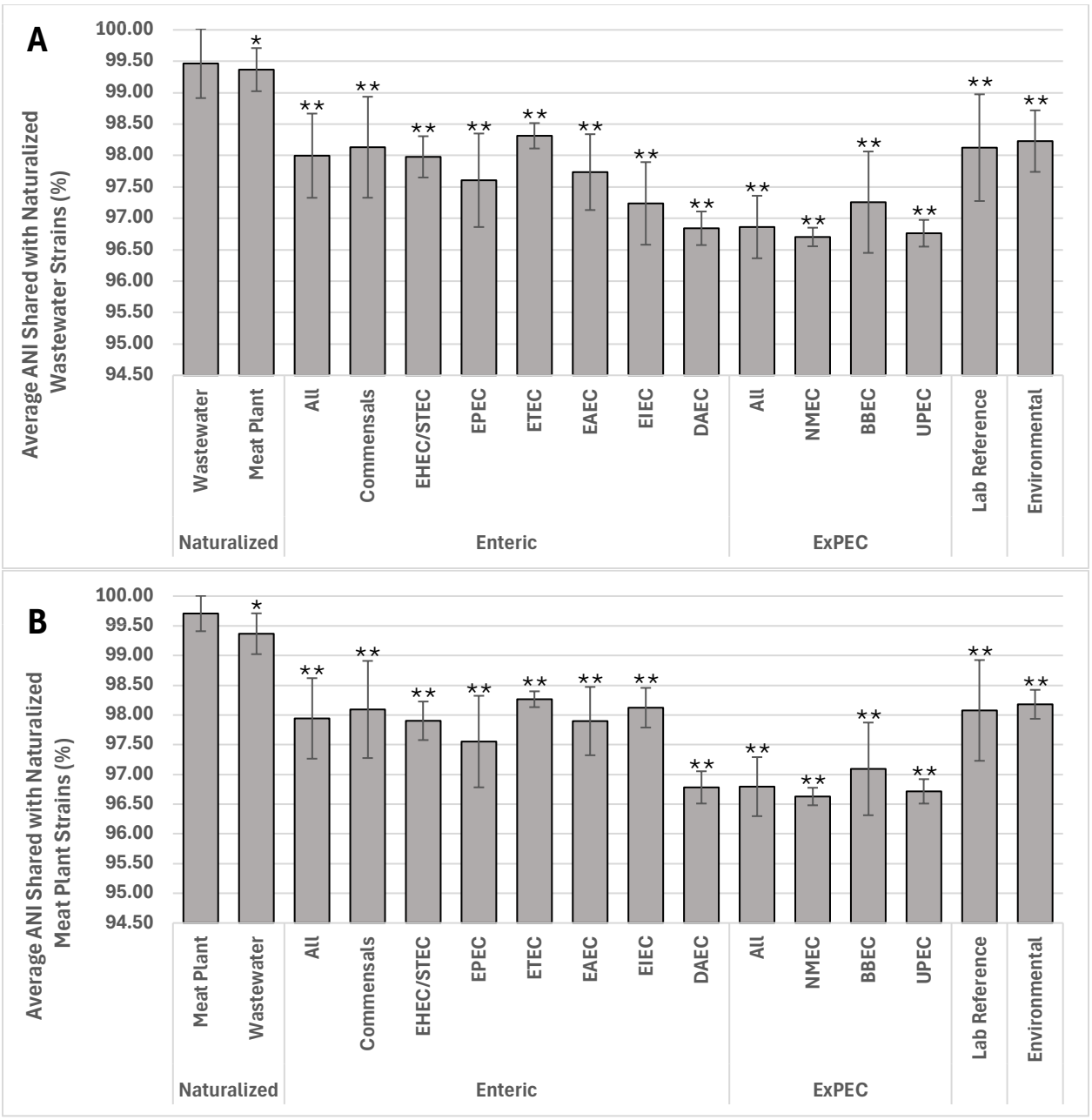



**Figure 3-1. Comparison of the average within-group ANI shared amongst naturalized wastewater and meat plant *E. coli* strains with the between-group ANI shared with other *E. coli* ecotypes.** FastANI was used to calculate the pairwise ANI values shared between each E. coli strain across all E. coli ecotypes in a pairwise fashion. The ANI values were then pooled according to each ecotypic group to calculate the average ANI shared amongst the **(A)** naturalized wastewater and **(B)** naturalized meat plant strains, respectively, and the average ANI that these naturalized groups shared with strains belonging to other *E. coli* ecotypic groups (i.e., enteric, ExPEC, lab reference, environmental). These average ANI values were then compared to each naturalized group's average within-group ANI using T-tests (*: $p < 0.05$; **: < 1E-5), to evaluate the degree of genomic similarity that exists between the naturalized strains and other *E. coli* strains belonging to other ecotypic groups.

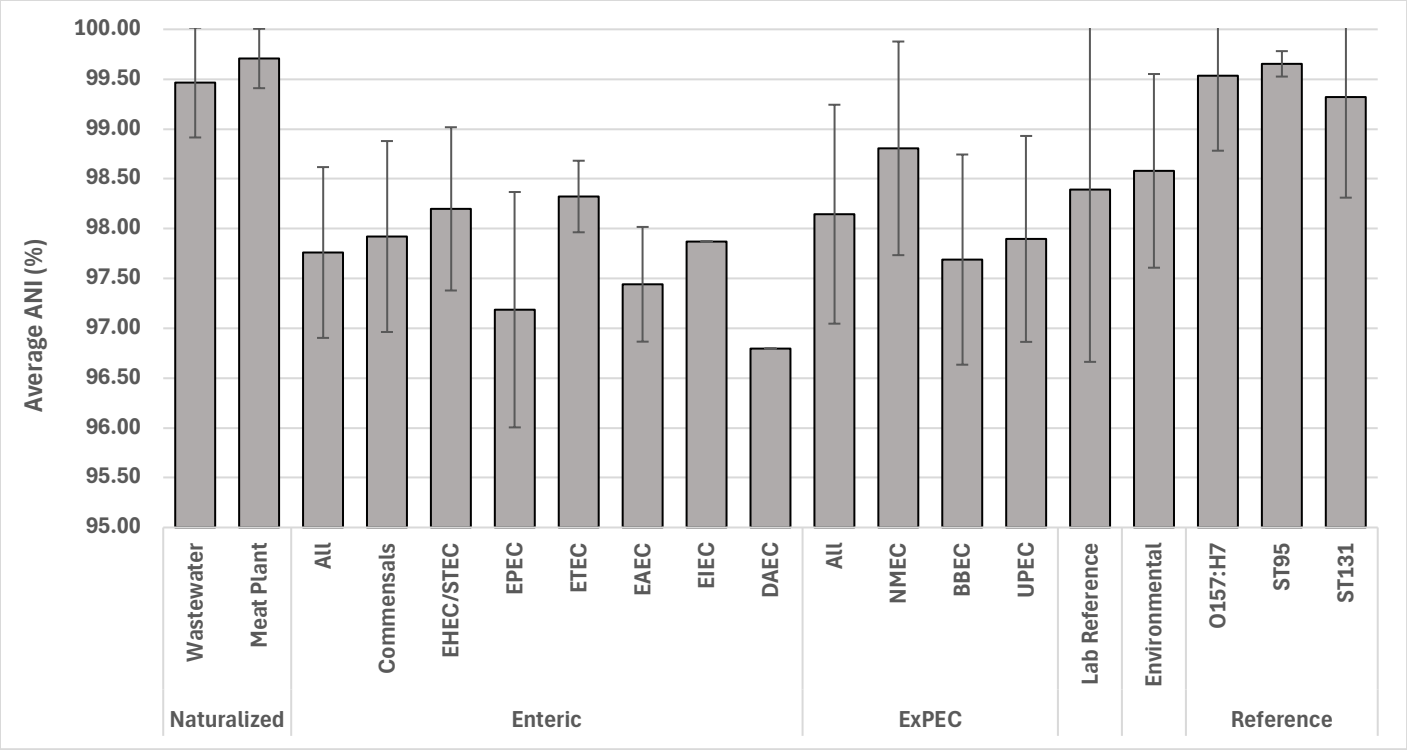



**Figure 3-2. Comparison of average within-group ANI values across *E. coli* ecotypes and lineages.** FastANI was used to calculate the pairwise ANI values shared between each strain across each *E. coli* ecotype (i.e., naturalized wastewater, naturalized meat plant, enteric, ExPEC, laboratory, environmental) and reference lineage (i.e., O157:H7, ST95, ST131) in a pairwise fashion. The ANI values were then pooled according to each ecotypic group/lineage to calculate the average ANI shared amongst *E. coli* strains belonging to the same ecotype or lineage to better evaluate whether the ANI values shared amongst the naturalized wastewater and meat plant strains could reflect shared ecotypic properties.

### 3.3.3 Phylogenetics and Typing of Naturalized Wastewater and Meat Plant *E. coli* Strains

The finding that the naturalized-engineered groups exhibited ANI values that were comparable to the average ANI values of other *E. coli* lineages with well-defined ecologies and lifestyles (i.e., pattern of pathogenesis) suggests that the wastewater and meat plant strains could represent distinct, novel *E. coli* lineages that have naturalized specifically within engineered environments. To assess the emergence of these potential naturalized-engineered *E. coli* lineages, a core-genome, maximum-likelihood phylogenetic analysis was performed with the 37 WWS and MPS strains alongside 45 representative *E. coli* strains across lifestyles (i.e., commensal, intestinal pathogenic, extraintestinal pathogenic, laboratory, environmental) and phylogroups, 5 *Escherichia* strains across the cryptic clades, and an *Escherichia albertii* strain as the outgroup.

All wastewater and meat plant strains were found to cluster within phylogroup A (Figure 3-3), with most grouping within a monophyletic clade separate from their host- and natural environment-derived counterparts. Aside from the wastewater strains SZ4 and WW38, which grouped closest to the InPEC strains ETEC_H101407 and 53638 respectively, the rest of the wastewater and meat plant strains formed a separate cluster that was largely exclusive to the naturalized-engineered strains except for the inclusion of Fec6, a presumptive human commensal isolate recovered from a fecal swab sample. Despite clustering exclusively within a single phylogroup, the naturalized-engineered strains were distributed across multiple sequence types (STs). Twenty-seven naturalized-engineered strains were designated as ST635, followed by 8 that were designated as ST399, and one wastewater strain each designated as ST216 and ST48 (Table 3-1).

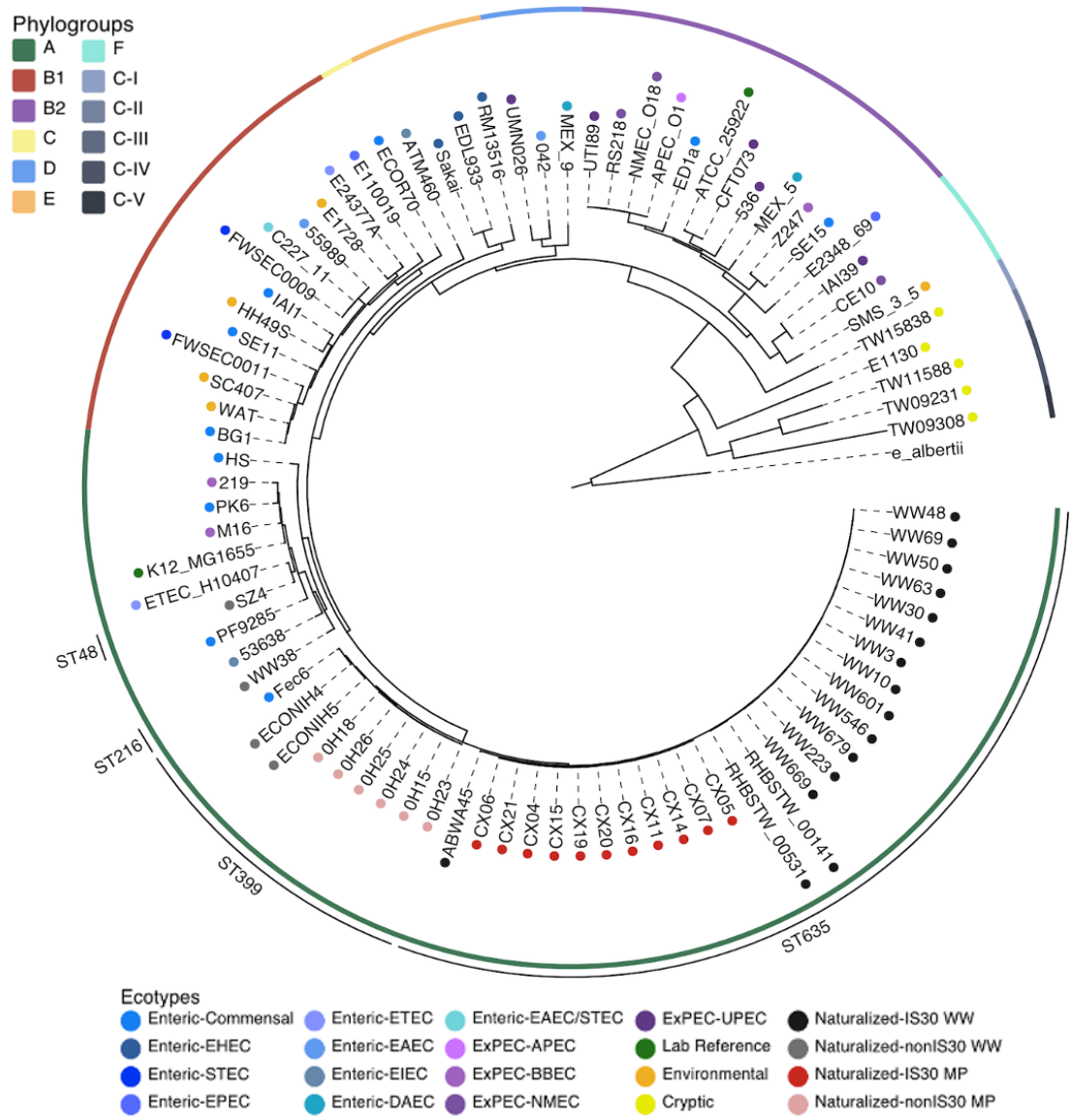



**Figure 3-3. Core genome maximum likelihood phylogenetic tree of naturalized wastewater and meat plant strains alongside other strains representative of the *E. coli* species and the cryptic *Escherichia* clades.** Naturalized wastewater and meat plant *E. coli* strains were screened and downloaded from NCBI. To evaluate the evolutionary history of these strains, the core genome sequence variation of the naturalized wastewater (black and grey circles corresponding to those possessing ['IS30 WW'] and lacking the *uspC*–IS30–*flhDC* biomarker ['nonIS30 WW'], respectively) and meat plant (red and pink circles corresponding to those possessing ['IS30 MP'] and lacking the *uspC*–IS30–*flhDC* biomarker ['nonIS30-MP'], respectively) strains were compared to enteric (blue circles), ExPEC (purple circles), lab reference *E. coli* (green circles), environmental *E. coli* (orange circles), and cryptic *Escherichia* (yellow circles) strains. Phylogroups in the phylogenetic tree are indicated by the inner ring and colored according to the upper legend. The main sequence types represented across the naturalized strains are indicated in the outermost ring. The tree is rooted against an *E. albertii* strain as the outgroup.

**Table 3-1.** Distribution of sequence types and serotypes across the naturalized wastewater and meat plant *E. coli* strains.

| Strains | Ecotype | Location of Isolation | Sequence Type | Serotype |
|---|---|---|---|---|
| WW10, WW223, WW3, WW30, WW41, WW48, WW40, WW546, WW601, WW63, WW669, WW679, WW69 | Naturalized Wastewater | Canada | ST635 | O11:H25 |
| ABWA45 | | Switzerland | | O9/27:H7 |
| RHBSTW_00141, RHBSTW_00531 | | England | | O166:H25 |
| CX04, CX05, CX06, CX07, CX11, CX14, CX15, CX16, CX19, CX20, CX21 | Naturalized Meat Plant | Canada | | O10:H25 |
| ECONIH4 | Naturalized Wastewater | USA | ST399 | O18/129_13_gp10:H30 |
| ECONIH5 | | | | O166:H30 |
| 0H15, 0H18, 0H23, 0H24, 0H25, 0H26 | Naturalized Meat Plant | Canada | | O154:H12 |
| WW38 | Naturalized Wastewater | Canada | ST216 | O11:H4 |
| SZ4 | | China | ST48 | O64:H20 |

*flhDC* locus. Indeed, the ST635 lineage included all naturalized-engineered strains carrying the *uspC*–IS30–flhDC locus while the ST399 cluster mainly consisted of naturalized-engineered strains lacking the biomarker.

Serotyping revealed further sub-structuring, with several serotypes represented across the larger ST clusters identified. For instance, while ST635 contained all wastewater and meat plant strains positive for the *uspC*–IS30–*flhDC* locus, four different serotypes were represented, appearing to coincide with the original geographical source of isolation of the naturalized-engineered strains. This included O11:25 for most of the Canadian wastewater strains, O166:H25 for the U.K. wastewater strains, O9/O27:H7 for the lone Swiss wastewater strain, and O10:H25 for the Canadian meat plant strains (Table 3-1). Similarly, amongst the *uspC*–IS30–*flhDC*–negative naturalized strains comprising the ST399 clade, the rest of the Canadian meat plant strains were assigned the serotype O154:H12, whereas the two U.S. wastewater isolates ECONIH4 and ECONIH5 were designated as O8/O129_13_gp10:H30 and O166:H30, respectively. The two divergent wastewater strains, SZ4 and WW38, were also found to belong to unique serotypes as they were designated as O64:H20 and O11:H4, respectively. Notably, none of the naturalized-engineered strains belonged to the same serotype as any of the host- or natural environment-associated strains included in the phylogenetic tree (Supplementary Table 3-S1).

### 3.3.4 Ecotype Prediction with Naturalized Wastewater and Meat Plant *E. coli* Strains

To evaluate whether the phylogenetically distinct wastewater and meat plant *E. coli* populations could also represent distinct *E. coli* ecotypes, ecotype prediction analyses were performed. Given their unique phylogenetic clustering, a phylogeny-based ecotype prediction

algorithm, Ecotype Simulation 2, was used first. Through this phylogeny-based approach, two naturalized-engineered ecotypes, 'Ecotype0004' and 'Ecotype0005', were identified corresponding to two main clusters of naturalized-engineered strains (Figure 3-4A). Interestingly, these ecotypes did not correspond with the original ecological niche from which the strains were isolated (i.e., wastewater versus meat plants), but rather reflected the division of strains based on the presence of the *uspC*–IS30–*flhDC* locus – where 'Ecotype0004' corresponded to the ST635 cluster consisting of wastewater and meat plant strains positive for the biomarker, while 'Ecotype0005' corresponded to the ST399 cluster of wastewater and meat plant strains lacking the biomarker and the enteric strain Fec6. Consequently, the two divergent wastewater strains, SZ4 and WW38, were not assigned to either of the two predicted naturalized ecotypes, but instead clustered into a separate group, 'Ecotype0006' (Figure 3-4A).

As an alternative ecotype prediction method, logic regression was also used to cluster the naturalized strains into putative ecotypes in a phylogeny-independent manner. Based on the sequence variation within the *asnS–ompF* and *csgDEFG–csgBAC* intergenic regions, most wastewater and meat plant strains could be clustered into a single naturalized-engineered *E. coli* ecotype despite their differing sources of isolation (Figure 3-4B). Across 5 independent classification trials, 35 of the 37 wastewater and meat plant strains were consistently determined to belong to the putative naturalized-engineered *E. coli* ecotype. Notably, the ST399 enteric strain Fec6 was also consistently classified as part of the putative naturalized-engineered ecotype, whereas the remaining two wastewater strains, SZ4 and WW38, were only classified as naturalized on a trial-by-trial basis in 3 and 2 of the 5 classification trials, respectively. Interestingly, the classifications of SZ4 and WW38 appeared to be mutually exclusive, as in any one trial only one of the two strains were classified as naturalized by logic regression.

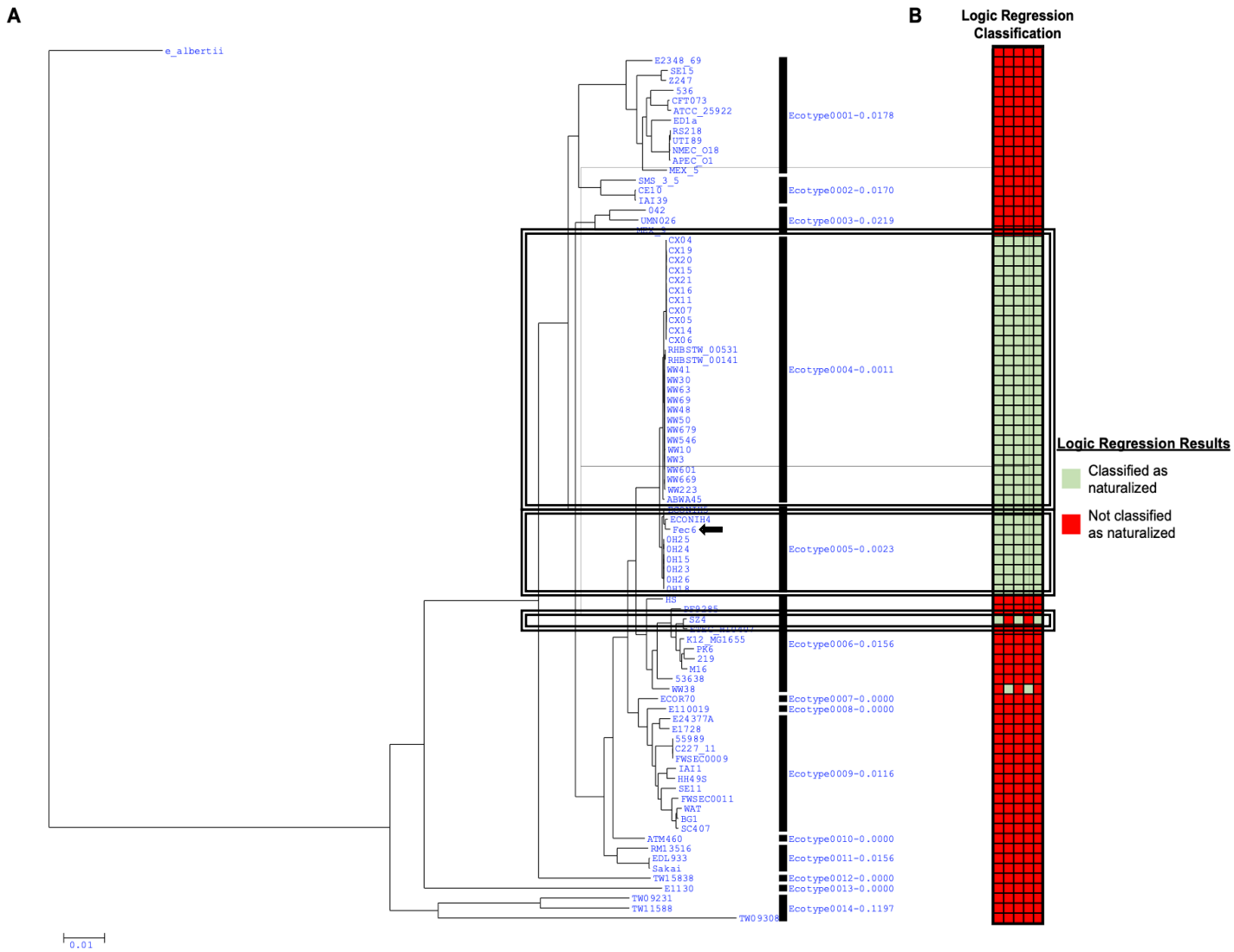



**Figure 3-4. Prediction of putative naturalized *E. coli* ecotypes.** Identification of putative naturalized *E. coli* ecotypes using a phylogeny-based (i.e., unsupervised) approach via the Ecotype Simulation 2 algorithm and a supervised learning approach via logic regression. According to the **(A)** phylogeny-based approach, two putative naturalized-associated ecotypes, labelled 'Ecotype0004' and 'Ecotype0005', were identified; however, two wastewater strains, SZ4 and WW38, were not identified as belonging to either of these ecotypes through this approach. In contrast, across 5 classification trials, the **(B)** logic regression-based approach clustered most naturalized strains (indicated with the black boxes) into a singular naturalized ecotype, and was also able to correctly classify SZ4 and WW38 on a case-by-case basis. Regardless of the ecotype prediction approach used, however, the enteric strain Fec6 (indicated with an arrow) was included in the naturalized ecotype(s) predicted.

### 3.3.5 Screening of Naturalized-Specific Intergenic Sequence Element Biomarkers

The close association between the *uspC*–IS30–*flhDC* locus and the ST635 lineage described above suggests that the naturalized-engineered ecotype may be characterized by additional genetic markers that might distinguish them from other ecotypic groups within the *E. coli* species. Moreover, considering that insertion elements have been documented to play an important role in facilitating bacterial evolution and niche adaptation (Consuegra *et al.* 2021; Durrant *et al.* 2020), there may be additional insertion element-related biomarkers indicative of *E. coli* niche-specificity that can be used to differentiate between the wastewater and meat plant groups. As such, a screening analysis was performed to assess whether the WWS- and MPS-*E. coli* strains harbored other ecotype-informative, potentially niche-specific, insertion element-related genetic markers.

Overall, 85 total putative naturalized-engineered-associated insertion element-related genetic markers were identified. Of these, 44 insertion markers were integrated at sites located between genes and were thus determined to be intergenic (Supplementary Table 3-S3), whereas 41 insertion events resulted in the disruption of a gene and were thus found to be intragenic (Supplementary Table 3-S4). Although a select number of insertion element markers involved hypothetical proteins with no known function, the majority of the identified insertion events occurred within or between genes with annotated functions. Notably, a subset of the intragenic and intergenic insertion markers appeared to be ecotype-informative, as they involved genes mediating functions that could facilitate the naturalization process. Several intergenic insertion element markers, for instance, were found to be integrated between genes that mediated functions relevant for survival outside of the host environment, including biofilm formation, microbial defense mechanisms (i.e., toxin-antitoxin systems, CRISPR-Cas systems), and stress resistance against

various environmental stressors such as DNA-damaging stimuli, oxidative stress, osmotic stress, and heavy metals (**Table 3-2.6**Table 3-2). In contrast, the majority of the ecotype-informative intragenic insertion element markers identified in the naturalized-engineered strains were found to disrupt genes related to colonization, virulence, and host-associated stress resistance (i.e., nutritional stress resistance, acid resistance) – functions that would otherwise be adaptive within the host gastrointestinal environment (Table 3-3).

Consistent with their apparent functional relevance to a naturalized lifestyle, the majority of the intergenic (Table 3-2) and intragenic (Table 3-3) insertion element markers were found to be specific to the naturalized-engineered strains. Indeed, screening of the candidate insertion markers sequences against NCBI databases revealed that, out of all publicly available *E. coli* genome sequences that have been sequenced to date, only strains derived from engineered contexts (i.e., hospital wastewater, municipal wastewater, meat processing plants) appeared to harbor most of the identified naturalized-engineered-associated insertion element biomarkers. Interestingly, while many of these insertion markers were found to be distributed across both wastewater and meat plant strains, a select number appeared to be unique to either the wastewater or meat plant groups specifically.

## 3.5 Discussion

The diversification of the *E. coli* species across different niche-specific ecotypes has primarily been explored through the restriction of different strains to various host species or natural environments; however, the evolution of *E. coli* niche-specificity may be most evident through the adaptation of distinct strains to man-made, engineered environments. Reflecting this, the recovery of distinct *E. coli* populations inhabiting wastewater treatment plants (Zhi *et al.* 2016a; Zhi *et al.*

**Table 3-2.** Ecotype-informative and niche-specific intergenic insertion element genetic markers identified in the naturalized-engineered strains

| Insertion Element Marker | Distribution Across Naturalized-Engineered Strains | Prevalence in Wastewater Strains | Prevalence in Meat Plant Strains | Associated Functions | Specific to Naturalized-Engineered Strains? |
|---|---|---|---|---|---|
| *hns*–IS2–*tdk* | Universal | 17/20 | 17/17 | Microbial Defense (cleavage of foreign DNA)<br><br>Stress Resistance (osmotic stress, acid stress) | N |
| *uspC*–IS30–*flhDC* | ST635-specific | 17/20 | 11/17 | Motility (flagellar biosynthesis)<br><br>Stress Resistance (DNA-damaging stimuli) | Y |
| *uspC*–ISEc33–*flhDC* | ST399-specific | 0/20 | 6/17 | | |
| *lsrK*–IS5–[CRISPR Protein] | ST635-specific | 15/20 | 11/17 | Adhesion and Biofilm Formation<br><br>Homeostasis (signaling, quorum sensing)<br><br>Microbial Defense (CRISPR-Cas system) | Y |
| *yjjP*–(IS1222-ISSen4)–*yjjQ* | ST635-specific | 15/20 | 11/17 | Motility (flagellar biosynthesis)<br><br>Stress Resistance (detoxification, oxidative stress)<br><br>Pathogenesis (capsule biosynthesis) | Y |
| *nupC*–IS186B–*pdeA* | ST635-specific | 15/20 | 11/17 | Stress Resistance (antibiotics, oxidative stress, reactive nitrogen species)<br><br>Pathogenesis (survival against macrophages) | N |
| *evgS*–(ISKpn43-IS903B-ISKpn43)–*yfdE* | ST635-specific | 11/20 | 11/17 | Stress Resistance (acid stress, antibiotics, heavy metals) | Y |
| *mgtS*–IS5–*dgcZ* | Wastewater-specific | 15/20 | 0/17 | Adhesion and Biofilm Formation (production of biofilm exopolysaccharides)<br><br>Motility (repression of swimming motility, flagella) | Y |

| | | | | | |
|---|---|---|---|---|---|
| *ygcB*–IS186B–*hokE* | Wastewater-specific | 12/20 | 0/17 | Microbial Defense (CRISPR-Cas system, toxin-antitoxin system) | Y |
| *yhiM*–ISEc78–*pcoE* | Wastewater-specific | 11/20 | 0/17 | Stress Resistance (acid stress, heavy metals) | Y |
| *caiT*–IS1R–*fixA* | Meat plant-specific | 0/20 | 11/17 | Stress Resistance (heat shock, cold shock, pressure stress, etc.) | Y |
| *sbmC*–ISKox3–*dacD* | Meat plant-specific | 0/20 | 6/17 | Microbial Defense (toxin neutralization)<br><br>Stress Resistance (antibiotics, DNA-damaging stimuli) | N |

**Table 3-3.** Ecotype-informative and niche-specific intragenic insertion element genetic markers identified in the naturalized-engineered strains

| Insertion Element Marker | Distribution Across Naturalized-Engineered Strains | Prevalence in Wastewater Strains | Prevalence in Meat Plant Strains | Associated Functions | Specific to Naturalized-Engineered Strains? |
|---|---|---|---|---|---|
| IS30::*ycjM* | Universal | 17/20 | 17/17 | Colonization<br><br>Stress Resistance (salt stress, nutritional stress) | N |
| IS30::*glnP* | Universal | 17/20 | 11/17 | Colonization<br><br>Stress Resistance (acid stress) | Y |
| IS30::*ynfF* | Universal | 17/20 | 11/17 | Metabolism (sulfur metabolism, nitrogen metabolism)<br><br>Respiration (electron transport – terminal reductase) | Y |
| IS1222-ISSen4::*yhbX* | ST635-specific | 16/20 | 11/17 | Homeostasis (cell wall homeostasis, LPS biosynthesis)<br><br>Virulence | Y |
| ISKpn26::*arnC* | Wastewater-specific | 16/20 | 0/17 | Stress Resistance (antibiotics) | Y |
| IS2::*cntO* | Meat plant-specific | 0/20 | 11/17 | Colonization (metal acquisition – nickel, zinc) | Y |
| IS903B::*evgS* | Meat plant-specific | 0/20 | 6/17 | Stress Resistance (acid stress, antibiotics) | Y |
| IS903B::*frc* | Meat plant-specific | 0/20 | 6/17 | Stress Resistance (acid stress) | Y |
| ISLead2::*fhuA* | Meat plant-specific | 0/20 | 6/17 | Colonization (metal acquisition – iron) | Y |

2017; Zhi *et al.* 2019) and meat processing facilities (Yang *et al.* 2021), suggests that additional *E. coli* ecotypes have emerged across various food- and water-associated engineered niches. This is a remarkable prospect, given that these engineered environments represent relatively recent technological advances in human history. Indeed, the evolutionary emergence and success of the WWS- and MPS-*E. coli* strains serve as a prominent reminder of the power of natural selection in microbial evolution and the ability of microbes to exploit almost any environmental niche.

The possibility that the wastewater and meat plant strains could represent novel naturalized-engineered *E. coli* ecotypes was first reflected through ANI analyses. Typically, ANI is used to demarcate the level of genomic similarity defining whether two strains could belong to the same species (Jain *et al.* 2018). While ANI values have not yet been used to define ecotypic boundaries, other studies have leveraged ANI to benchmark taxonomic groups below the species level (Pearce *et al.* 2021). Furthermore, from a conceptual standpoint, strains belonging to the same ecotype (i.e., that share a common ecological niche or role) would be expected to exhibit higher degrees of genomic similarity than when compared to strains belonging to different ecotypic groups. Reflecting this, the wastewater and meat plant strains appeared to be genetically distinct from other *E. coli* ecotypes as they were found to consistently share significantly higher within-group ANI similarity than commensal, InPEC, ExPEC, laboratory reference, and environmental strains. Interestingly, the within-group ANI values shared amongst the wastewater and meat plant strains were comparable to the within-group ANI values of other reference *E. coli* groups, including O157, ST131, and ST95 strains. Considering that these reference groups are each defined by shared ecotypic characteristics and lifestyle patterns (i.e., patterns of pathogenesis), the high within-group ANI of the naturalized-engineered strains could similarly reflect a common host-independent lifestyle acquired specifically within engineered niches.

Given that ecotypes have been suggested to represent the fundamental units of bacterial diversity (Cohan and Perry 2007; Koeppel *et al.* 2008), the emergence of a naturalized-engineered *E. coli* ecotype could also be reflected in phylogeny. Confirming previous analyses (Yang *et al.* 2021; Zhi *et al.* 2019) all wastewater and meat plant strains clustered within phylogroup A. Interestingly, phylogroup A has been previously found to be negatively associated with various *E. coli* pathotypes (Escobar-Páramo *et al.* 2004; Hutton *et al.* 2018; Li *et al.* 2010), suggesting that the naturalized-engineered strains may largely be non-pathogenic. Despite this, the close grouping of the wastewater strains WW38 and SZ4 with a host-associated counterpart, coupled with the phylogenetic proximity of the enteric strain Fec6, suggest that these naturalized-engineered populations might have evolved from a host-derived ancestor after repeated passage through wastewater treatment plants and/or meat processing facilities.

Within phylogroup A, the wastewater and meat plant strains were distributed across two main sequence type clusters, consisting of ST635 strains possessing the *uspC*–IS30–*flhDC* biomarker and ST399 strains without the biomarker, regardless of their original source of isolation. Interestingly, these sequence types have been previously found to be associated with *E. coli* populations recovered from other man-made environments, particularly those related to water sanitation systems. For instance, *E. coli* isolates recovered from septic tanks by Behruznia *et al.* (2022a; 2022b) were found to be non-randomly distributed across 3 main lineages, including clonal complex 10, which was proposed to be mainly host- and freshwater-associated, as well as clonal complex 399 and ST401, which were found to be strongly associated with the septic tank niche. Similarly, Constantinides *et al.* (2020) were able to cluster non-clinical *E. coli* isolates colonizing hospital sink drains into 4 sequence type lineages, including ST635, ST401, ST472 and ST399. As such, *E. coli* populations belonging to the ST635 and ST399 lineages appear to be

particularly predisposed to becoming naturalized, and notably within engineered environments associated with food and water sanitation. Importantly, although the high genomic and phylogenetic similarity shared amongst the wastewater and meat plant strains could imply a high degree of clonality, the representation of multiple serotypes suggests that several *E. coli* lineages may have independently naturalized across food- and water-associated engineered environments.

Reinforcing their classification as distinct *E. coli* ecotypes, two independent ecotype prediction approaches were able to distinguish the naturalized strains from their host- and natural environment-associated counterparts – though the specific ecotypes predicted differed depending on the approach used. Of the two ecotype prediction approaches used, logic regression appeared to exhibit greater classification power as it could classify a greater number of wastewater and meat plant strains into naturalized-engineered-associated ecotypic groups. As discussed in the previous chapter, these findings again highlight the utility of logic regression as a reliable approach for the identification and classification of putative ecotypes within a bacterial species. Notably, however, regardless of the ecotype prediction approach used, the wastewater and meat plant strains could not be distinguished. While this could suggest that these two populations collectively represent one generalized *E. coli* ecotype that has dispersed across various engineered environments, this finding could also be due to the current limitations of the ecotype prediction approaches used. For instance, the ecotypes predicted with the Ecotype Simulation 2 algorithm were inherently linked with the phylogeny of the strains; however, unsupervised learning approaches akin to phylogenetic clustering have limited utility for evaluating *E. coli* niche-specificity and may thus not be appropriate for demarcating putative *E. coli* ecotypes (Zhi *et al.* 2015). Alternatively, while logic regression correctly classified a higher proportion of the wastewater and meat plant strains as naturalized, the classifications were made using the sequence variation contained within only two

intergenic regions (i.e., *asnS–ompF* and *csgDEFG–csgBAC*). Given that the previous chapter demonstrated that different intergenic regions encode varying degrees of niche-relevant information, the selection of alternative or a greater number of intergenic loci may improve the discrimination power of the logic regression algorithm, thereby allowing for the sub-classification of the wastewater and meat plant strains into distinct ecotypes.

Interestingly, the wastewater and meat plant strains could also be distinguished from the other *E. coli* ecotypes based on the presence of several niche-specific insertion element genetic markers. Beyond their potential application as genetic markers for tracking environmental contamination derived from engineered environmental sources (i.e., wastewater or sewage run-off), these insertion elements also appear to reflect specific evolutionary adaptations that could underlie the evolution and niche-specificity of the wastewater and meat plant strains. The integration of insertion elements within intergenic regions, for instance, may influence the expression of the surrounding genes (Zhang *et al.* 2022; Zhi *et al.* 2017), thereby allowing a bacterial cell to adjust to the specific environmental conditions of a given niche. Several of the naturalized-engineered-associated intergenic insertion markers involved genes mediating functions including biofilm formation, microbial defense (i.e., toxin-antitoxin systems, CRISPR-Cas systems, cleavage of foreign invading genetic material), and stress resistance (i.e., DNA-damaging stimuli, oxidative stress, heavy metals, etc.) – functions that would presumably need to tightly regulated to respond to the harsh conditions encountered in wastewater treatment plants and/or meat processing facilities (i.e., competition, environmental stressors). In contrast, intragenic insertions lead to gene disruption and loss-of-function, reflecting the evolutionary trade-offs that accompany the process of niche-specialization (Vamosi *et al.* 2014). Interestingly, several intragenic insertion markers that were identified within the naturalized-engineered strains involved

the disruption of genes associated with colonization, virulence, and resistance mechanisms against the specific stressors that could be encountered within a host gastrointestinal environment (i.e., acid stress in the stomach, nutritional stress due to the competing gut microbiome, antibiotic stress, etc.). As such, these intragenic insertions appear to reflect the functional trade-offs associated with the naturalization process as a given strain adapts to the engineered environment and away from the original host niche.

Remarkably, several of these insertion element markers were found to be niche-specific. A subset of the wastewater strains, for instance, were uniquely characterized by the *ygcB*–IS186B–*hokE* and *yhiM*–ISEc78–*pcoE* biomarkers. The role of the flanking genes in microbial defense (*ygcB*, *hokE*) and heavy metal resistance (*pcoE*) suggests that these biomarkers could represent important wastewater-specific adaptations required to survive the intense microbial competition (Cydzik-Kwiatkowska and Zielińska 2016), high phage load (Ballesté *et al.* 2022; Runa *et al.* 2021; Strange *et al.* 2021), and toxic heavy metal contaminants (Qasem *et al.* 2021) present within the wastewater matrix. In contrast, some of the meat plant strains were characterised by the *caiT*–IS1R–*fixA* biomarker. Interestingly, the flanking genes, *caiT* and *fixA*, appear to mediate the uptake and metabolism of carnitine, an important osmoprotectant utilized by bacteria that can also enhance thermotolerance, cryotolerance, and barotolerance (Meadows and Wargo 2015). Considering that carnitine is found at particularly high concentrations in animal tissues, the *caiT*–IS1R–*fixA* biomarker could reflect a niche-specific adaptation that the meat plant strains acquired by exploiting the specific resources available within the meat plant environment. As such, although phylogenetic and ecotype prediction approaches alone failed to distinguish between the WWS- and MPS-*E. coli* strains, they each appear to be characterized by distinct niche-adaptive signatures and thus could still represent distinct ecotypic groups.

Collectively, the genomic, phylogenetic, and ecotypic evidence presented in this chapter suggest that distinct populations of *E. coli* have evolved to exploit various food- and water-associated engineered environments as their primary niche. To date, these naturalized-engineered strains appear to have only been isolated from food and water industrial contexts, suggesting they may represent a distinct ecotype that has diverged from other host-associated and environmental *E. coli* ecotypic groups. The apparent adaptation of these WWS- and MPS-*E. coli* groups to wastewater treatment plants and meat processing facilities is puzzling, however, as these environments typically employ a variety of strategies (i.e., disinfection) specifically designed to eliminate microbes. While the evidence presented in this chapter indicate that these *E. coli* populations seem to thrive in such environments, the specific adaptive mechanisms facilitating their survival remain unclear. The following two chapters of this thesis will thus aim to recapitulate the phylogenetic and ecotypic evidence presented in this chapter with specific genetic and phenotypic adaptations that could underlie the evolutionary success of these *E. coli* populations within their respective engineered environments.

# Chapter Four: Pan-Genomic Characterization of the Genetic Features Underlying the Niche-Adaptation of Naturalized Wastewater and Meat Plant *Escherichia coli* Strains[4]

## 4.1 Introduction

The evidence presented in Chapters Two and Three of this thesis describe novel *E. coli* strains that appear to be specifically adapted to engineered environments as a primary niche. Indeed, *E. coli* strains isolated from wastewater treatment plants and meat processing facilities were reliably distinguished from their host-associated and environmental counterparts using a variety of genotypic, phylogenetic, and ecotypic means. While this collective evidence suggests that these wastewater- and meat plant-derived *E. coli* strains constitute unique ecotypic groups that have evolved to exploit man-made engineered environments, the specific adaptive mechanisms underlying their evolutionary success remain unclear.

While the insertion element markers identified in the previous chapter seem to reflect some of the functional adaptations that may have been acquired by the wastewater and meat plant strains, their niche-specificity might also be explained by presence of niche-adaptive genes, as facilitated by the *E. coli* pan-genome. *E. coli* is characterized by an 'open' pan-genomic structure (Rasko *et al.* 2008), which affords the species the ability to rapidly acquire and exchange the genetic determinants required to exploit any given niche. In particular, the continually expanding accessory genome effectively supplies *E. coli* strains with an extensive repertoire of genes that will

[4] A version of this chapter has been published as: Yu, D., Stothard, P., and Neumann, N.F. 2024. Emergence of potentially disinfection resistant, naturalized *Escherichia coli* populations across food- and water-associated engineered environments. Sci. Rep. 14(13478): 1–14. doi:10.1038/s41598-024-64241-y

be particularly adaptive within a given environment (Mira *et al.* 2010). The host-specificity of various host-derived *E. coli* ecotypes, for instance, appears to be mediated in part by the presence of key colonization factors that can bind to the gut receptors that are differentially expressed across different host species (Dubreuil *et al.* 2016; von Mentzer and Svennerholm 2023; Ron 2006). Similarly, environmental *E. coli* ecotypes appear to have acquired distinct genetic repertoires in response to the environmental stressors (i.e., desiccation, exposure to UV, cold shock, etc.) encountered outside of the host gastrointestinal environment. Indeed, compared to their host-derived counterparts, environmentally-derived *E. coli* strains appear to be enriched in genes mediating various functions, including alternative nutritional pathways (Luo *et al.* 2011), microbial defense (Luo *et al.* 2011; Tymensen *et al.* 2015), and the production of stress-resistant capsules (Touchon *et al.* 2020; Power *et al.* 2005), that could underlie their survival within, and specialization to, non-host environmental niches.

Compared to the host gastrointestinal tract and various natural environments (i.e, soil, sediment, water, etc.), engineered niches present a myriad of unique and often extreme environmental stressors (i.e., disinfection, sanitation) that could drive the divergence of resident *E. coli* strains away from their host- and environmental counterparts. Consequently, we sought to perform a comprehensive pan-genomic analysis to identify the specific genetic features that might explain the adaptation and evolutionary success of the WWS- and MPS-*E. coli* strains within their respective engineered environments.

## 4.2 Materials and Methods

### 4.2.1 Bacterial Strains

All pan-genomic analyses were performed using the same collection of *E. coli* genome

sequences that were used for the comparative genomic, phylogenetic, and ecotypic analyses conducted in the previous chapter, including: (i) naturalized wastewater and meat plant strains harboring the *uspC*–IS30–*flhDC* biomarker; (ii) naturalized wastewater and meat plant strains lacking the *uspC*–IS30–*flhDC* biomarker; (iii) enteric strains, including human and animal commensal isolates and pathogenic strains belonging to the major InPEC pathotypes (i.e., EHEC, STEC, EPEC, ETEC, EAEC, EIEC, DAEC, etc.); (iv) pathogenic strains belonging to the major ExPEC pathotypes (i.e., BBEC, NMEC, UPEC, APEC); (v) lab reference strains; (vi) environmental strains (i.e., naturalized strains within natural environments); and (vii) genus *Escherichia* strains belonging to the cryptic clades. All information related to the bacterial strains assessed in this chapter can be found in Supplementary Table 3–S1.

### 4.2.2 Comparative Genomic Alignments of Naturalized Wastewater and Meat Plant Strains Against Enteric, Extraintestinal Pathogenic, and Environmental *E. coli*

As an initial assessment into whether the naturalized wastewater and meat plant strains could harbor unique, potentially niche-adaptive genetic regions, comparative genomic alignments were performed. Specifically, pairwise whole genome alignments were conducted to visualize and identify any genetic regions that could be uniquely characteristic to a given ecotypic group (i.e., naturalized, host-associated, etc.). Strains were selected to represent five putative *E. coli* ecotypes, including a naturalized wastewater group (i.e., WW10, ABWA45, RHBSTW_00141, WW38), a naturalized meat plant group (i.e., CX20, CX05, 0H24), an enteric group including both commensal and InPEC strains (i.e., HS, Fec6, SE11, SE15, IAI1, EDL933, E2348_69), an ExPEC group (i.e., 219, UTI89, CFT073, 536, CE10), and an environmental group (i.e., WAT, SMS_3_5), as indicated in Supplementary Table 3–S1. Three sets of serial pairwise alignments were

performed with BLAST, with the genome alignment for each strain rooted against a reference naturalized wastewater strain (i.e., *E. coli* WW10), a reference naturalized meat plant strain (i.e., *E. coli* CX20), or a reference host-associated strain (i.e., *E. coli* HS). The genome alignment maps were then visualized and annotated with the reference strain's coding sequences (CDS) and GC content using the Proksee (Grant *et al.* 2023) webserver.

### 4.2.3 Pan-Genome Dynamics of Naturalized Wastewater and Meat Plant Strains with Enteric, Extraintestinal Pathogenic, and Environmental *E. coli*

All naturalized wastewater and meat plant *E. coli*, enteric *E. coli* (i.e., commensal and InPEC), ExPEC, laboratory reference *E. coli*, environmental *E. coli*, and cryptic *Escherichia* strains (n = 113) were functionally annotated using Prokka v1.14.6 (Seemann 2014), after which a pan-genome was estimated using Roary v3.13.0 (Page *et al.* 2015). Genes that were left unannotated were screened against all bacterial protein sequences available on the NCBI Protein database with BLAST, and their functions were inferred based on sequence homology to closely-related protein families and the identification of conserved functional domains. To assess the degree of genetic homogeneity amongst the strains analysed (i.e., whether certain subsets of strains, corresponding to distinct ecotypes, could be characterized by unique genes), the distribution of genes within the estimated pan-genome was evaluated using a pan-genome spectrum function (Baumdicker *et al.* 2012; Collins and Higgs 2012; Gordienko *et al.* 2013; Moldovan and Gelfand 2018). Briefly, a pan-genome spectrum function plots the number of genes within a pan-genome against the number of genomes each gene is found in (Baumdicker *et al.* 2012; Collins and Higgs 2012) to evaluate the degree of genetic homogeneity across a set of bacterial strains within a sample. Genetically homogenous samples will produce a smooth 'U'-

shaped curve, whereas non-homogenous samples will produce internal peaks within the spectrum function corresponding to the number of genes that are unique to a subset of strains (Moldovan and Gelfand 2018). While typically used to differentiate the gene pools of different species (Gordienko *et al.* 2013; Moldovan and Gelfand 2018), this approach could also be leveraged for the identification of ecologically-relevant, niche-adaptive genes within a given *E. coli* ecotype.

### 4.2.4 Identification of Ecotype-Informative, Niche-Adaptive Genes Within Naturalized Wastewater and Meat Plant *E. coli* Strains through Pan-Genome Wide-Association Studies

To characterize the specific genomic features underlying the naturalization and niche-adaptation of the wastewater and meat plant strains, a pan-genome-wide association study (pan-GWAS) was performed. Scoary v1.6.16 (Brynildsrud *et al.* 2016) was used to score every gene in the pan-genome to identify genes that were statistically over-represented and under-represented across the naturalized strains. Three separate analyses were performed, to determine genes that were correlated with: a) the naturalized wastewater group specifically; b) the naturalized meat plant group specifically; and c) the naturalized group as a whole, encompassing both wastewater and meat plant strains. All Scoary runs were performed with the '--no_pairwise' flag and used the Benjamini-Hochberg correction method with a *p*-value cut-off of 1E-5, as recommended by the developers of the program (https://github.com/AdmiralenOla/Scoary).

The results from each run were combined, after which duplicate gene entries, truncated genes, and genes present in fewer than 75% of the naturalized strains (as a lower limit for 'over-represented' genes) were screened out. The remaining genes were then broadly classified according to their prevalence amongst the naturalized strains, as either: a) 'absent', if they were

not present in any of the naturalized strains; b) ‘duplicate’, if the gene of interest in the naturalized strains appeared to be a copy of another gene that was already widely prevalent across the strains included in the analysis; c) ‘shared’, if the gene of interest was the only copy in the naturalized strains, but was still shared amongst other strains in the analysis; d) ‘unique’, for gene entries that were present only in the naturalized strains, and; e) ‘variant’, if there were multiple entries for a given gene, but for which specific entries appeared to be particularly over- or under-represented in the naturalized strains. Additionally, the genes were also categorized based on their distribution across the strains analysed, as either: a) ‘wastewater-dominant’, if the gene entry’s prevalence was 40% higher in the wastewater strains compared to the meat plant strains; b) ‘meat plant-dominant’, if the gene entry’s prevalence was 40% higher in the meat plant strains compared to the wastewater strains; c) ‘common across wastewater and meat plant’, if the gene entry exhibited greater than 50% prevalence in both the wastewater and meat plant strains, but with no significant difference in sensitivity between the two groups; and d) ‘lacking in wastewater and meat plant’, if the gene entry was under-represented in both groups of strains. All genes that were found to be statistically correlated with the naturalized groups by Scoary were functionally annotated after reference to the UniProt (Bateman *et al.* 2017) and EcoCyc (Karp *et al.* 2018) databases. The distribution of these genes across the strains analyzed was then visualized through a presence/absence heatmap produced with R software using the ggplot2 v3.4.2 package (Wickham 2009).

### 4.2.5 Gene-Gene Interaction Mapping of Naturalized-Associated Accessory Genes

The presence of certain accessory genes may impact the presence of other accessory genes within the pan-genome. These gene-gene interactions can lead to genes co-occurring within the same genomic background if they exhibit a synergistic effect on fitness, or dissociating (i.e.,

antagonistic pleiotropy) if they antagonize each other's fitness contribution. To assess these gene-gene relationships, Coinfinder v1.2.1 (Whelan *et al.* 2020) was used to evaluate gene association (i.e., the presence of one gene is linked to the presence of another) and dissociation (i.e., one gene is present specifically when another is absent) interactions occurring within the genomic background of the naturalized strains. Pan-genome association and dissociation networks were produced using Coinfinder based on the estimated pan-genome produced by Roary and a core genome phylogenetic tree produced with FastTree v2.1.11 (Price *et al.* 2010), with the Bonferroni correction method, as recommended by the developers of the program (https://github.com/fwhelan/coinfinder). The gene-gene association and dissociation network maps were then visualized using the Fruchterman Reingold layout with the Gephi platform (Bastian *et al.* 2009) and annotated using Inkscape software.

### 4.2.6 Localization of Naturalized-Associated Resistance and Defense Genes on Mobile Genetic Elements

As the wastewater and meat plant strains appear to be specifically adapted to niches representing potential environmental hotspots for the evolution of antibiotic, and potentially disinfection, resistance, a screening analysis was performed to assess whether they could serve as reservoirs of key resistance determinants in their respective engineered environments. First, all naturalized strains were screened for putative plasmid sequences using ABRicate v1.0.1 (https://github.com/tseemann/abricate) against the PLASMIDFINDER database (Carattoli and Hasman 2020). All identified plasmids were then typed using the Plasmid MLST tool (https://pubmlst.org/organisms/plasmid-mlst) on the PubMLST webserver (Jolley *et al.* 2018). To identify whether the naturalized strains could be carrying key antibiotic resistance, stress

resistance, and microbial defense genes on plasmids, the RFPlasmid webserver (van der Graaf-Van Bloois *et al.* 2021) was used to identify the specific contigs for each strain that correlated with plasmid sequences. Each strain's plasmid-associated contigs were then compared to the CARD 2020 database (Alcock *et al.* 2020) with ABRicate to screen for key, clinically-relevant antibiotic resistance genes, while the remaining genes were functionally annotated using the UniProt (Bateman *et al.* 2017) and EcoCyc (Karp *et al.* 2018) databases to identify potentially mobile genes associated with microbial defense systems or stress resistance (i.e., against disinfection).

## 4.3 Results

### 4.3.1 Comparative Genomic Alignments of Naturalized Wastewater and Meat Plant Strains Against Enteric, Extraintestinal Pathogenic, and Environmental *E. coli*

As an initial assessment into whether the naturalized wastewater and meat plant strains could harbor unique, potentially niche-adaptive genetic regions, comparative whole-genome alignments were performed. Overall, all genome alignment maps revealed extensive commonality amongst the strains compared, regardless of the reference strain used; however, several gaps, representing unique naturalized- or host-associated genetic regions. were observed in the alignments (Figure 4-1). Most gaps were observed in the alignments rooted against the naturalized wastewater (Figure 4-1A) and meat plant (Figure 4-1B) reference strains, representing unique genetic regions found in the wastewater and meat plant strains, respectively, that were absent in their host- and natural environment-associated counterparts. Interestingly, the wastewater- and meat plant-rooted maps also contained regions that were commonly unique or over-represented in both wastewater and meat plant groups, representing genetic regions that were generally characteristic of the naturalized strains. To a lesser extent, gaps were also observed in the map

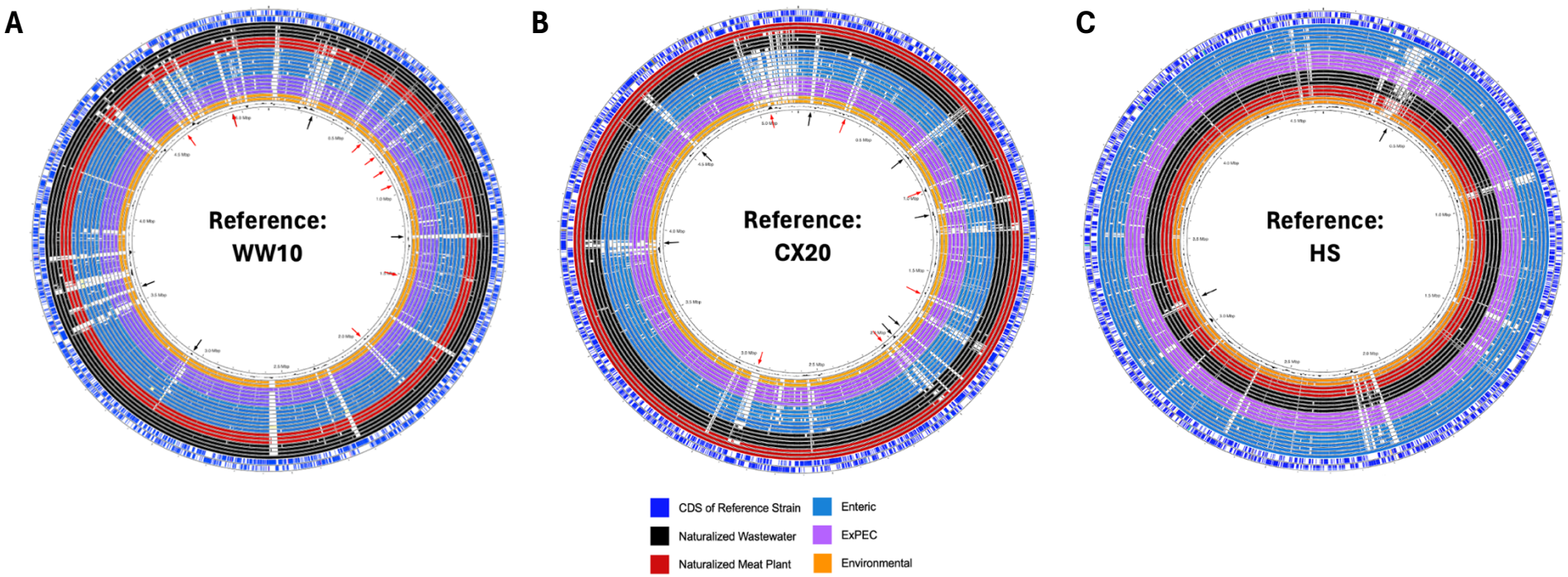



**Figure 4-1. Serial pairwise genomic alignment maps of naturalized wastewater and meat plant strains with enteric, extraintestinal pathogenic, and environmental *E. coli* strains.** Genome alignment maps of representative *E. coli* strains belonging to five ecotypic groups, including naturalized wastewater, naturalized meat plant, enteric (including both commensal and InPEC strains), ExPEC, and environmental *E. coli*, rooted against **(A)** a reference naturalized wastewater strain, WW10; **(B)** a reference naturalized meat plant strain, CX20; and **(C)** a reference host-associated strain, HS. In each alignment map, the reference strain's coding sequences (CDS), tRNA genes and rRNA genes are depicted as the outer two blue rings, with the reference genome included as the solid ring just inside the two outer rings. Against each reference genome, strains belonging to different ecotypes are aligned, including: i) the naturalized wastewater strains (black rings) WW10, ABWA45 and RHBSTW_00141 as *uspC*–IS30–*flhDC*–positive strains, and WW38 as a *uspC*–IS30–*flhDC*–negative strain; ii) the naturalized meat plant strains (red rings) CX20 and CX05 as *uspC*–IS30–*flhDC*–positive strains and 0H24 as a *uspC*–IS30–*flhDC* negative strain; iii) the enteric strains (blue rings), including HS, Fec6, SE11, SE15, and IAI1 as commensal strains, and EDL933 and E2348_69 as intestinal pathogenic strains; iv) the ExPEC strains (purple rings) 219, UTI89, CFT073, 536, and CE10; and v) the environmental strains WAT and SMS_3_5. Gaps in each alignment indicating genetic sequences unique to the reference strain for each ecotype (black arrows), as well as those unique to the naturalized ecotype generally (red arrows), are depicted in the center of the maps.

rooted against the host-associated strain (Figure 4-1C), representing genetic regions that were missing in the naturalized strains when compared to their host-associated counterparts, suggesting that the naturalized strains may lack certain genetic features that are otherwise characteristic of their host-associated counterparts.

### 4.4.2 Pan-Genome Dynamics of Naturalized Wastewater and Meat Plant *E. coli* Strains with Other Strains Representative of the *E. coli* Species

A pan-genome was calculated for the 37 naturalized wastewater and meat plant strains alongside 76 representative commensal *E. coli*, InPEC, ExPEC, laboratory reference *E. coli*. environmental *E. coli* and cryptic *Escherichia* strains (Supplementary Table 3–S1). The pan-genome was estimated to consist of 37,502 total genes, including 1885 that were 'core' and shared by ≥ 98% of the strains included in the analysis (Figure 4-2). To specifically evaluate the distribution of genes within the estimated pan-genome, a pan-genome spectrum function analysis (Baumdicker *et al.* 2012; Collins and Higgs 2012) was performed. In this analysis, the spectrum function produced a curve containing slight internal peaks (Figure 4-3), suggesting that the genes were non-homogenously distributed across the strains in the pan-genome (Gordienko *et al.* 2013; Moldovan and Gelfand 2018). Interestingly, these internal peaks appeared to coincide with sets of genes that were unique to the naturalized-engineered strains, suggesting that they may possess distinct genetic adaptations reflective of their respective engineered niches when compared to their host-associated and environmental counterparts.

| Type | No. Genes | % |
|---|---|---|
| Cloud (< 15% Strains) | 31335 | 83.5555437 |
| Shell (15–95% Strains) | 3590 | 9.57282278 |
| Soft-Core (95–98% Strains) | 692 | 1.84523492 |
| Core (≥98% Strains) | 1885 | 5.02639859 |
| **TOTAL** | **37502** | **100** |

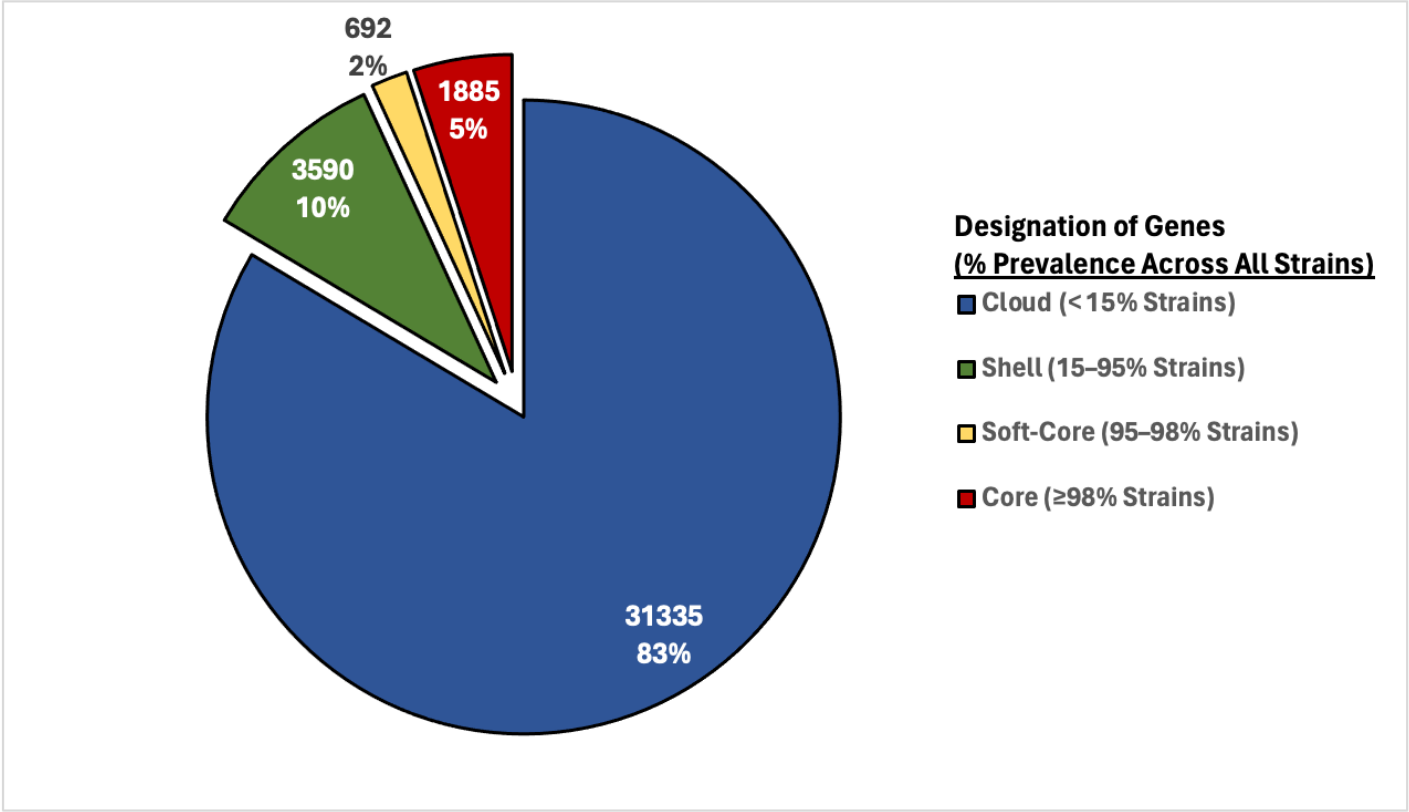



**Figure 4-2. Pan-genome dynamics of the naturalized *E. coli* strains.** Pan-genome estimated using Roary for the naturalized wastewater and meat plant *E. coli* strains alongside representative enteric *E. coli*, InPEC, ExPEC, laboratory reference *E. coli*, natural environment *E. coli*, and cryptic *Escherichia* strains. Genes within the pan-genome were categorized according to the prevalence across the strains analysed in this study, as: (i) 'cloud' genes if they were present in less than 15% of all strains (blue); (ii) 'shell' genes if they were present in anywhere between 15-95% of all strains (green); (iii) 'soft-core' genes if they were present in 95-98% of strains (yellow); and (iv) 'core' genes if they were present in at least 98% of all strains (red).

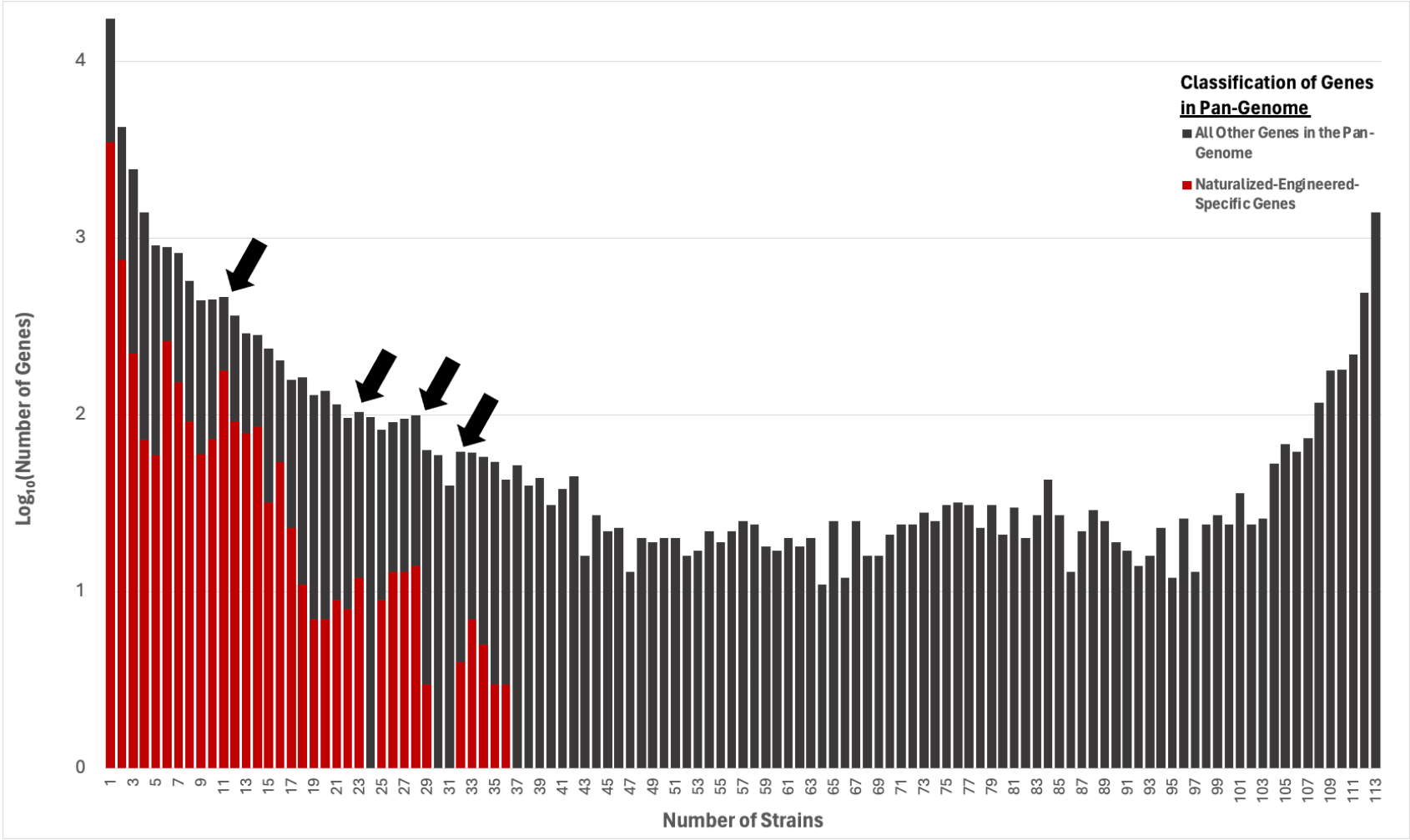



**Figure 4-3. Pan-genome spectrum function depicting the distribution of genes across the estimated pan-genome.** The prevalence of each gene within the estimated pan-genome was determined based on the number of strains that each was found in, according to Roary. The number of genes that were found in each number of strains within the sample were then plotted to produce a pan-genome spectrum function, to visualize the degree of genetic heterogeneity amongst the strains analysed in this chapter. Note the internal peaks present within the spectrum function curve indicating the heterogeneity of the gene pool, with the peaks (shown with black arrows) roughly coinciding with genes specifically found in the naturalized-engineered strains (depicted in red).

### 4.4.3 Identification of Ecologically-Relevant, Niche-Adaptive Genes Within Naturalized Wastewater and Meat Plant *E. coli* Strains

To identify the specific genes that could be associated with the evolution and adaptation of the wastewater and meat plant strains towards their respective engineered niches, a pan-genome-wide association study was performed. Of the 2501 genes identified by Scoary to be statistically correlated with the naturalized-engineered strains, 2082 (83%) were found to be over-represented amongst the WWS and MPS strains when compared to their host-associated and environmental counterparts, whereas 419 (17%) were under-represented (Figure 4-4A). Although 736 genes were left unannotated with no known function, the remaining 1765 were found to be distributed across several functional categories, including those that could be particularly relevant for a naturalized lifestyle within the engineered environment (Figure 4-4B). Notably, when compared to *E. coli* strains belonging to other host- and natural environment-associated ecotypic groups, the wastewater and meat plant strains appeared to be particularly enriched in genes involved in adhesion and biofilm formation, microbial defense, and stress resistance, although the number of genes that were comparatively under-represented within these strains were found to be disproportionately associated with virulence and colonization (Figure 4-5).

Regarding specific genes important for adhesion and biofilm formation, both wastewater and meat plant strains were found to encode the alternative Yfc fimbrial system (‘*yfcOPQRS’*) which, while normally cryptic in reference laboratory strains, appears to play a role in adhesion to environmental surfaces (Korea *et al.* 2010) (Supplementary Table 4-S1). Beyond this, both the wastewater and meat plant strains were individually characterized by additional genes that could enhance their ability to adhere to surfaces and form biofilms within the environment. For instance, the wastewater strains encoded components of other alternative fimbrial systems, including Yeh

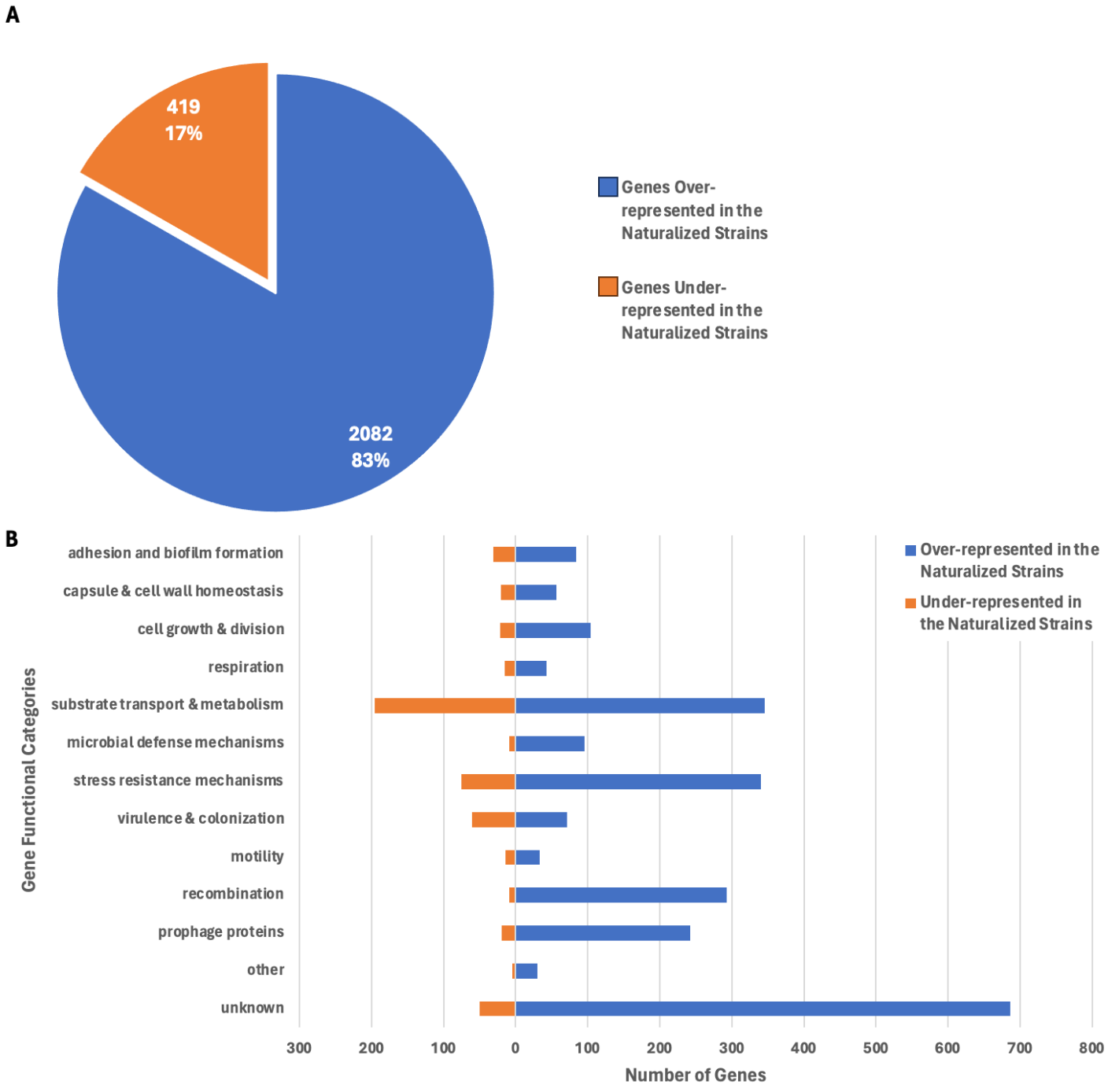



**Figure 4-4. Summary statistics of pan-genome wide association study results. (A)** Numeric breakdown and **(B)** functional distribution of genes that were found to be statistically over-represented (blue) and under-represented (orange) in the naturalized wastewater and meat plant strains when compared to other strains representative of the *E. coli* species, as determined by Scoary.

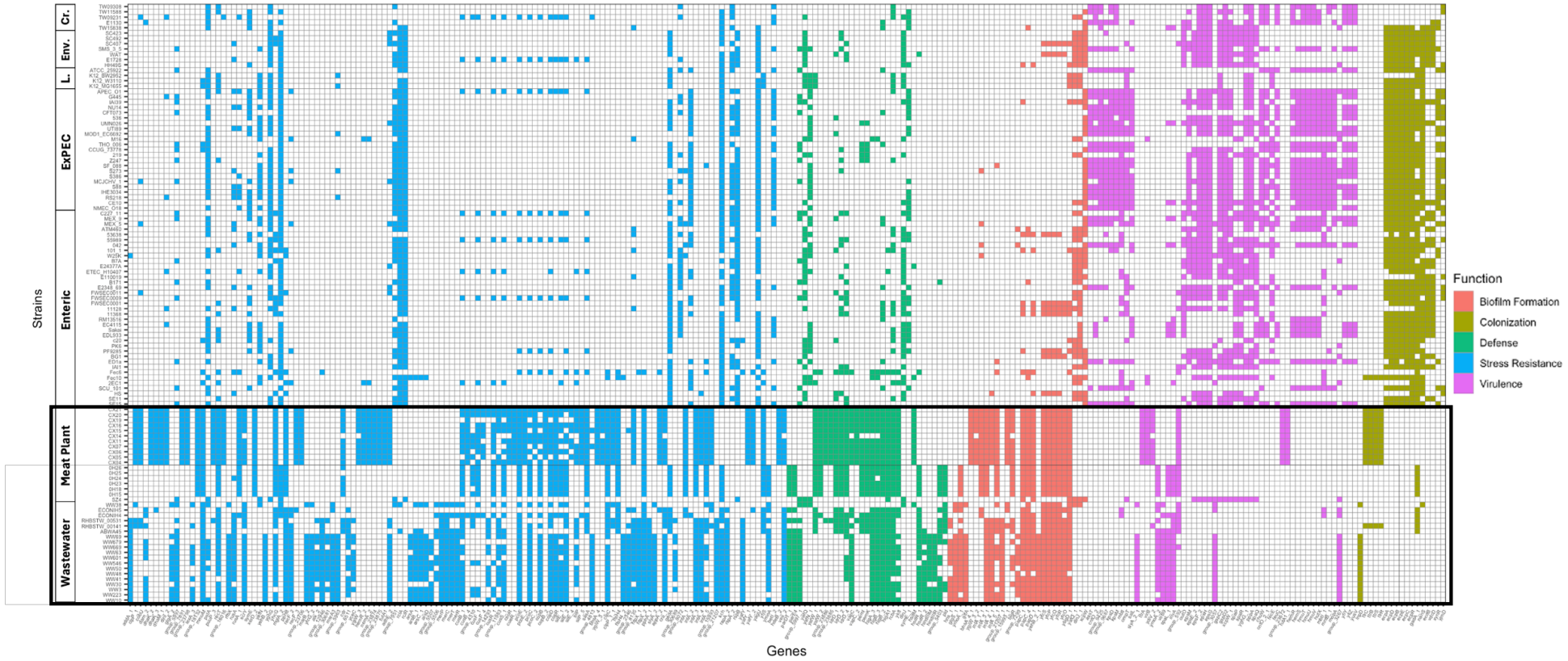



**Figure 4-5. Presence/absence heatmap of genes statistically correlated with the wastewater and meat plant strains when compared to other strains representative of the *E. coli* species.** All genes in the estimated pan-genome were statistically scored ($p < 1E\text{-}5$, with Benjamini-Hochberg correction) with Scoary to determine genes that could be associated with the distinct ecology of the wastewater and meat plant strains, especially in comparison to strains belonging to the other major *E. coli* ecotypes, including: "Enteric" (i.e., commensal and InPEC) strains, "ExPEC" strains, lab reference (labelled as "L.") strains, natural environmental (labelled as "Env.") strains, and cryptic Escherichia (labelled as "Cr.") strains. The function of each annotated gene was determined after cross-referencing against the UniProt and EcoCyc databases, while the functions of unannotated genes were inferred through sequence homology shared with protein entries in NCBI databases. Several functions were represented across the genes identified by Scoary and based on their characteristic distribution across the strains analyzed, the wastewater and meat plant strains (indicated with the black box) appeared to be enriched in genes associated with adhesion and biofilm formation (red), microbial defense mechanisms (green) and stress resistance (blue), but were relatively lacking in genes related to colonization (yellow) and virulence (purple).

and Elf fimbriae, that also appear to be involved in environmental adhesion (Korea *et al.* 2010, as well as duplicates of the genes *hns* (‘*hns_1*’) and *glgS* (‘*glgS_1*’), which appear to play regulatory roles in the control of pili and fimbrial systems and the production of biofilm polysaccharides, respectively. In comparison, the meat plant strains were characterized by the presence of additional biofilm regulators, such as *bhsA* (‘*bhsA_4*’) and *bigR*, which could enhance their control over the formation of biofilms, especially in response to changing environmental conditions.

The naturalized-engineered strains were also found to be enriched in genes associated with microbial defense mechanisms. For instance, these strains harbored an abundance of toxin-antitoxin system genes, including those shared between the wastewater and meat plant strains such as *higAB* (‘*higA_3*’, ‘*higA1*’, ‘*higB_2*’, ‘*higB_3*’), *vapC*, *rcbA*, and *ykfI*, as well as toxin-antitoxin genes specifically associated with the wastewater strains, such as *parDE* (‘*parD1_1*’, ‘*parE4*’), and meat plant strains, including *pemIK, ccdAB* (‘group_23885 (*ccdB_2*)’, ‘group_23886 (*ccdA*)’), *yafW* (‘*yafW_1*’, ‘*yafW_2*’), and *ldrD (‘ldrD_1*’, ‘*ldrD_2*’, ‘*ldrD_4*’), respectively (Supplementary Table 4-S1). Furthermore, the naturalized-engineered strains also possessed a myriad of defense genes against phages and other invasive mobile genetic elements, including the common anti-phage defense protein *pld*, the meat plant-associated restriction-modification system protein *hsdM*, and the wastewater-associated restriction-modification system proteins *hsdR* (‘*hsdR_1*’), *hindIIIM*, and *hindIIIR,* and CRISPR-Cas system protein *casC* (‘group_5403 (*casC*)’).

Most notably, the naturalized-engineered strains were characterized by an abundance of stress resistance genes. Both wastewater and meat plant strains, for instance, were enriched in genes related to responses to DNA-damaging stimuli (i.e., UV radiation), including various DNA repair and SOS response genes such as *recT*, *recF* (‘*recF_2*’, ‘*recF_3*’, ‘group_23706 (*recF_2*)’), *dam* (‘*dam_2*’), *lexA* (‘*lexA_3*’, ‘group_18132 (*lexA_2*)’), *rusA* (‘*rusA’*, ‘*rusA_1*’), *addA*

(‘*addA_1*’), *yhcG*, *dinI* (‘*dinI_2*’), *umuD* (‘group_23798 (*umuD_1*)’) and *umuC* (‘group_7891 (*umuC_2*)’) (Supplementary Table 4-S1). These strains also possessed various oxidative stress resistance genes, including common antioxidant proteins such as *adhE* (‘*adhE_2*’) and oxidative damage repair chaperones such as *msrA* (‘*msrA3*’, ‘group_23875 (*msrA3*)’) and *msrB* (‘*msrB_2*’, ‘group_23874 (*msrB_2*)’). Furthermore, the wastewater strains were also found to harbor the redox modulator *alx* (*‘alx_2*’), the antioxidant proteins *yfcG* (‘*yfcG_2’*), oxidative stress proteins including *stiP* (‘group_30681 (*stiP*)’) and *yceC*, and the electrophilic stress protein *kefC* (‘group_12946 (*kefC_1*)’), whereas the meat plant strains were characterized by the chlorine resistance proteins *nemA* (*‘nemA_2*’) and *nemR* (‘*nemR_2*’). Interestingly, the naturalized-engineered strains appeared to lack the chlorine resistance genes *rclR* and *rclA.*

Beyond DNA-damaging stimuli and oxidative stress, the naturalized-engineered strains also harbored various heavy metal resistance systems, including the *cus* (‘*cusB_1*’, ‘group_4310 (*cusC_1*)’, ‘*cusF_1*’, ‘*cusR_1*’, ‘*cusB_4*’, ‘*cusF_4*’, ‘group_14293 (*cusF_1*)’, ‘*cusS_1*’, ‘group_11983 (*cusR_1*)’), *pco* (‘*pcoC*’, ‘*pcoC_2*’, ‘*pcoE*’, ‘*pcoE_2*’), and *cop* (‘*copB*’, ‘*copB_2*’, ‘*copD*’, ‘*copR*’) systems involved in copper resistance, as well as the *sil* (‘*silE_1*’, ‘*silE_2*’, ‘*silE_4*’, ‘*silP*’, ‘*silP_5*’) system involved in silver resistance (Supplementary Table 4-S1). Furthermore, while the wastewater strains possessed the *ars* (‘*arsA’*, ‘*arsB_1*’, ‘*arsC_1*’, ‘*arsD*’) arsenical resistance system and the *mer* (‘group_5506 (*merA*)’, ‘group_5507 (*merA*)’, ‘*merC*’, ‘*merP*’, ‘*merR*’, ‘*merT*’) mercuric resistance system, the meat plant strains harbored additional copper resistance genes including *cueR* (‘*cueR_1*’) and *csoR*. The naturalized-engineered strains additionally possessed various genes involved in other stress responses, including several heat shock proteins such as *htpX* (‘*htpX_1*’, ‘*htpX_2*’), *clpC, ftsH4* (‘*ftsH4*’, ‘*ftsH4_2*’), *hspA* (‘*hspA*’, ‘group_14438 (*hspA*)’), *pphA* (‘*pphA_1*’) and *clpP* (*‘clpP_1*’), and the cold shock protein *ves*

(‘*ves_2*’). Remarkably, beyond these annotated genes, the wastewater and meat plant strains also harbored a myriad of hypothetical proteins that, based on sequence homology, appear to further augment functions related to biofilm formation, microbial defense, and stress resistance (Supplementary Table 4-S1).

### 4.4.4 Identification of Host-Adaptive Genes Lacking in the Naturalized Wastewater and Meat Plant *E. coli* Strains

While the naturalized-engineered strains harbored an abundance of genes associated with adaptive functions (i.e., biofilm formation, microbial defense, stress resistance) within engineered contexts, they also appeared to lack genes likely required to survive within the original host environment. For instance, these strains were found to lack various virulence factors, including the ExPEC-associated *kps* (‘*kpsD*’, ‘*kpsF*’, ‘*kpsM*’, ‘*kpsU*’) capsule biosynthesis genes (Sarowska *et al.* 2019), various secretion system structural and effector proteins (‘*outO*’, ‘*epsE_2*’, ‘*epsF_2*’, ‘*epsG*’, ‘*epsH*’, ‘*epsM*’, *‘gspC2*’, ‘*gspD2*’, ‘*spaP*’, ‘*spaQ*’, *‘spaR*’, *‘pppA*’, ‘*xcpW*’, ‘*yghG*’), and a myriad of sequestration proteins (*‘chuW*’, ‘*feuC*’, ‘*fecE*’, ‘*hemR*’, ‘*hemS*’, ‘*hmuT*’, ‘*hmuU*’, ‘*hmuV*’, ‘*hutX*’, ‘*yfiY*’, ‘*yfiZ*’, ‘*yusV*’, ‘*mopA*’) involved in the acquisition of iron and other essential metals (Supplementary Table 4-S1). Furthermore, the wastewater and meat plant strains were also found to lack various key genes that could be involved in host-colonization, including the acid resistance gene *gadA* which facilitates survival during passage through the stomach (Tramonti *et al.* 2006), key colonization factors such as the *E. coli* common pilus (Garnet *et al.* 2012) encoded by the *ecp* operon (*‘ecpRABCDE*’) and an esterase (‘*nanS*’) that cleaves sialic acid residues commonly found lining mammalian mucosal sites (Steenbergen *et al.* 2009), and detoxification genes such as *ecdB* and *vdcD* providing resistance against the toxic phenolic acid

derivatives that are commonly produced as metabolites within the gastrointestinal environment (Cueva *et al.* 2010).

### 4.4.5 Associative and Dissociative Gene-Gene Interactions within the Naturalized-Engineered Strains Reflect Niche-Adaptation and Antagonistic Pleiotropy

The characteristic over- and under-representation of certain functions within the wastewater and meat plant strains suggests that their adaptation towards their respective engineered niches may have come at a cost of reduced fitness within the original host environment. To assess the specific gene-gene interactions between the over-represented naturalized niche-adaptive genes and the under-represented host-adaptive genes within the genomic background of the naturalized strains, gene association and dissociation networks were generated (Supplementary Figure 4-S1). Generally, associative gene-gene interactions reflected ecotype, as genes that co-occurred with each other in the WWS and MPS strains typically involved functions relevant for survival within the engineered niche, including biofilm formation, microbial defense, and stress resistance (Figure 4-6). Conversely, co-occurring genes that were lacking in the naturalized strains were associated with virulence and colonization – functions typically associated with the host environment. Interestingly, interactions between these two groups of genes appeared to be antagonistic as select genes linked with colonization (i.e., '*ecpRABCDE*' and *'nanS*') that were under-represented in the naturalized strains formed large dissociation networks with genes involved in stress resistance, biofilm formation, and microbial defense that were otherwise over-represented and presumably niche-adaptive in the naturalized strains (Figure 4-7).

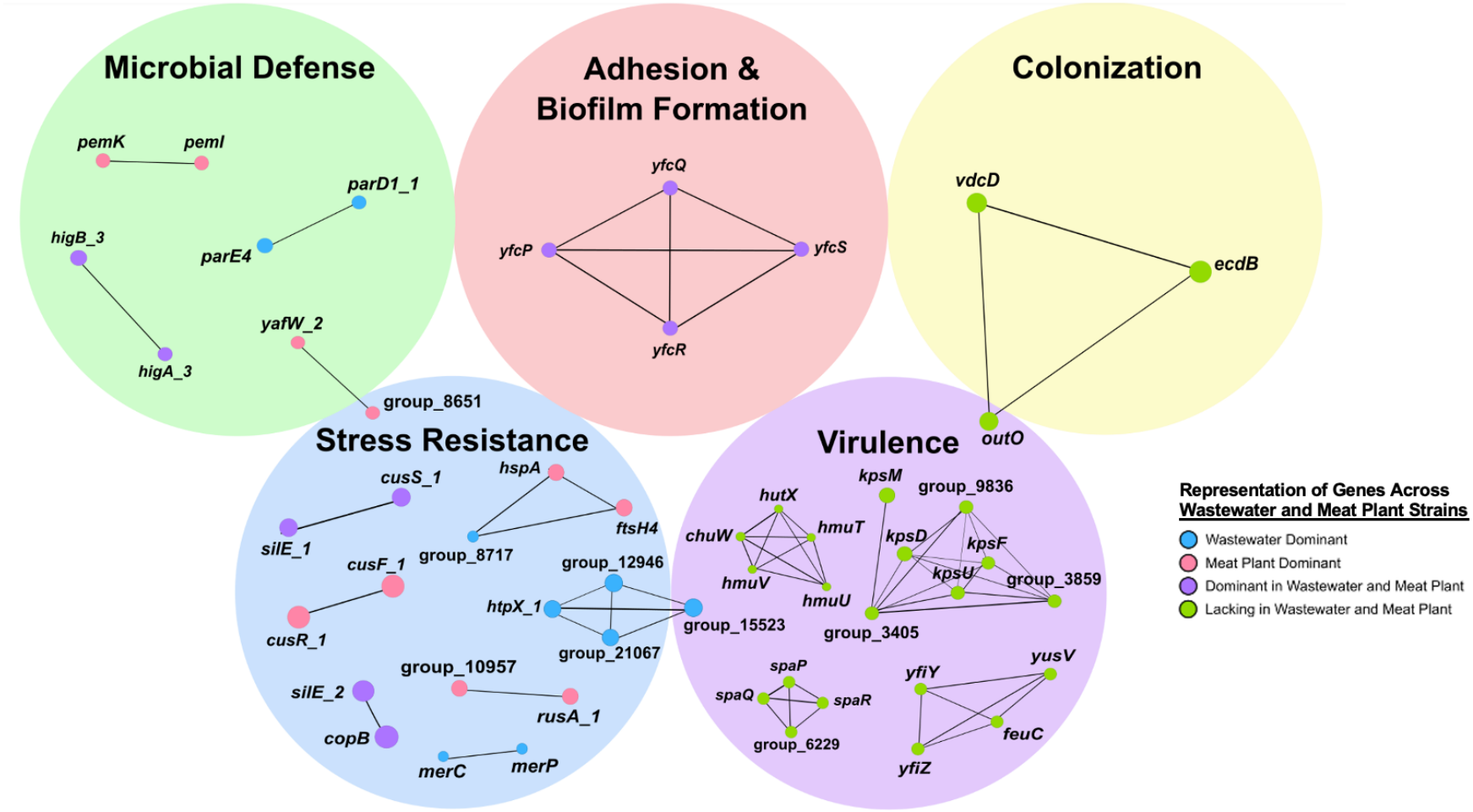



**Figure 4-6. Concurrent gene-gene interaction networks associated with the wastewater and meat plant *E. coli* strains.** Coinfinder was used to evaluate all gene-gene interactions within the estimated pan-genome. Concurrent gene-gene association networks (i.e., depicting instances where genes would co-occur with one another) that positively and negatively correlated with the wastewater and meat plant strains were identified. Each gene, represented by the circular nodes, were color-coded according to their representation across the strains including those that were wastewater-dominant (blue circles), meat plant-dominant (red circles), common across all naturalized strains (purple circles), and lacking amongst the naturalized strains (green circles), as determined previously through Scoary.

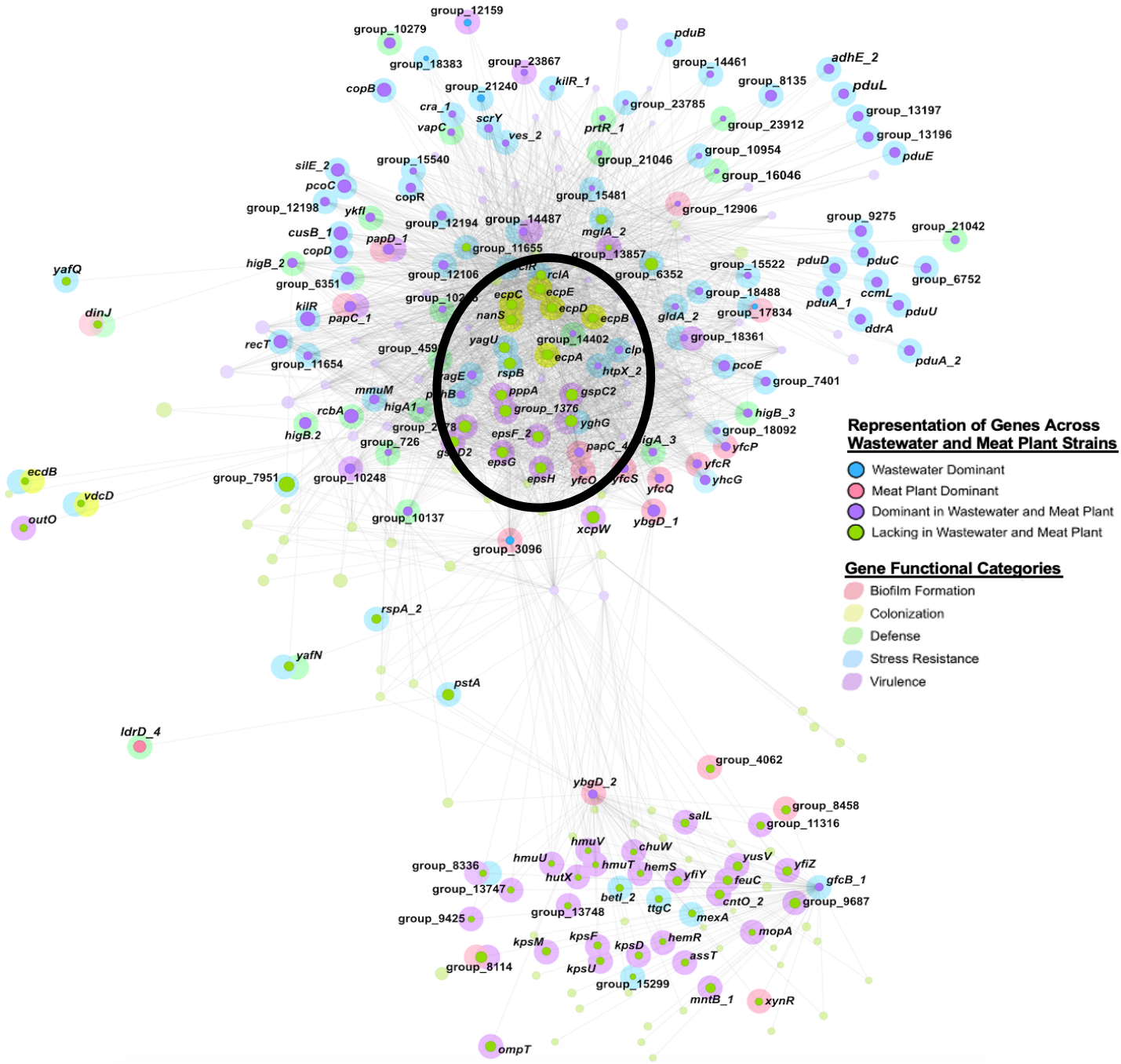



**Figure 4-7. Discordant gene-gene interaction networks associated with the wastewater and meat plant *E. coli* strains.** Coinfinder was used to evaluate all gene-gene interactions within the estimated pan-genome that were discordant (i.e., when one was present, the other was absent). Each gene, represented by the circular nodes, was first classified as: wastewater-dominant (blue circles), meat plant-dominant (red circles), common across all naturalized strains (purple circles), or lacking amongst the naturalized strains (green circles), as determined with Scoary. Each gene was then color-coded according to function, represented by the lighter-shaded clouds surrounding each node, including biofilm formation (red), colonization (yellow), microbial defense (green), stress resistance (blue), and virulence (purple). Note the central cluster of colonization and virulence genes (indicated with the black ring) surrounded by the radiation of mostly stress resistance and defense genes that were found to be over-represented in the wastewater and meat plant strains.

### 4.4.6 Identification of Mobile Antibiotic Resistance, Microbial Defense, and Stress Resistance Genes within the Naturalized Wastewater and Meat Plant *E. coli* Strains

Engineered environments such as wastewater treatment plants and meat product processing facilities represent emerging hotspots for the evolution and dissemination of antibiotic resistance within the microbial world (Kunhikannan *et al.* 2021; Rizzo *et al.* 2013). Considering that the naturalized strains appear to be specifically adapted to these environments, they may serve as resident populations that can act as important reservoirs for the mobilization of key resistance determinants. Reflecting this, the naturalized-engineered strains were found to possess a myriad of putative plasmid sequences (Table 4-1). Although various plasmid types were represented across the different plasmid incompatibility groups, most of the naturalized strains appeared to be characterized by a unique IncF plasmid type, 'F100:A-:B-'. Aside from this IncF plasmid, various other plasmid types were represented across the naturalized strains, including Type 3 IncA/C plasmids within the wastewater strains ECONIH4 and ECONIH5, various IncH1-type plasmids across both wastewater (ECONIH5) and meat plant (0H15, 0H18, 0H23, 0H24, 0H25, 0H26) groups, and a Type 1 IncN plasmid within the wastewater strain ABWA45.

Several naturalized-engineered strains were also found to harbor various antibiotic resistance genes, many of which appear to be of growing clinical concern (Van Hoek *et al.* 2011). Interestingly, all naturalized-engineered strains were characterized by a distinct chromosomal beta-lactamase gene, $bla_{EC\text{-}15}$ (Figure 4-8); however, outside of this unique beta-lactamase, the majority of the wastewater and meat plant strains did not appear to possess any key mobile antibiotic resistance genes. Despite this, a select number of the wastewater strains were found to harbor clinically-relevant antibiotic resistance genes across several antibiotic classes. For instance, the Swiss naturalized wastewater strain ABWA45 (Zurfluh *et al.* 2017), which was isolated from

**Table 4-1.** Screening of putative plasmids harbored by the naturalized wastewater and meat plant *E. coli* strains.

| Plasmid Incompatibility Group | Plasmid Type (pMLST) | Ecotype | Strains |
|---|---|---|---|
| IncA/C | Type 3 | Naturalized Wastewater | ECONIH4, ECONIH5 |
| IncF | F100:A-:B- | Naturalized Wastewater | WW10, WW223, WW3, WW30, WW41, WW48, WW50, WW546, WW601, WW63, WW669, WW679, WW69, ABWA45, RHBSTW00141, RHBSTW00531 |
| | | Naturalized Meat Plant | CX04, CX05, CX06, CX07, CX11, CX15, CX16, CX19, CX20, CX21, 0H15, 0H18, 0H23, 0H24, 0H25, 0H26 |
| | F95:A-:B- | Naturalized Wastewater | ECONIH4 |
| IncH1 | Type 2 | Naturalized Meat Plant | 0H18, 0H25 |
| | Type 6 | | 0H24 |
| | Type 7 | | 0H15 |
| | Type 10 | | 0H23, 0H26 |
| | Type 11 | Naturalized Wastewater | ECONIH5 |
| IncH2 | - | - | - |
| IncI | - | - | - |
| IncN | Type 1 | Naturalized Wastewater | ABWA45 |

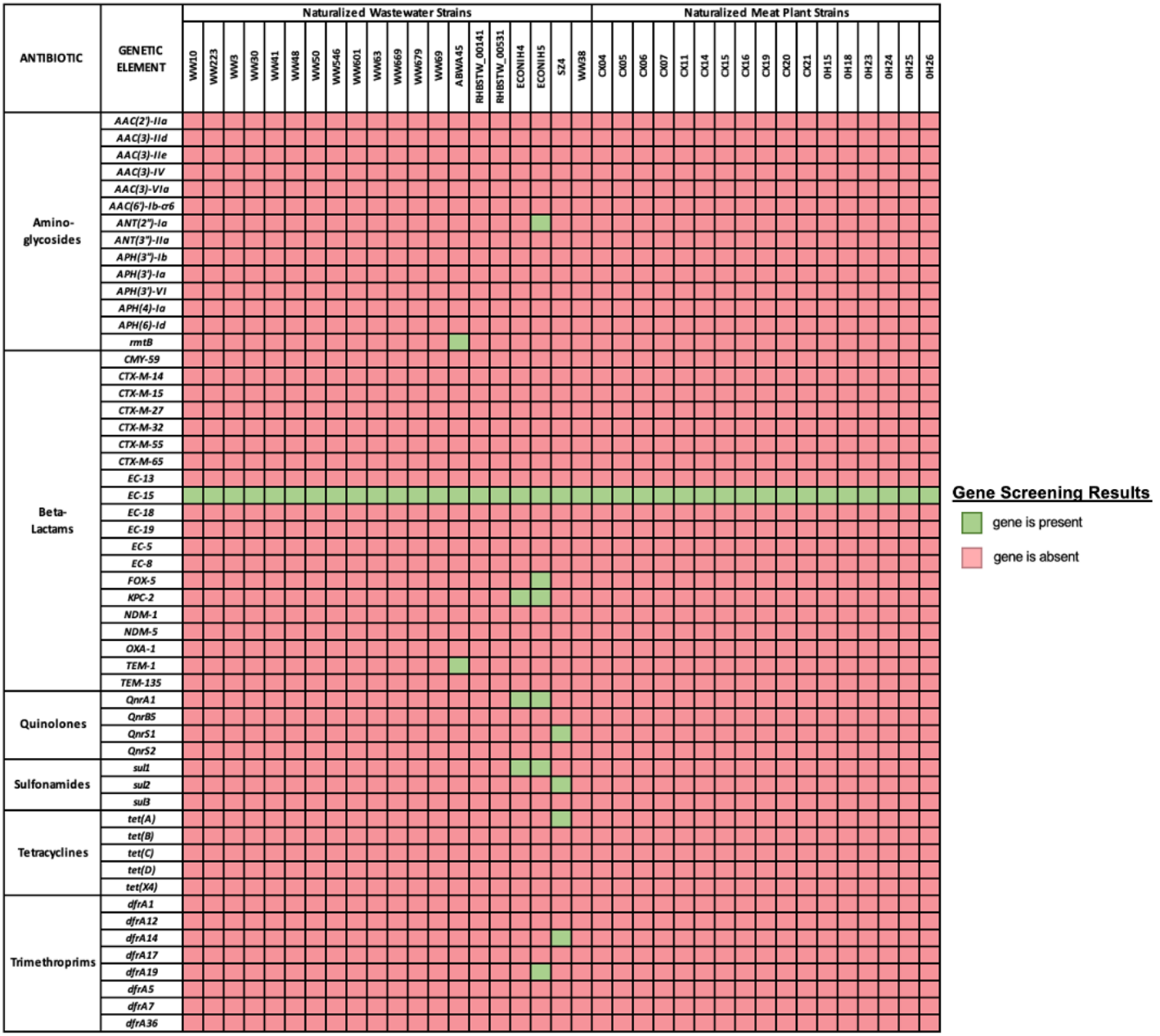



**Figure 4-8. Presence/absence screening of antibiotic resistance genes across the naturalized wastewater and meat plant *E. coli* strains.** All wastewater and meat plant strains were screened against the CARD database for key, acquired antibiotic resistance genes using ABRicate (https://github.com/tseemann/abricate). Only gene sequence matches meeting or exceeding 95% identity and coverage to the reference antibiotic resistance gene sequences in the CARD database were treated as positive hits within the heatmap.

hospital wastewater, was found to harbor the aminoglycoside resistance gene *rmtB*, as well as the beta-lactamase $bla_{TEM-1}$. The wastewater strains ECONIH4 and ECONIH5, isolated from the United States, appeared to possess the beta-lactamase $bla_{KPC-2,}$ the quinolone resistance gene *QnrA1*, and the sulfonamide resistance gene *sul1*, with the latter strain also harboring the aminoglycoside resistance gene *ANT(2'')-Ia*, the beta-lactamase $bla_{FOX-5,}$ and the trimethoprim resistance gene *dfrA19*. Finally, the Chinese wastewater strain SZ4 was found to harbor the quinolone resistance gene *QnrS1*, the sulfonamide resistance gene *sul2*, the tetracycline resistance gene *tet(A)*, and the trimethoprim resistance gene *dfrA14*. Importantly, aside from $bla_{EC-15}$, all antibiotic resistance genes identified in the wastewater strains (i.e., ABWA45, ECONIH4, ECONIH5, SZ4) were determined to be localized onto a plasmid sequence (Table 4-2). Interestingly, this included all antibiotic resistance genes that were identified in SZ4, despite this strain seemingly lacking plasmids when screened against the PLASMIDFINDER database.

The naturalized wastewater strains also appeared to harbor a variety of plasmid-localized microbial defense and stress resistance genes (Table 4-3). This included a wide array of naturalized-associated toxin-antitoxin genes (*ccdAB*, *higAB*, *parDE*, *pemIK*, *vapC*) and resistance genes involved in responses against DNA-damaging stimuli (*dam, recF*, *umuCD*), heat shock (*clpC, ftsH4, hspA, htpX*), heavy metals (*arsABCD*, *bhsA*, *copABDR*, *cusABCFRS*, *merACPRT*, *pcoCE*, *silEP*), oxidative stress (*alx*, *msrAB*, *namA*, *nemAR*, *stiP*, *ydiV*) and other environmental stressors (*ves*). Additionally, other plasmid-localized genes not previously identified through the pan-genome analyses were discovered, including several genetic determinants that could underscore the naturalization and niche-specificity of the wastewater and meat plant strains within their respective engineered niches. Both wastewater and meat plant groups, for instance, were found to harbor the *hha* regulatory protein involved in facilitating the transition of cells into

**Table 4-2.** Localization of key antibiotic resistance genes identified in naturalized strains as either chromosomal or plasmid in origin.

| Strain | Antibiotic Resistance Gene | RFPlasmid Prediction | | |
|---|---|---|---|---|
| | | **Localization Prediction** | **Prediction Probability for Chromosome** | **Prediction Probability for Plasmid** |
| ABWA45 | *rmtB* | Plasmid | 0.022 | 0.978 |
| | $bla_{TEM-1}$ | Plasmid | 0.022 | 0.978 |
| | $bla_{EC-15}$ | Chromosomal | 0.878 | 0.122 |
| ECONIH4 | $bla_{KPC-2}$ | Plasmid | 0.015 | 0.985 |
| | *QnrA1* | Plasmid | 0.003 | 0.997 |
| | *sul1* | Plasmid | 0.003 | 0.997 |
| | $bla_{EC-15}$ | Chromosomal | 0.969 | 0.031 |
| ECONIH5 | *ANT(2'')-Ia* | Plasmid | 0.005 | 0.995 |
| | $bla_{FOX-5}$ | Plasmid | 0.005 | 0.995 |
| | $bla_{KPC-2}$ | Plasmid | 0.021 | 0.979 |
| | *QnrA1* | Plasmid | 0.005 | 0.995 |
| | *sul1* | Plasmid | 0.005 | 0.995 |
| | *dfrA19* | Plasmid | 0.005 | 0.995 |
| | $bla_{EC-15}$ | Chromosomal | 0.969 | 0.031 |
| SZ4 | *QnrS1* | Plasmid | 0.028 | 0.972 |
| | *sul2* | Plasmid | 0.080 | 0.920 |
| | *tet(A)* | Plasmid | 0.028 | 0.972 |
| | *dfrA14* | Plasmid | 0.028 | 0.972 |
| | $bla_{EC-15}$ | Chromosomal | 1 | 0 |

**Table 4-3.** List of plasmid-localized microbial defense and stress resistance genes identified in the naturalized strains.

| Functional Category | Gene(s) | Description | Representation Across Naturalized Strains' Plasmids |
|---|---|---|---|
| **Microbial Defense Mechanisms** | | | |
| Restriction-Modification Systems | *dcm* | DNA cytosine methyltransferase that interferes with cleavage by the *EcoRII* restriction enzyme | Meat Plant Dominant |
| Toxin-Antitoxin Systems | *ccdAB*[1] | DNA gyrase inhibiting toxin (*ccdB*) and its cognate antitoxin (*ccdA*) that also functions in plasmid maintenance through post-segregational killing of cells (i.e., plasmid addiction system) | Meat-Plant Dominant |
| | *higAB*[1] | mRNA interferase toxin (*higB*) and its cognate antitoxin (*higA*) | Common Across Wastewater and Meat Plant |
| | *hokC* | Membrane potential-disrupting toxin that also functions in plasmid maintenance through post-segregational killing of cells | Meat-Plant Dominant |
| | *parDE*[1] | DNA gyrase inhibiting toxin (*parE*) and its cognate antitoxin (*parD*) that also functions in plasmid maintenance through post-segregational killing of cells | Common Across Wastewater and Meat Plant |
| | *pemIK*[1] | Endoribonuclease toxin (*pemK*) and its cognate antitoxin (*pemI*) that also functions in plasmid maintenance through post-segregational killing of cells | Wastewater-Dominant |
| | *vapC*[1] | tRNA(fMet)-specific endonuclease toxin that inhibits translation | Common Across Wastewater and Meat Plant |
| **Stress Resistance** | | | |
| DNA-Damaging Stimuli | *blc* | Outer membrane lipoprotein involved in membrane maintenance under stressful conditions | Meat Plant-Dominant |
| | *dam*[1] | DNA adenine methylase that plays a role in DNA repair | Common Across Wastewater and Meat Plant |
| | *recF*[1] | DNA replication and repair protein involved in the induction of the SOS response | Wastewater-Dominant |

| | | | |
|---|---|---|---|
| | *umuCD*[1] | Putative proteins involved in the SOS response and responses to DNA damaging stimuli such as UV, particularly in mediating DNA repair | Common Across Wastewater and Meat Plant |
| | *uspAB* | Universal stress proteins involved in mediating resistance to DNA-damaging agents | Wastewater-Specific |
| Heat Shock | *clpC*[1] | Member of a stress-induced multi-chaperone system involved in the recovery of the cell from heat-induced damage | Common Across Wastewater and Meat Plant |
| | *ftsH4*[1] | ATP-dependent zinc metallopeptidase that plays a role in the quality control of cytoplasmic and integral membrane proteins | Common Across Wastewater and Meat Plant |
| | *hspA*[1] | Small heat shock protein that appears to play a role as a chaperone that prevents protein aggregation and facilitates proper protein refolding | Common Across Wastewater and Meat Plant |
| | *htpX*[1] | Membrane-localized protease involved in the quality control of integral membrane proteins | Wastewater-Dominant |
| Heavy Metals | *arsABCD*[1] | Arsenic reductase and efflux pump system that mediates resistance to arsenical compounds | Wastewater-Dominant |
| | *bhsA*[1] | Stress resistance protein that reduces membrane permeability to copper; also appears to regulate biofilm formation | Meat Plant-Dominant |
| | *copABDR*[1] | Members of a copper sequestration and efflux system that mediates resistance to copper | Wastewater-Dominant |
| | *cusABCFRS*[1] | Structural and regulatory components of a cation efflux system that mediates resistance to copper and silver | Common Across Wastewater and Meat Plant |
| | *merACPRT*[1] | Structural and regulatory components of a mercury efflux system that mediates resistance to mercuric compounds | Wastewater-Dominant |
| | *pcoCE*[1] | Copper binding and sequestration proteins involved in mediating resistance to copper | Common Across Wastewater and Meat Plant |
| | *silEP*[1] | Silver-binding and exporting proteins involved in a cation efflux system that mediates resistance to silver | Common Across Wastewater and Meat Plant |
| Oxidative Stress | *alx*[1] | Putative membrane-bound redox modulator | Wastewater-Dominant |

| | | | |
|---|---|---|---|
| | *gor* | Glutathione reductase that acts to maintain high levels of reduced glutathione, an antioxidant, in the cell cytosol | Wastewater-Dominant |
| | *msrAB*[1] | Peptide methionine sulfoxide reductases that function as repair enzymes for proteins that have been inactivated by oxidation | Common Across Wastewater and Meat Plant |
| | *mshA* | Involved in the biosynthesis of mycothiol, a major protective thiol found in *Actinobacteria* that acts as an important antioxidant, like glutathione, for the detoxification of alkylating agents, reactive oxygen and nitrogen species, and antibiotics | Wastewater-Dominant |
| | *namA*[1] | NADPH dehydrogenase that appears to be involved in detoxification and the oxidative stress response | Meat Plant-Dominant |
| | *nemAR*[1] | N-ethylmaleimide reductase (*nemA*) and associated transcriptional repressor (*nemR*) involved in responses against reactive electrophilic and chlorine species | Meat Plant-Dominant |
| | *rclC* | Inner membrane protein involved in mediating resistance to reactive chlorine species | Wastewater-Dominant |
| | *stiP*[1] | Cysteine protease that plays a role in regulating cell morphology in response to oxidative damage | Wastewater-Dominant |
| | *ydiV*[1] | Anti-flagellar regulatory protein that appears to mediate resistance to host immune response-associated oxidative stress | Meat Plant-Dominant |
| | *yfcG* | Disulfide-bond oxidoreductase involved in responses against oxidative stress | Wastewater-Dominant |
| Misc. | *hha* | Regulatory protein that binds DNA in response to environmental stimuli to facilitate the formation of persister cells | Common Across Wastewater and Meat Plant |
| | *ves*[1] | Putative protein that appears to mediate cellular responses to cold shock | Wastewater-Dominant |
| | *ydeO* | Transcriptional regulator that appears to regulate the expression of genes involved in acid resistance | Meat-Plant Dominant |

[1] Genes previously determined to be naturalized-associated via Scoary

persister states that can enhance survival against various environmental stressors (Berkvens *et al.* 2022). Furthermore, several wastewater strains were found to harbor additional plasmid-localized stress resistance genes against DNA-damaging agents (*uspAB*) and oxidative stress (*gor*, *mshA*, *rclC*, *yfcG*), which could be important for surviving the extreme stressors employed during wastewater treatment (e.g., UV disinfection). Similarly, several meat plant strains possessed additional microbial defense genes associated with restriction-modification systems (*dcm*) and toxins (*hokC*), as well as stress resistance genes against DNA-damaging agents (*blc*) and acid stress (*ydeO*), which could promote their survival during meat processing.

## 4.4 Discussion

The evidence presented in Chapters Two and Three of this thesis have described a novel, emerging *E. coli* ecotype that appears to have evolved to exploit various engineered environments, such as wastewater treatment plants (Zhi *et al.* 2019) and meat processing facilities (Yang *et al.* 2021), as primary niches. The genotypic, phylogenetic, and ecotypic evidence suggest that *E. coli* strains belonging to this naturalized-engineered-specific ecotype have diverged from their host-associated and environmental counterparts; however, the specific adaptive processes allowing these wastewater- and meat plant-associated strains to survive and persist within man-made, built environments were unclear. As such, we sought to characterize the specific genetic features that could underlie the adaptation and potential niche-specificity of the naturalized, WWS- and MPS-*E. coli* strains.

Genome mapping revealed that the naturalized strains harbored several unique genetic regions, potentially representing general genetic adaptations that could be required for survival within non-host, engineered environments. A subset of these genetic regions were found to be

unique to either of the wastewater or meat plant strains, suggesting that these ecotypic groups may be further characterized by additional niche-specific genetic adaptations corresponding to the specific conditions and stressors encountered within the wastewater or meat plant niche. Supporting this, pan-genome spectrum function analyses revealed that the wastewater and meat plant strains were genetically heterogenous when compared to the other *E. coli* ecotypes (i.e., enteric, InPEC, ExPEC, environmental, etc.), suggesting they harbored unique, and presumably ecologically relevant, genetic repertoires.

Indeed, pan-genome-wide association studies demonstrated that the wastewater and meat plant strains were enriched in a multitude of genes associated with biofilm formation, microbial defense, and stress resistance – functions that could be advantageous in non-host contexts and especially within the harsh conditions of man-made environments. For instance, the use of alternative fimbrial systems and additional biofilm regulators could enhance the ability of the naturalized strains to form biofilms, thereby increasing their tolerance to the extreme stressors (i.e., low temperatures, disinfecting agents [i.e., chlorine, advanced oxidants, UV], etc.) encountered in engineered environments (Chattopadhyay *et al.* 2022; Fernández-Gómez *et al.* 2022; Yin *et al.* 2019). In line with this, previous work conducted by Zhi *et al.* (2017) found that naturalized wastewater *E. coli* strains were particularly robust biofilm producers, forming biofilms at roughly three times the capacity of their enteric counterparts. The naturalized-engineered strains also harbored an abundance of microbial defense genes, including various toxin-antitoxin systems to survive the intense inter-microbial competition against the complex microbial communities within wastewater matrices (Cydzik-Kwiatkowska and Zielińska 2016) and meat processing facilities (Zwirzitz *et al.* 2020). Moreover, these strains were most notably characterized by an over-abundance of stress resistance genes, including those mediating responses against DNA-damaging

stimuli (i.e., UV radiation), oxidative stress (i.e., oxidants, reactive oxygen species, reactive electrophilic species, chlorine), heat shock (i.e., composting of human biosolids, steam pasteurization), and heavy metals. Importantly, these genotypic features appear to be reflected in phenotype as these strains have been found to exhibit enhanced resistance to disinfection-related stressors, including against high temperatures of up to 60°C (Wang *et al.* 2020; Yang *et al.* 2021), as well as advanced oxidants and chlorine (Wang *et al.* 2020).

Beyond these shared genetic signatures, the wastewater and meat plant strains were each found to exhibit additional, niche-specific genetic adaptations. For instance, while both groups were found to be collectively enriched with microbial defense genes, the wastewater strains were enriched with restriction-modification and CRISPR-Cas systems that could confer protection against the heavy load of phages and other invasive mobile genetic species present in sewage and wastewater (Ballesté *et al.* 2022; Runa *et al.* 2021; Strange *et al.* 2021). Similarly, each group appeared to possess specific stress resistance mechanisms reflective of their corresponding niches. The wastewater strains, for example, were found to harbor additional heavy metal resistance systems against arsenic and mercury – common heavy metal constituents within wastewater (Qasem *et al.* 2021; Suess *et al.* 2020; Ungureanu *et al.* 2015). In contrast, the meat plant strains were found to harbor the chlorine resistance genes *nemA* and *nemR*, which could enhance resistance against the bleach-based sanitizers commonly used in food processing operations.

In conjunction with these niche-relevant genetic adaptations, the naturalized-engineered strains were found to simultaneously lack several host-adaptive genetic features. The wastewater and meat plant strains, for instance, lacked various metal acquisition and secretion system genes that could enhance fitness within a host (Garcia *et al.* 2011; Ho *et al.* 2008; Serapio-Palacios *et al.* 2022; Slater *et al.* 2018). Furthermore, these strains also lacked various genes that would be

required for colonizing the gastrointestinal tract, including for survival during passage through the stomach (Tramonti *et al.* 2006), key colonization factors (Garnett *et al.* 2012; Steenbergen *et al.* 2009), and protection against the toxic metabolites produced within by the gut microbiome (Cueva *et al.* 2010). Interestingly, some of these host-adaptive genes were found to be negatively correlated with the biofilm formation, defense, and stress resistance genes that were over-represented amongst the naturalized strains. As such, it appears that the genetic adaptations acquired to tolerate the harsh conditions encountered within the engineered environment may have come at the cost of fitness within the host (i.e., antagonistic pleiotropy).

While the wastewater and meat plant strains assessed in this study do not appear to be pathogenic or even host-associating, their characterization highlights a concerning public health prospect: that microbes could be evolving resistance to disinfection. Concerningly, studies have provided evidence reinforcing this possibility, as *E. coli* populations repeatedly exposed to monochloramine water treatment were found to develop resistance to disinfection (Daer *et al.* 2022), with >60% of cells remaining viable after treatment (Daer *et al.* 2021). Importantly, while this present study focuses on naturalized *E. coli*, the same selective pressures are likely operating for the other microbes present in the microbiomes that are resident within these engineered environments, including pathogenic *E. coli*. For instance, given that many of the genes over-represented in the naturalized strains were duplicates within the pan-genome (Supplementary Table 4-S1), other *E. coli* populations could similarly acquire and amplify these genes to modify their capacity to respond to the extreme stressors (i.e., disinfection) encountered in the engineered environment (Kondrashov 2012).

Alternatively, naturalized-engineered *E. coli* populations that are resident within wastewater or meat plants may serve as reservoirs for the dissemination of resistance determinants

to other microbial populations through horizontal gene transfer. Indeed, engineered environments are presently recognized as emerging hotspots for microbial evolution, particularly through the exchange of genetic material amongst resident microbial communities (Hutinel *et al.* 2019; Vinayamohan *et al.* 2022). This is most apparent for antibiotic resistance, with wastewater being an increasingly important environmental hotspot for the evolution and mobilization of antibiotic resistance within the microbial world (Berglund *et al.* 2023; Uluseker *et al.* 2021). Reflecting this, aside from a distinct $bla_{EC\text{-}15}$ variant of the chromosomal *blaEC* family of beta-lactamases (Schmidt *et al.* 2023) that appeared to be associated with *E. coli* strains specifically derived from engineered environments, a select number of the wastewater strains were found to harbor additional antibiotic resistance genes. This included several beta-lactamases ($bla_{TEM\text{-}1}$, $bla_{KPC\text{-}2}$, $bla_{FOX\text{-}5}$), as well as various genes mediating resistance against aminoglycosides (*rmtB*, *ANT(2'')-Ia*), quinolones (*qnrA1*, *qnrS1*), sulfonamides (*sul1*, *sul2*), tetracyclines (*tet(A)*) and trimethoprims (*dfrA19*, *dfrA14*) – many of which are common antibiotic resistance constituents within wastewater (Karkman *et al.* 2018; Mutuku *et al.* 2022). Importantly, these antibiotic resistance genes appeared to be localized on plasmid sequences, suggesting these WWS strains might participate in the dissemination of antibiotic resistance determinants within the wastewater environment.

An important extension to this finding is that these naturalized-engineered strains could also be transmitting key genetic determinants that could enhance the survival of other *E. coli* populations within engineered environments. In line with this, several genes that appeared to be niche-adaptive within the wastewater and meat plant strains were also found to be located on plasmid sequences, suggesting that they may also be transferable to other microbes present within the wastewater or meat plant niche. Concerningly, several of these genetic adaptations appear to enhance survival within these engineered environments, including cold shock (i.e., low ambient

temperatures), heavy metals, microbial competition (i.e., sludge digestion during wastewater treatment), DNA-damaging agents (i.e., UV radiation), oxidizing agents (i.e., advanced oxidants, chlorine, bleach-based sanitizers), and heat (i.e., sludge digestion, biosolids composting). As such, beyond their role in the dissemination of antibiotic resistance, these findings indicate that naturalized *E. coli* populations within engineered environments could also be facilitating the emergence and evolution of disinfection resistance within the microbial world.

The findings presented in this chapter recapitulate the hypothesis that naturalized *E. coli* populations residing within wastewater treatment and meat processing plants represent unique ecotypes. Indeed, both WWS- and MPS-*E. coli* groups were found to harbor distinct genetic adaptations that appear to enhance their survival against the specific stressors encountered within the wastewater or meat plant environment. In the next chapter we link these genotypic findings with discernable phenotypes, as we demonstrate that these naturalized-engineered strains (i.e., specifically the WWS-*E. coli*) can be characterized by various phenotypic adaptations that appear to enhance their persistence and evolutionary success within the engineered environment.

# Chapter Five: Characterization of Thermotolerance, Temperature-Dependent Growth Kinetics and Biofilm Formation of Naturalized Wastewater *Escherichia coli* Strains

## 5.1 Introduction

While the diversification of the *E. coli* species into niche-specific ecotypes can be readily traced through phylogenetic, genotypic, and ecotypic means, the polyphasic approach emphasizes that these signatures of niche-specificity should be recapitulated through discernable, niche-adaptive phenotypes. For instance, the differential expression of various species-specific colonization factors (Dubreuil *et al.* 2016; von Mentzer and Svennerholm 2023; Ron 2006), appears to coincide with the host-preference or host-restriction of various host-adapted *E. coli* strains (Allen *et al.* 2006; Ron 2006; Tobe and Sasakawa 2002; Wenzel *et al.* 2017; Zhang *et al.* 2019a). Similarly, the genotypic divergence of environmental *E. coli* strains from enteric populations (Byappanahalli *et al.* 2012; Jang *et al.* 2011; Jang *et al.* 2015; Tymensen *et al.* 2015) appears to be reflected in various environmentally adaptive traits including enhanced survival (Ishii *et al.* 2006; Ishii *et al.* 2010) and even growth (Brennan *et al.* 2013) at lower temperatures, as well as the production of capsules that confer protection against environmental stressors such as UV radiation and desiccation (Power *et al.* 2005; Touchon *et al.* 2020).

The preceding chapters in this thesis have provided extensive phylogenetic, genotypic, and pan-genomic evidence describing a novel *E. coli* ecotype that has emerged within various engineered environments. The logic regression analyses presented in Chapter Two, for instance, demonstrated that wastewater-derived *E. coli* strains harbor highly informative and ecotype-

specific SNP-SNP biomarkers within several ITGRs across the *E. coli* genome. In Chapter Three, these wastewater strains were found to cluster into distinct sequence type lineages, including ST635 and ST399, and several predicted naturalized ecotypic groups that were largely separate from other host-associated (i.e., enteric, ExPEC) and environmental ecotypes. Interestingly, although the wastewater strains could not be definitively distinguished from their meat plant-derived counterparts through phylogenetic or ecotype prediction means, the pan-genomic analyses presented in Chapter Four revealed that these strains still harbored various niche-specific genetic adaptations, particularly relating to microbial defense and resistance to stressors including heat shock, DNA damaging stimuli, oxidative stress, and heavy metals.

Ecologically, the wastewater niche represents a distinctly stressful environment that could drive the divergence and evolution of WWS-*E. coli* strains towards niche-specificity. Microbes that are present in wastewater are exposed to various contaminants, including antibiotics and antibiotic residues (Rodriguez-Mozaz *et al.* 2020; Samrot *et al.* 2023), pharmaceuticals, detergents (Mousavi and Khodadoost 2019), and heavy metals (Qasem *et al.* 2021), as well as biological hazards including microbial competitors, predators (Cydzik-Kwiatkowska and Zielińska 2016) and infective phages (Ballesté *et al.* 2022; Runa *et al.* 2021; Strange *et al.* 2021), that threaten their survival. The wastewater treatment train itself presents additional challenges, including the removal and deprivation of nutrients from wastewater (Chahal *et al.* 2016), biosolids composting (i.e., heat treatment) (Gerba and Pepper 2009), as well as tertiary treatment (i.e., oxidative stress, UV irradiation, and chlorination) that can exert strong selective pressures on the evolution of wastewater-borne microbes. Thus, as the WWS-*E. coli* strains appear to harbor various genotypic characteristics that are specific to the wastewater niche, these genetic signatures should be recapitulated phenotypically through specific adaptations and survival strategies that would allow

them to tolerate the stressors encountered within wastewater and during wastewater treatment. Importantly, previous studies have reinforced some of these genotypic findings in phenotype, as naturalized WWS-*E. coli* strains have been found to exhibit enhanced resistance to chlorine and other oxidants (Wang *et al.* 2020; Zhi *et al.* 2017), suggesting they may have evolved resistance to the disinfection practices that have been designed to eliminate them.

Beyond resistance to disinfection, however, the success of the WWS strains within the wastewater niche suggests that they have likely acquired additional phenotypic adaptations, not necessarily associated with tertiary wastewater treatment, that could underlie their niche-specificity. This appears to include the ability to withstand heat (Wang *et al.* 2020), such as used in the management of biosolids, as well as other physicochemical stressors associated with the wastewater environment, including low temperature and low nutrient conditions. As such, we sought to explore whether additional phenotypic adaptations could underlie the survival and evolutionary success of the naturalized wastewater strains within the sewage treatment plant niche.

## 5.2 Materials and Methods

### 5.2.1 Bacterial Strains

A set of representative naturalized wastewater, host-associated, and reference *E. coli* strains were phenotypically characterized to assess whether the naturalized wastewater *E. coli* could be better adapted to a non-host niche than their host-associated counterparts. Included in this representative set were: (i) four naturalized wastewater *E. coli* strains, including two (WW10, WW69) that were previously found to harbor the *uspC*–IS30–*flhDC* locus (Zhi *et al.* 2016a; Zhi *et al.* 2019) and the tLST (Wang *et al.* 2020), and two (WW2, WW9) that lacked both the *uspC*–IS30–*flhDC* biomarker and tLST but were still classified as naturalized wastewater strains through

logic regression (Zhi *et al.* 2016a; Zhi *et al.* 2019); (ii) three host-associated *E. coli* strains, including one human enteric strain deficient in the *rpoS*-mediated generalized stress response (H51), one human enteric strain with a robust *rpoS*–mediated generalized stress response (H54), and one reference clinical ExPEC strain (CFT073); and (iii) one laboratory reference strain (ATCC 25922). All relevant information on the *E. coli* strains assessed in this chapter are included in Table 5-1.

### 5.2.2 Heat Resistance and Upper Thermal Tolerance Assays

Part of the wastewater treatment process involves the generation and subsequent treatment of sludge and biosolid byproducts. In particular, processes such as sludge digestion and biosolids composting can employ temperatures ranging from 30–80°C (Gerba and Pepper 2009), suggesting that heat could be an additional selective pressure driving the evolution and adaptation of microbes that are resident within the wastewater matrix. As such, building on previous work (Wang *et al.* 2020) we further explored heat resistance as a possible additional phenotypic adaptation that could underlie the niche-adaptation and specificity of the naturalized wastewater strains.

All *E. coli* isolates were first cultured for 24 hours in 5mL of tryptic soy broth (TSB, BD Difco™) at 37°C with shaking at 200 rpm in a shaking incubator (Innova™ 42 Incubator Shakers, New Brunswick Scientific™, Eppendorf). To harvest the cells, 1mL aliquots of each overnight culture were centrifuged at 7,000 x g for 5 minutes. The resulting bacterial pellets were then washed twice with phosphate-buffered saline (PBS, Cytiva, HyClone™), and then re-suspended and diluted to a final concentration of approximately $10^{8\text{-}9}$ cells per mL. in 1mL of PBS. One-hundred µL of each diluted bacterial suspension was aliquoted into 0.2mL PCR tubes in triplicate, and then subject to heat treatment. To evaluate the relative degrees of heat resistance between the

**Table 5-11.** Description of *E. coli* strains selected for phenotypic characterization in this chapter

| Strain | Ecotype | Description | Reference |
|---|---|---|---|
| WW10 | Naturalized Wastewater | *uspC*–IS30–*flhDC* and tLST-positive wastewater strain | Zhi *et al.* 2016a |
| WW69 | Naturalized Wastewater | *uspC*–IS30–*flhDC* and tLST-positive wastewater strain | Zhi *et al.* 2016a |
| WW9 | Naturalized Wastewater | *uspC*–IS30–*flhDC* and tLST-negative wastewater strain | Zhi *et al.* 2016a |
| WW2 | Naturalized Wastewater | *uspC*–IS30–*flhDC* and tLST-negative wastewater strain | Zhi *et al.* 2016a |
| H51 | Host-Associated | *rpoS*-negative human enteric strain | Zhi *et al.* 2017 |
| H54 | Host-Associated | *rpoS*-positive human enteric strain | Zhi *et al.* 2017 |
| CFT073 | Host-Associated | clinical ExPEC strain | - |
| ATCC 25922 | Reference | laboratory reference strain | - |

naturalized wastewater and host-associated *E. coli* strains, two independent approaches were used based on the methods laid out by Wang *et al.* (2020) with some slight modifications. First, a general assessment of heat resistance was performed by subjecting the bacterial suspensions to elevated temperatures ranging from 60–66°C for 5 minutes using an Applied Biosystems® 2720 Thermal Cycler, in triplicate. In addition to general heat resistance, the upper thermal tolerance (i.e., highest temperature of survival following short-term exposure) of each strain was also evaluated. To evaluate upper thermal tolerance, in an independent set of trials the same strains were exposed to temperatures ranging from 40–74°C for 30 seconds with an Applied Biosystems® 2720 Thermal Cycler, again in triplicate.

To enumerate the surviving cells following heat treatment, both heat resistance assays utilized a modified three-tube most probable number (MPN) spot assay. Briefly, each heat-treated suspension was aliquoted into a 96-well plate, after which ten-fold serial dilutions were performed with PBS. Five µL aliquots of each dilution were then spotted onto tryptic soy agar (TSA, BD Difco™) plates and incubated at 37°C for 24 hours. The presence of colonies was taken to be indicative of surviving cells capable of growth, and the number and pattern of spotted areas displaying colony growth was then converted to MPN counts (Banting *et al.* 2016) and back-calculated to a final cell count in MPN/mL (see Figure 5-1 for example of spot plates). For both the general heat resistance and upper thermal tolerance assays, the surviving cell counts in MPN/mL after each heat treatment for each strain were then pooled to form three comparison groups: (i) a naturalized wastewater group (WW10, WW69, WW9, WW2); (ii) a host-associated group (H51, H54, CFT073); and (iii) a reference group (ATCC 25922). The average surviving cell counts at each temperature between these three groups were then compared and statistically evaluated serially in a pairwise fashion using ANOVA tests.

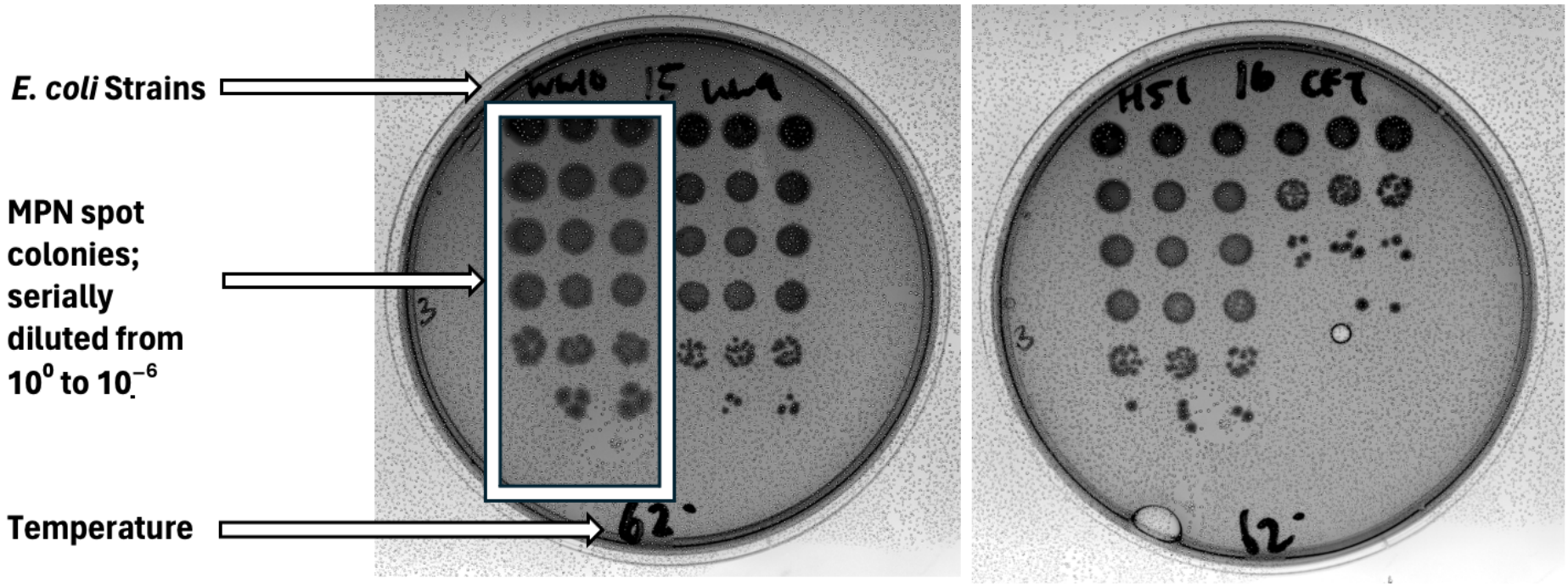



**Figure 5-1. Example of MPN spot plates that were used to enumerate surviving cells following heat treatment.** Following heat treatment and serial dilutions, 5µL of each bacterial suspension was spotted on tryptic soy agar plates in triplicate, as shown. Following incubation at 37°C for 24 hours, any surviving cells would presumably grow and form colonies in the spotted areas. The pattern of positive spots would then be correlated to MPN counts using the MPN conversion tables used for the standard three-tube MPN assay, but adapted for this spot assay (i.e., substituting each column of spots as a set of tubes). The resulting MPN counts would then be used for back-calculations to determine the surviving concentration of cells in MPN/mL.

### 5.2.3 Temperature- and Media-Dependent Growth Assays

The restriction of the WWS-*E. coli* to the wastewater niche implies that they may have acquired the ability to grow within the wastewater environment. Indeed, other naturalized *E. coli* populations have previously been reported to grow in natural environmental contexts (Brennan *et al.* 2013; Jang *et al.* 2017; Power *et al.* 2005), suggesting that the wastewater strains may be better able to sustain their population numbers in non-host (i.e., low temperature, low nutrient) conditions compared to their host-associated counterparts. To evaluate this, each strain was grown in different growth media and at different temperatures. As a proxy for a nutrient-rich (i.e., host-like) environment, TSB was selected as a general nutrient-rich enrichment medium. In contrast, to represent a nutrient-limiting (i.e., non-host-like) environment, all strains were also grown in Davis Minimal Media broth (MM, Millipore®). Additionally, strains were also grown at two different temperatures, including 37°C, representing the typical temperature of the mammalian gut environment, and 25°C, representing non-host, ambient temperatures.

To set up the growth assays, each strain was first cultured for 24 hours at 37°C in 5mL of the specific growth medium to be used to generate each growth curve, with shaking at 200 rpm in a shaking incubator. Cells were harvested by taking 1mL aliquots of each overnight culture and centrifuging them at 7,000 x g for 5 minutes, following which the resulting bacterial pellets were washed twice and then re-suspended in 1mL of PBS. The bacterial suspensions were diluted to achieve a concentration of ~$10^7$ cells/mL in PBS, after which 20 µL of each were inoculated into 180µL of growth medium in 96-well microplates (Fisher Scientific, Costar™), in replicates of 5, to achieve a final starting concentration of ~$10^6$ cells/mL for each strain. The plates were then incubated at either 37°C or 25°C for 24 hours with shaking at 200 rpm in a fluorescent microplate reader (BMG Labtech, FLUOstar® Omega). Growth assays were then run twice for each strain in

each growth medium and at each growth temperature.

Cell density measurements, in OD600 (i.e., optical density measured via absorbance at 600nm), were taken every 10 minutes over the 24-hour period for each strain. For each time period, the OD600 values for the strains were pooled according to ecotype and then plotted to generate growth curves for each of the naturalized wastewater, host-associated, and reference (ATCC 25922) comparison groups. To assess whether each of these groups displayed differences in their growth characteristics, the Compare Groups of Growth Curves (CGCC) test was performed. Briefly, the CGCC statistical test performs a series of T-tests to compare two growth curves at each time point, after which the resulting T-statistics are averaged to obtain a permutation p-value. CGGC statistical tests were performed in a pairwise fashion between the growth curves for each set of growth conditions (i.e., medium and temperature) with 1000 permutations using the CGCC webserver (Baldwin *et al.* 2007; Elso *et al.* 2004).

### 5.2.4 Temperature- and Media-Dependent Biofilm Formation Assays

In Chapter 4, the WWS-*E. coli* strains were found to harbor an abundance of biofilm formation genes. Thus, these strains could maintain their persistence within the wastewater environment through the production of biofilms that not only offer protection against wastewater-associated environmental stressors, but also allow for the establishment of a resident population within the wastewater niche. While previous studies have evaluated the biofilm formation capacity of wastewater-derived *E. coli* strains (Zhi *et al.* 2017), we specifically sought to assess the effects of nutrient availability and temperature on biofilm formation between the naturalized wastewater and host-associated strains.

Each strain was grown for 24 hours in 5mL of TSB at 37°C with shaking at 200 rpm in a

shaking incubator. One-mL aliquots of each culture were taken and centrifuged at 7,000 x g for 5 minutes, following which the resulting bacterial pellets were washed twice, re-suspended, and then diluted in 1mL of PBS to achieve a concentration of ~$10^8$ cells/mL. The bacterial suspensions were inoculated into either TSB or MM in triplicate in 96-well microplates (Sigma-Aldrich, Nunc®) to achieve a starting concentration of ~$10^7$ cells/mL. A 96-well pegged-lid (DOJINDO Laboratories) was then submerged into each cell suspension, and the microplates were incubated at either 37°C (according to the manufacturer's instructions) or 25°C for 24 hours. Following incubation, the pegs were washed two times with PBS and then submerged and incubated in crystal violet dye for 30 minutes to stain the resultant biofilms. The pegs were then washed again twice in PBS and then submerged and incubated in 100% ethanol for 15 minutes to elute the captured crystal violet dye. This was performed twice for each strain in each growth medium and temperature.

To measure the relative biofilm formation capacity of each strain, the absorbance at 590nm for each well was measured using a fluorescent microplate reader (BMG Labtech, FLUOstar® Omega). The absorbance values at 590nm were pooled into groups corresponding to the naturalized wastewater, host-associated, and reference (ATCC 25922) groups. To compare biofilm formation capacity, the average absorbance values of each group were plotted for each growth medium and incubation temperature and then statistically evaluated serially for each growth condition and temperature using ANOVA tests.

## 5.3 Results

### 5.3.1 Survivability of Naturalized Wastewater Versus Host-Associated *E. coli* Strains During Prolonged Exposure to Elevated Temperatures

Differences in survivability following prolonged exposure of 5 minutes to elevated

temperatures ranging from 60–66°C were observed between the naturalized wastewater, host-associated, and reference strains. While the WWS strains (i.e., WW10, WW9, WW69, and WW2) could still be cultured following heat exposure, all other host-associated (i.e., H51, H54, and CFT073) and reference (i.e., ATCC 25922) strains appeared to be completely inactivated at all elevated temperatures (Figure 5-2). Interestingly, despite their collective grouping into the 'naturalized wastewater' ecotype, there was considerable variability in the survivability of each individual wastewater strain. For instance, among the wastewater strains, WW2 appeared to be the most susceptible to heat treatment. While all wastewater strains were found to exhibit growth after exposure to 64°C for 5 minutes, WW2 appeared to be completely inactivated at the same temperature (Figure 5-2D). Differences in heat resistance were also observed amongst the remaining wastewater strains. Although WW10, WW9, and WW69 all exhibited growth following heat treatment at 60–64°C, each strain appeared to experience different degrees of inactivation between each temperature increment. WW69, for example, displayed consistent reductions in cell counts as the temperatures increased, including a ~1 $\log_{10}$ reduction at 60°C, followed by a ~3 $\log_{10}$ reduction at 62°C, and then a ~2 $\log_{10}$ reduction at 64°C (Figure 5-2C). In contrast, the cell counts for WW10 and WW9 remained relatively stable up to 62°C; however, cell counts appeared to decrease at 64°C resulting in a ~3 $\log_{10}$ reduction for WW10 (Figure 5-2A) and a ~5 $\log_{10}$ reduction for WW9 (Figure 5-2B). Beyond these individual differences, all strains appeared to be completely inactivated after 5 minutes of exposure to 66°C (Figure 5-2).

Although each strain, regardless of their ecotypic designation, displayed differences in their heat tolerance profiles, a clear distinction in survivability could be observed between the naturalized wastewater, host-associated, and reference groups. When pooled into their respective ecotypic groups, the wastewater strains were the only group that survived following exposure to

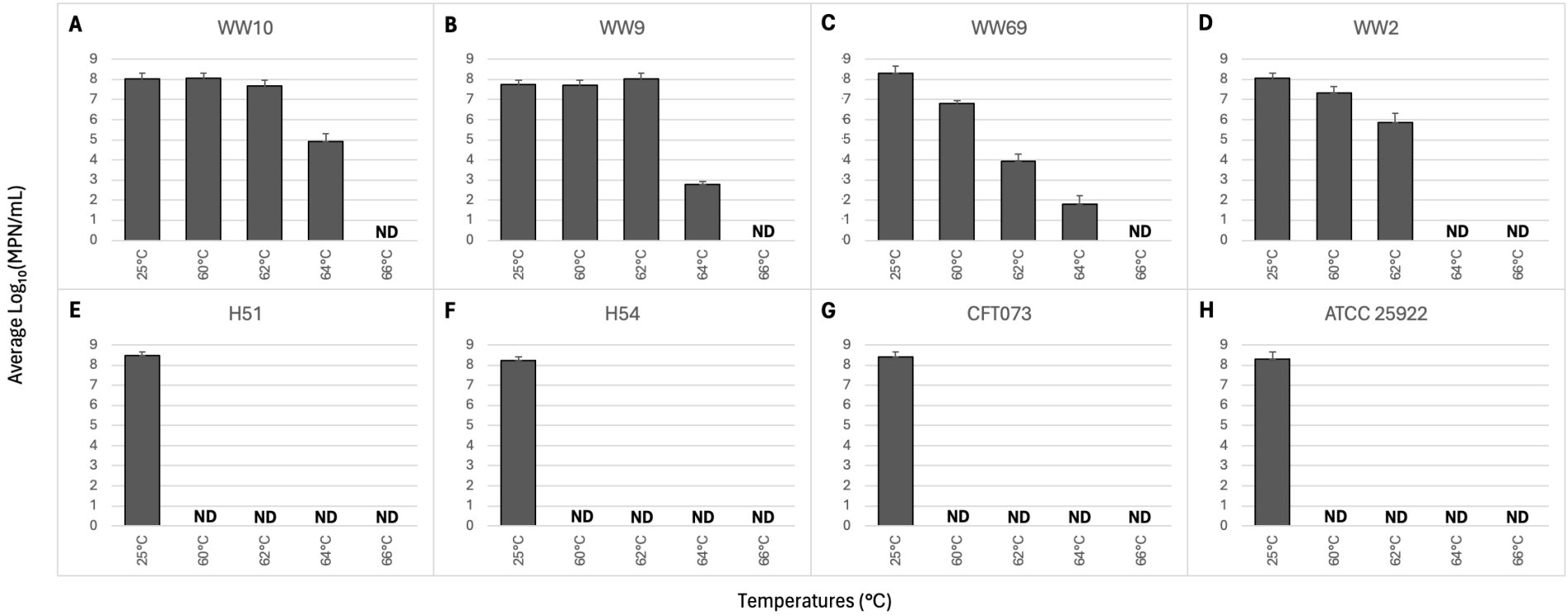



**Figure 5-2. Survival of naturalized wastewater, host-associated, and reference *E. coli* strains after 5-minute exposure to elevated temperatures of 60–66°C.** All naturalized wastewater **(A-D)**, host-associated **(E-G)**, and reference **(H)** *E. coli* strains were cultured overnight in TSB, washed twice and then re-suspended in PBS prior to heat treatment for 5 minutes, which was performed in triplicate. A modified MPN-spot assay was used to enumerate the surviving cells after each temperature in $\log_{10}$(MPN/mL), with the error bars indicating the standard deviation across the three independent heat treatment trials performed. Instances where no surviving cells were detected are indicated with 'ND' (not detected).

60°C and above, until 66°C (Figure 5-3). Indeed, none of the host-associated or reference strains survived following heat treatment, such that the wastewater group was found to exhibit significantly higher surviving cell counts at 60°C ($p = 5.414E-21$), 62°C ($p = 2.565E-7$), and 64°C ($p = 0.008$) compared to their host-associated and reference counterparts.

### 5.3.2 Upper Thermal Tolerance of Naturalized Wastewater Versus Host-Associated *E. coli* Strains

Differences were also observed in the culturability of the strains during the upper thermal tolerance assays, as the naturalized wastewater strains were found to exhibit higher upper thermal tolerances compared to the host-associated and reference strains (Figure 5-4. In particular, the host-associated strain CFT073 (Figure 5-4G) and reference strain ATCC 25922 (Figure 5-4H) appeared to exhibit the lowest heat tolerance, as these strains appeared to survive short-term exposure (30 seconds) to a maximum of only 62°C. Of the two, ATCC 25922 appeared to be more heat-susceptible as it exhibited slight reductions in cell counts (~1 $\log_{10}$) at temperatures as low as 58°C and displayed a greater reduction in surviving cell counts at 62°C (i.e., ~5 $\log_{1o}$ difference from baseline compared to only ~2 $\log_{10}$ for CFT073). The other host-associated strains, H51 (Figure 5-4E) and H54 (Figure 5-4F), were found to survive at slightly higher temperatures, as they could still be cultured following short-term exposure to 64°C. Overall, the naturalized wastewater strains were found to exhibit the highest upper thermal tolerances out of all strains analyzed, though some slight variability was observed between the individual wastewater strains. Out of all wastewater strains, WW69 (Figure 5-4C) and WW2 (Figure 5-4D) were characterized by the lowest maximum tolerable temperature, as these strains could only survive short-term exposure to a maximum temperature of 68°C. Furthermore, between these two strains WW69 appeared to be slightly more

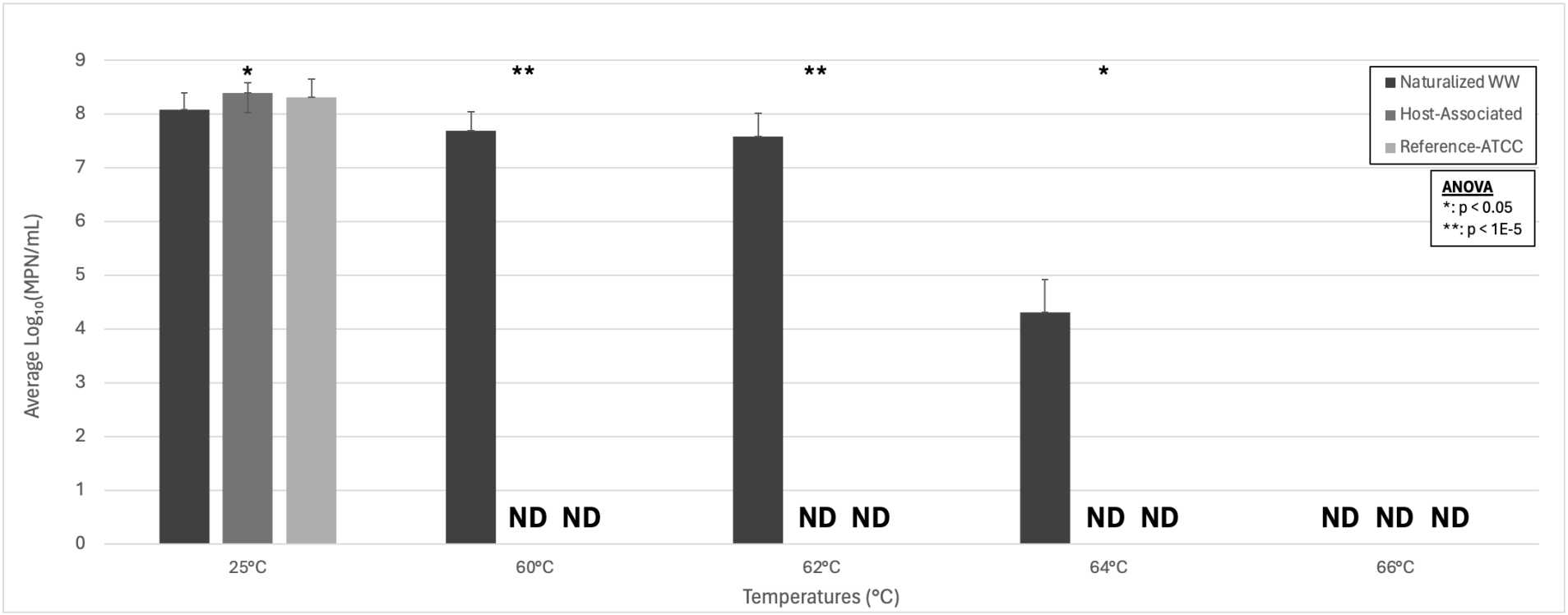



**Figure 5-3. Differences in heat resistance profiles between the naturalized wastewater, host-associated and reference *E. coli* ecotypes.** The surviving cell counts (in $\log_{10}$(MPN/mL)) for each naturalized wastewater (black), host-associated (grey), and reference (light grey) strain following heat treatment (5 minute exposure at each temperature) were pooled into their corresponding ecotypic groups. The average surviving cell counts for each group were then compared at each temperature using Single Factor ANOVA tests, with statistically significant differences between the wastewater, host-associated and reference groups indicated with either a single asterix (*, $p < 0.05$) or double asterixes (**, $p > 1E\text{-}5$). ND – not detected.

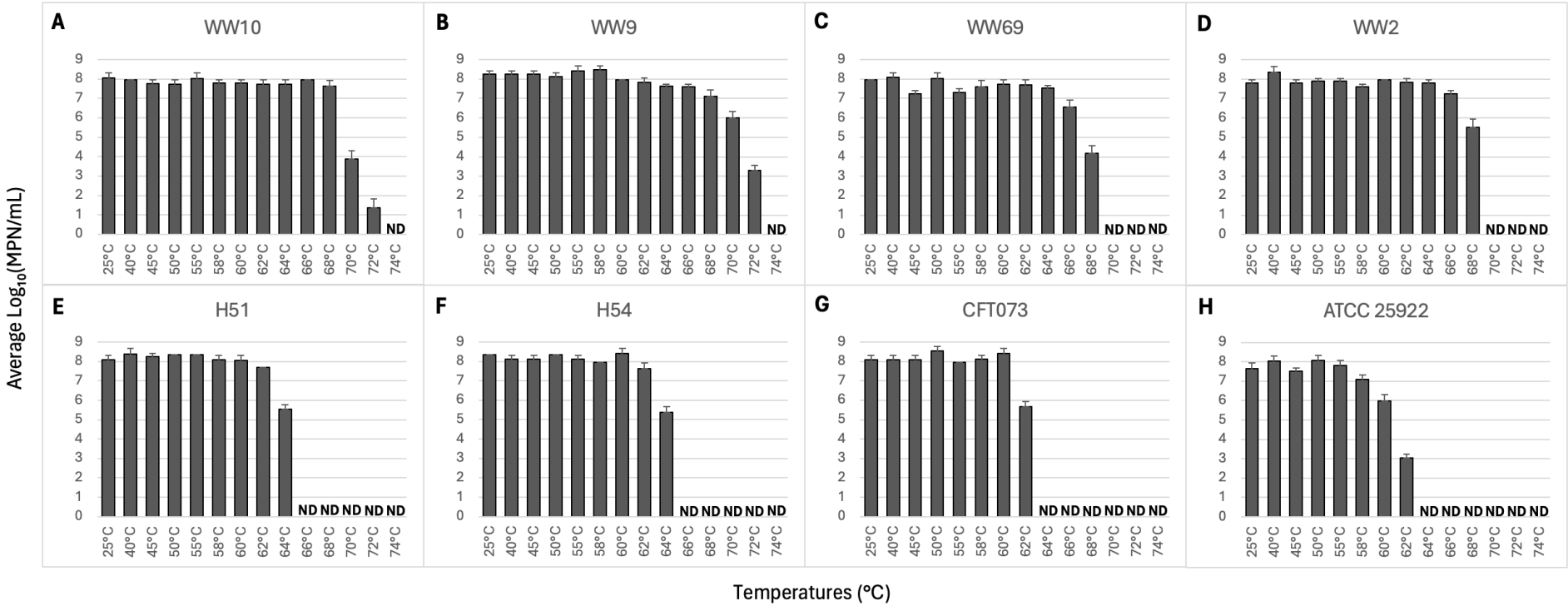



**Figure 5-4. Maximum tolerated temperatures of naturalized wastewater, host-associated, and reference *E. coli* strains following 30-second exposure times.** All naturalized wastewater **(A-D)**, host-associated **(E-G)**, and reference **(H)** *E. coli* strains were cultured overnight in TSB, washed twice and then re-suspended in PBS prior to heat treatment for 30 seconds, which was performed in triplicate. A modified MPN-spot assay was used to enumerate the surviving cells after each temperature in $\log_{10}$(MPN/mL), with the error bars indicating the standard deviation across the three independent heat treatment trials performed. Instances where no surviving cells were detected are indicated with 'ND' (not detected).

susceptible to heat as it exhibited greater reductions in surviving cell counts compared to WW2 (~4 $\log_{10}$ versus ~2-3 $\log_{10}$) at 68°C. In contrast, WW10 (Figure 5-4A) and WW9 (Figure 5-4B) appeared to exhibit the highest upper thermal tolerances, as both strains were able to withstand 30-second exposures to 72°C. WW9, however, appeared to be slightly more tolerant to these elevated temperatures, as it had a higher surviving cell count (~$10^3$ MPN/mL) compared to WW10 (~$10^1$ MPN/mL). Outside of these differences, no strains could be cultured following short-term exposure to 74°C (Figure 5-4).

As a group, the naturalized wastewater strains were found to exhibit higher upper thermal tolerances compared to their host-associated and reference counterparts (Figure 5-5). Although all groups were able to tolerate short-term exposure to temperatures of up to 62°C, differences in the overall survival between the naturalized wastewater, host-associated, and reference groups were observed. Specifically, the reference group appeared to be the least heat resistant, as its surviving cell counts were found to be significantly lower than the other two groups after short-term exposure to 58°C ($p = 0.003$), 60°C ($p = 2.811E-9$), and 62°C ($p = 6.860E-9$), following which the reference group could no longer be cultured at temperatures of 64°C ($p = 2.770E-7$) and above. Following this, as the temperature increased, only the wastewater group continued to survive, as the host-associated and reference groups appeared to be completely inactivated at temperatures past 66°C. Indeed, the surviving cell counts of the wastewater group were found to be significantly higher than the host-associated and reference groups at 66°C ($p = 1.026E-18$), 68°C ($p = 2.092E-8$), and 70°C ($p = 0.031$) – and while the cell counts for the wastewater group at 72°C were not found to differ significantly compared to the other two groups ($p = 0.225$), only strains belonging to the wastewater group could survive at that temperature.

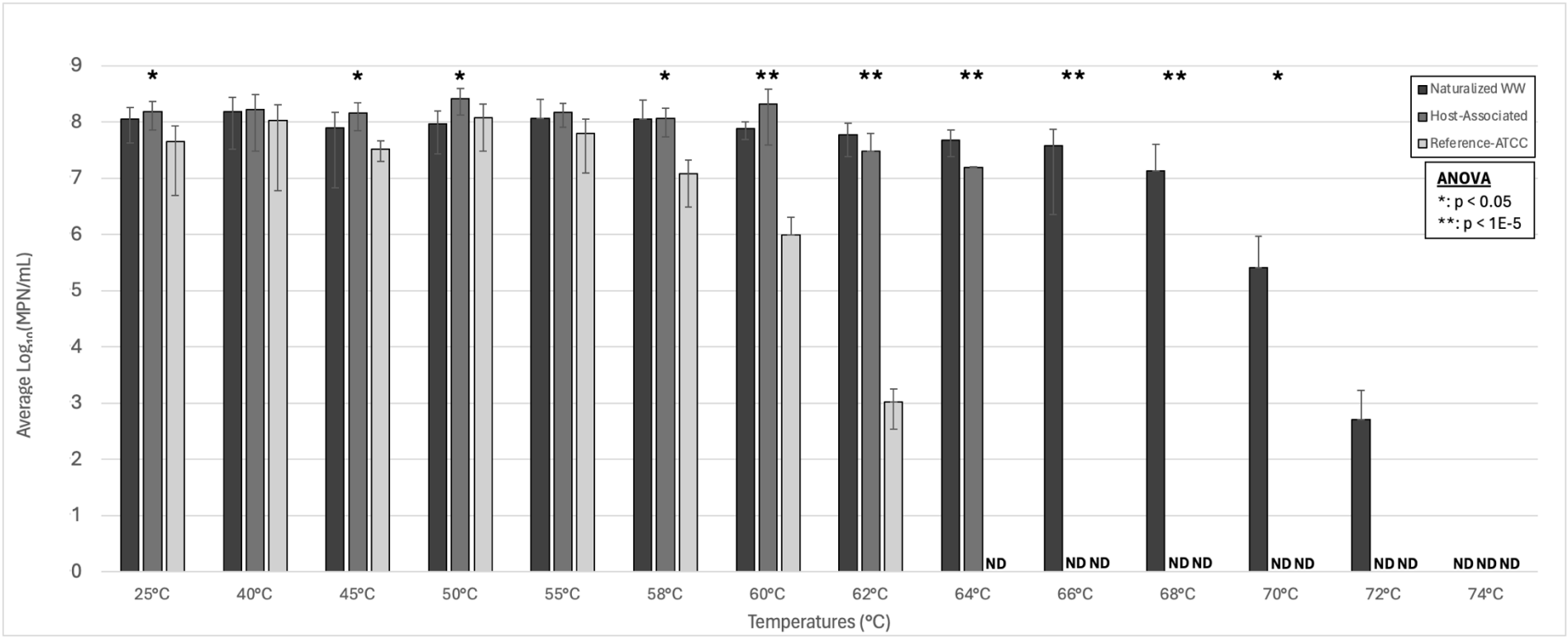



**Figure 5-5. Differences in upper thermal tolerances between the naturalized wastewater, host-associated, and reference *E. coli* ecotypes.** The surviving cell counts (in $\log_{10}$(MPN/mL)) for each strain following heat treatment (30 second exposure at each temperature) were pooled into naturalized wastewater (black), host-associated (grey), and reference (light grey) groups. The average surviving cell counts for each group were then compared at each temperature using Single Factor ANOVA tests, with statistically significant differences between the wastewater, host-associated and reference groups indicated with either a single asterix (*, $p < 0.05$) or double asterixes (**, $p > 1E\text{-}5$). ND – not detected.

### 5.3.3 Temperature-Dependent Growth Kinetics of Naturalized Wastewater, Host-Associated and Reference *E. coli* Strains in Nutrient Rich and Minimal Conditions

Despite their apparent specificity to the wastewater niche, the growth patterns exhibited by the wastewater strains were not found to necessarily favor low temperature or low nutrient conditions. In TSB, for instance, all strains exhibited typical patterns of growth (i.e., the standard sigmoidal 'S' curve) at both 25°C and 37°C conditions (Figure 5-6), although each strain appeared to initiate growth earlier, and also reached a higher peak cell density faster, when grown specifically at 37°C. Despite these overall similarities, some slight differences in the individual growth curves were observed. WW69, for instance, consistently exhibited the lowest peak cell densities, regardless of the growth temperature. Indeed, while the growth curves of most strains were found to display an average peak OD600 of ~1.8, WW69 appeared to exhibit a peak OD600 of ~1.4 at 37°C and only 0.6 at 25°C (Figure 5-6C). In contrast, WW2 was found to exhibit the highest peak cell densities during growth, with a peak OD600 ranging between ~2 to 2.2 (Figure 5-6D). Additionally, select strains were also found to exhibit a 'two-tiered' growth curve at 25°C in TSB, including WW9 (Figure 5-6B), H51 (Figure 5-6E) and CFT073 (Figure 5-6G). Importantly, while there was a slight degree of variability in the growth curves of each strain in TSB at 37°C and 25°C, these differences did not appear to be associated with ecotype.

While all strains were able to grow in TSB, regardless of the growth temperature, growth appeared to be hindered overall in minimal media. Indeed, none of the strains exhibited significant growth in MM when incubated at 25°C (Figure 5-7), with select strains, including the host-associated strain CFT073 (Figure 5-7G) and lab reference strain ATCC 25922 (Figure 5-7H), also experiencing a complete lack of growth at 37°C. Of the strains that were able to grow in MM at 37°C, growth appeared to be considerably stunted as each strain appeared to initiate growth much

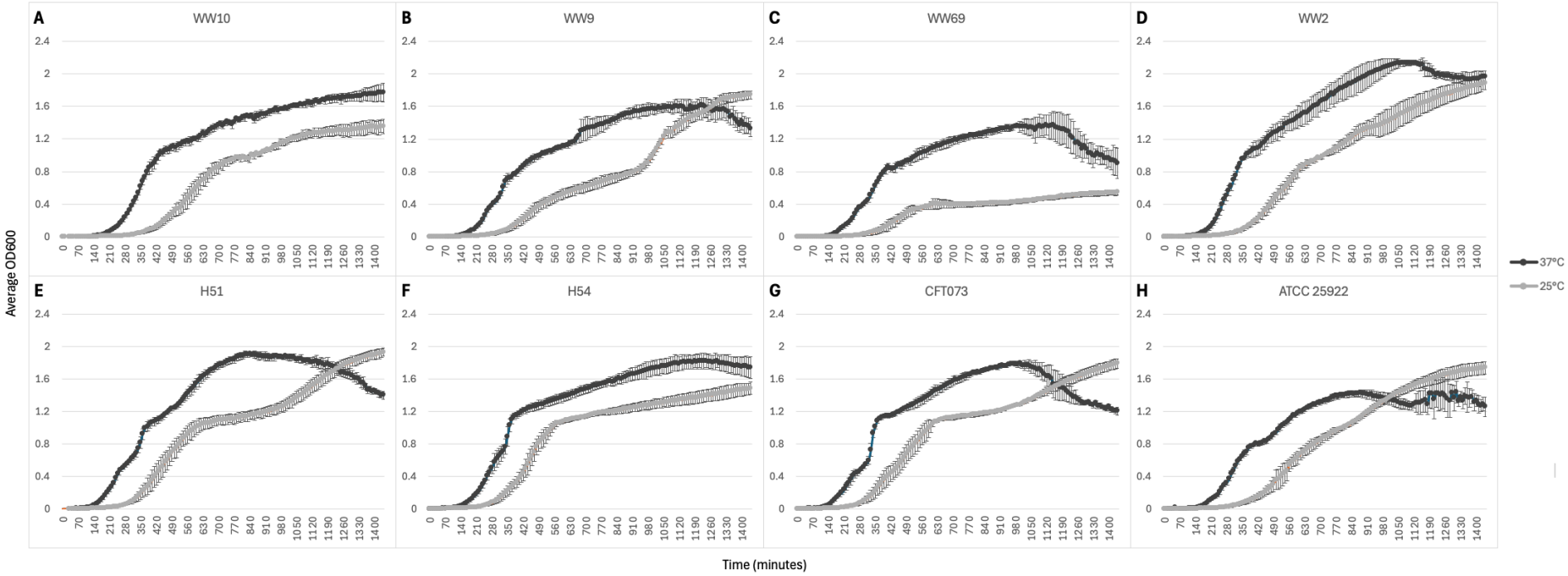



**Figure 5-6. Growth curves of naturalized wastewater, host-associated, and reference *E. coli* strains in TSB at 37°C and 25°C.** All naturalized wastewater **(A-D)**, host-associated **(E-G)**, and reference (**H)** *E. coli* strains were cultured overnight in TSB, washed twice and then re-suspended in PBS, and then diluted to achieve a starting concentration of ~$10^6$ cells/mL. Five replicates of each strain in TSB were aliquoted into a 96-well plate, and then incubated at either 37°C (black lines) or 25°C (grey lines) for 24 hours with shaking at 200 rpm in a fluorescent microplate reader (BMG Labtech, FLUOstar® Omega). Cell density measurements, in OD600 were taken every 10 minutes over the 24-hour period, and the average OD600 at each time point was calculated for each strain, with the error bars indicating the standard deviation in OD600 measurements across two independent growth assays that were performed for each strain at each temperature condition.

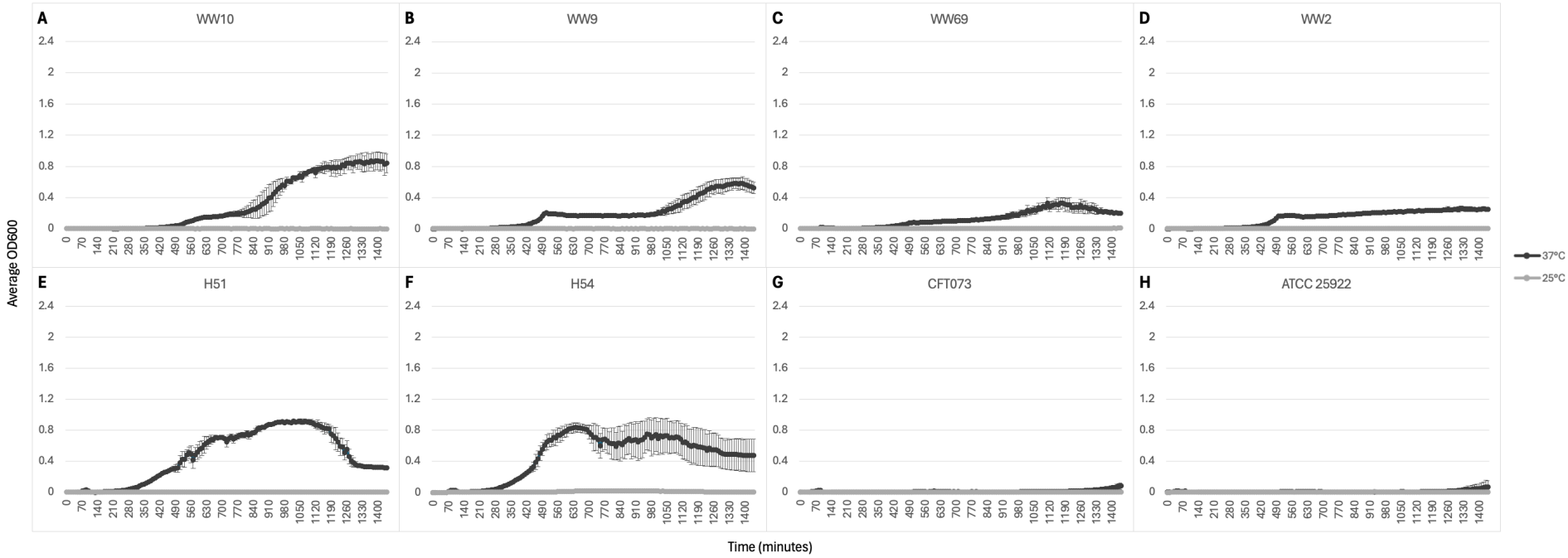



**Figure 5-7. Growth curves of naturalized wastewater, host-associated, and reference *E. coli* strains in MM at 37°C and 25°C.** All naturalized wastewater **(A-D)**, host-associated (**E-G)**, and reference **(H)** *E. coli* strains were cultured overnight in TSB, washed twice and then re-suspended in PBS, and then diluted to achieve a starting concentration of ~$10^6$ cells/mL. Five replicates of each strain in MM were aliquoted into a 96-well plate, and then incubated at either 37°C (black lines) or 25°C (grey lines) for 24 hours with shaking at 200 rpm in a fluorescent microplate reader (BMG Labtech, FLUOstar® Omega). Cell density measurements, in OD600 were taken every 10 minutes over the 24-hour period, and the average OD600 at each time point was calculated for each strain, with the error bars indicating the standard deviation in OD600 measurements across two independent growth assays that were performed for each strain at each temperature condition.

slower (i.e., at a later time point) in MM than compared to their growth patterns in TSB. Furthermore, the peak cell densities achieved during growth within MM was much lower than what was observed during growth in TSB, with OD600 values reaching a maximum of only ~0.8 for WW10 (Figure 5-7A), H51 (Figure 5-7E), and H54 (Figure 5-7F), and a minimum of ~0.2 for WW69 (Figure 5-7C) and WW2 (Figure 5-7D).

Overall, all strains appeared to grow more favorably in a nutrient-rich medium, and at temperatures reflecting the 'optimal' host environment. Despite this, when the strains were pooled according to ecotype, some general trends could be observed across the growth conditions. For instance, while all ecotypic groups generally exhibited similar growth curves in TSB at 37°C (Figure 5-8A), the host-associated strains appeared to initiate growth faster and reach a higher peak average cell density compared to the naturalized wastewater strains ($p < 0.05$) and reference strain ($p < 0.05$). Interestingly, although there were differences in the final cell densities between these latter two groups, the growth curve of the naturalized wastewater strains did not appear to differ significantly from the reference strain ($p = 0.061$). A similar trend was observed for the growth curves produced in TSB at 25°C. Again, while the general growth patterns appeared to be similar between the three ecotypes, the host-associated strains appeared to initiate growth significantly faster compared to the naturalized wastewater ($p < 0.05$) and reference ($p < 0.05$) strains, whereas the latter two groups did not appear to display significant differences ($p = 0.109$) in their growth kinetics (Figure 5-8B). In contrast, all groups were found to exhibit distinct growth patterns in MM at 37°C (Figure 5-8C), as the host-associated strains appeared to initiate growth earlier compared to the naturalized wastewater ($p << 0.05$) and reference ($p << 0.05$) strains, whereas the wastewater group was found to exhibit significantly greater growth overall (i.e., higher peak OD600) compared to the reference strain ($p < 0.05$). None of the naturalized wastewater, host-associated,

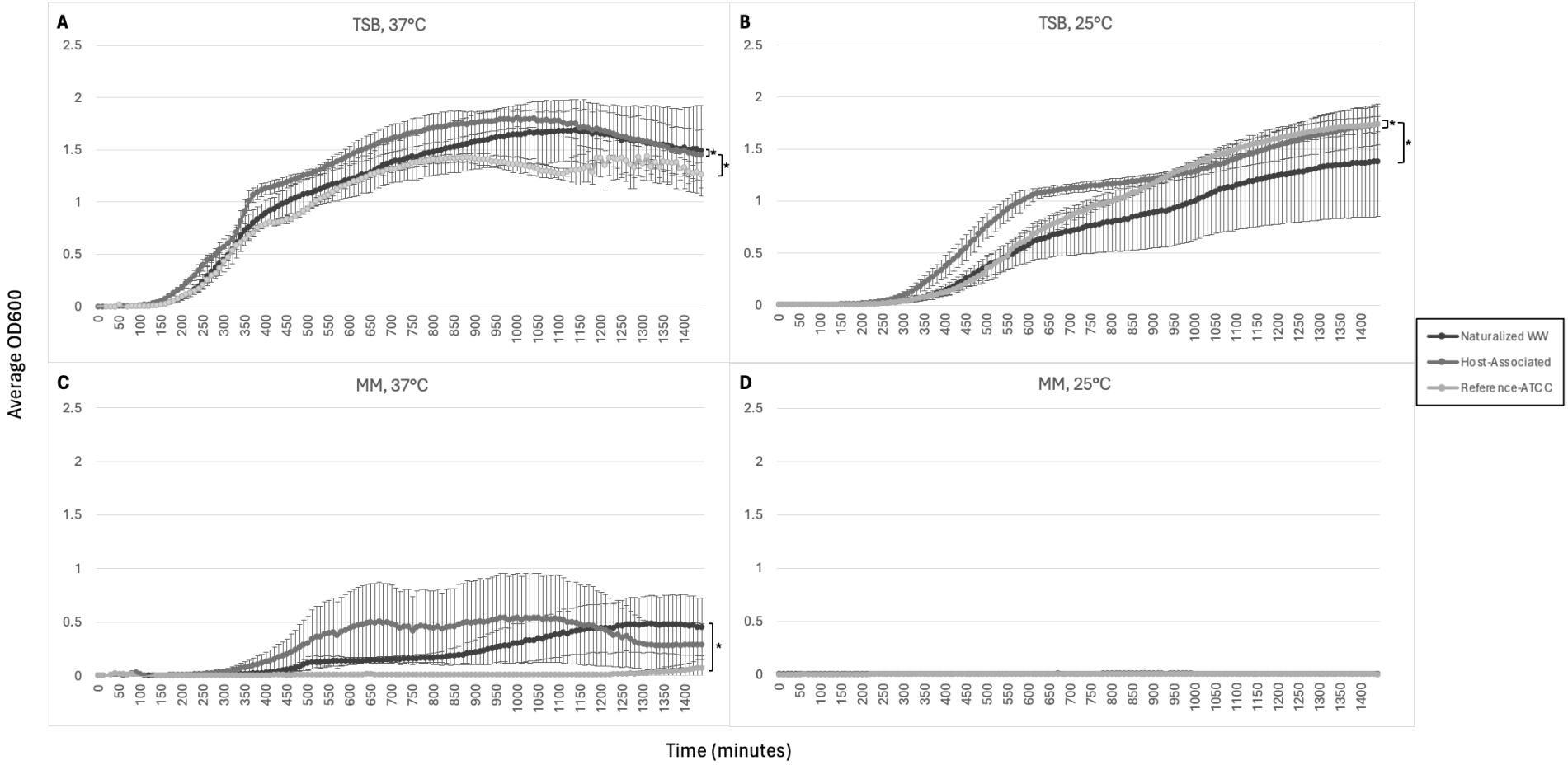



**Figure 5-8. Temperature- and medium-dependent trends in growth kinetics between the naturalized wastewater, host-associated, and reference *E. coli* ecotypes.** The individual OD600 values for each strain at each time point were pooled so that an average growth curve can each be calculated for the naturalized wastewater (black lines), host-associated (dark grey lines), and reference (light grey lines) groups in TSB and MM at 37°C and 25°C. The average growth curve for each group in TSB at 37°C **(A)**, TSB at 25°C **(B)**, MM at 37°C **(C)**, and MM at 25°C **(D)**, were compared in a pairwise fashion using the Compare Groups of Growth Curves statistical test. Any differences in growth curves between any two of the naturalized wastewater, host-associated, or reference groups that were found to be statistically significant are indicated by a single asterix (*, $p < 0.05$).

and reference groups, however, were found to exhibit any significant growth or differences in their growth patterns (all pairwise comparisons produced p-values > 0.05), in MM at 25°C (Figure 5-8D).

### 5.3.4 Temperature-Dependent Biofilm Formation of Naturalized Wastewater, Host-Associated and Reference *E. coli* Strains in Nutrient Rich and Minimal Conditions

Although the naturalized wastewater strains did not appear to grow more favorably in low-nutrient, lower-temperature conditions, their persistence within the wastewater environment could alternatively be facilitated through the production of biofilms. Interestingly, the biofilms that were produced by the naturalized wastewater, host-associated, and reference strains appeared to vary widely depending on the specific medium and temperature that the strains were incubated in (Figure 5-9). Aside from the host-associated strain H54, which did not appear to form biofilms at any significant capacity regardless of the incubation conditions (Figure 5-9F), all strains appeared to produce more robust biofilms at 25°C than at 37°C. There were a select number of strains, including WW9 (Figure 5-9B), WW69 (Figure 5-9C), and WW2 (Figure 5-9D) that were still able to show considerable biofilm formation capacity at 37°C (i.e., with the eluted biofilms producing an average absorbance at 590nm of ~0.6 to ~0.7); however, most other strains predominantly produced biofilms only after incubation at 25°C. The specific medium that favored biofilm formation at 25°C varied according to each strain. The host-associated strains H51 (Figure 5-9E) and CFT073 (Figure **5-9** 5-9G) alongside the reference strain ATCC 25922 (Figure 5-9H) appeared to form biofilms best in TSB. In contrast, the wastewater strains were generally found to produce better biofilms in MM, with an exception being WW2 (Figure 5-9D) as it was found to produce more robust biofilms in TSB similar to the host-associated and reference strains.

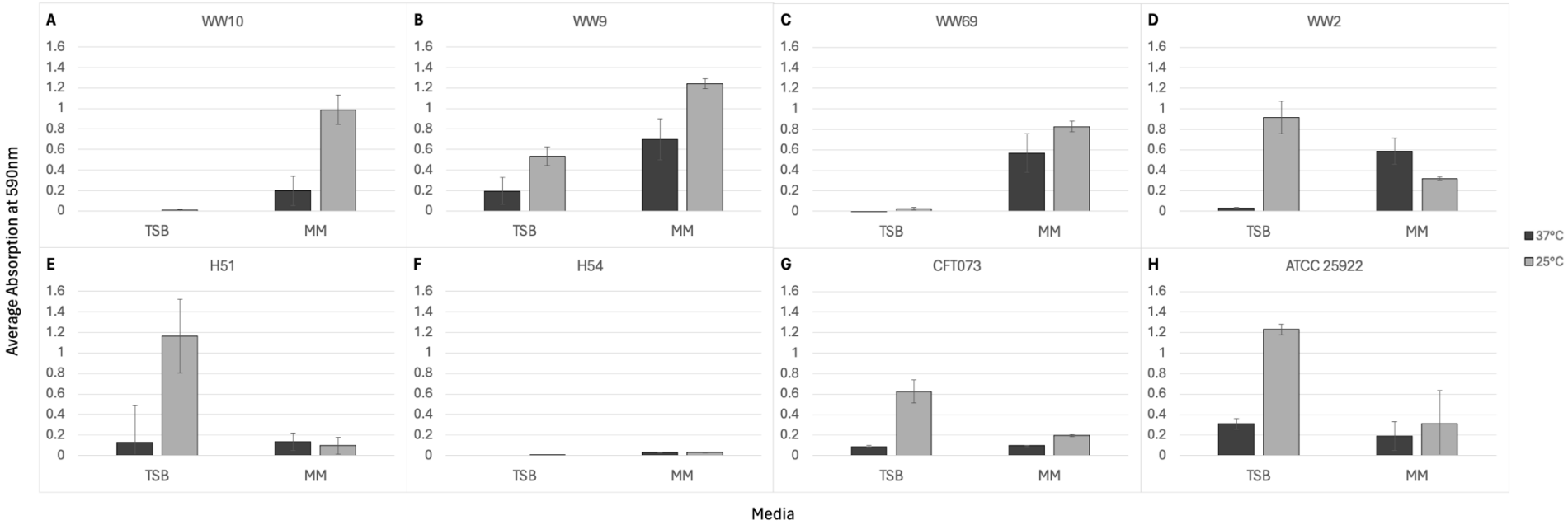



**Figure 5-9. Temperature- and media-dependent biofilm production in naturalized wastewater, host-associated, and reference *E. coli* strains.** All naturalized wastewater **(A-D)**, host-associated **(E-G)**, and reference **(H)** *E. coli* strains were cultured overnight in TSB, washed twice and then re-suspended in PBS, and then diluted to achieve a starting concentration of ~$10^7$ cells/mL. Three replicates of each strain were prepared in either TSB or MM in a 96-well plate, following which a 96-well pegged-lid (DOJINDO Laboratories) was submerged in each suspension and incubated for 24 hours at either 37°C (black bars) or 25°C (grey bars). Following incubation, the pegs were soaked in PBS twice, stained with crystal violet, soaked in PBS twice again, and then incubated in 100% ethanol to elute the dye that was captured by any biofilms formed on the pegs. The absorbance of the resulting stained solutions at 590nm were then measured using a fluorescent microplate reader (BMG Labtech, FLUOstar® Omega), with the error bars indicating the standard deviation in absorption measurements across two independent biofilm formation assays that were performed for each strain at each incubation temperature and medium.

Although each strain appeared to vary in their capacity to form biofilms, some general trends were observed when the strains were pooled according to ecotype. Generally, the reference strain appeared exhibit the greatest biofilm production in TSB compared to the host-associated and naturalized wastewater groups (Figure 5-10). Indeed, although biofilm formation was generally weaker in TSB at 37°C, the reference strain was still found to produce significantly more robust biofilms ($p = 1.124E-7$), at roughly 5 times the capacity, compared to their host-associated and naturalized wastewater counterparts. Even when all groups exhibited improved biofilm formation capacity in TSB at 25°C, the reference strain was again found to exhibit the greatest biofilm production ($p = 3.477E-4$), at ~4 times and ~2 times the capacity of the naturalized wastewater and host-associated groups, respectively. In minimal media, however, the naturalized wastewater strains appeared to be the best biofilm producers. At 37°C, the wastewater group was found to produce biofilms at roughly ~6 times the capacity of the host-associated group, and ~2.5 times the capacity of the reference strain ($p = 8.280E-9$). These differences appeared to be even more pronounced at 25°C, with an ~8-fold and ~3-fold difference in biofilm formation between the wastewater group their host-associated and reference counterparts, respectively ($p = 2.518E-10$).

## 5.4 Discussion

As described in the preceding chapters of this thesis, naturalized WWS-*E. coli* strains appear to be characterized by various ecotypic and genotypic characteristics that reflect their ability to exploit wastewater as a primary niche. Supporting these ecotype-informative features, and reflecting the polyphasic approach, the WWS strains were found to exhibit several phenotypic traits that seem to underlie their adaptation to the wastewater environment. For instance, in Chapter Four of this thesis, the wastewater strains were found to harbor an abundance of stress resistance

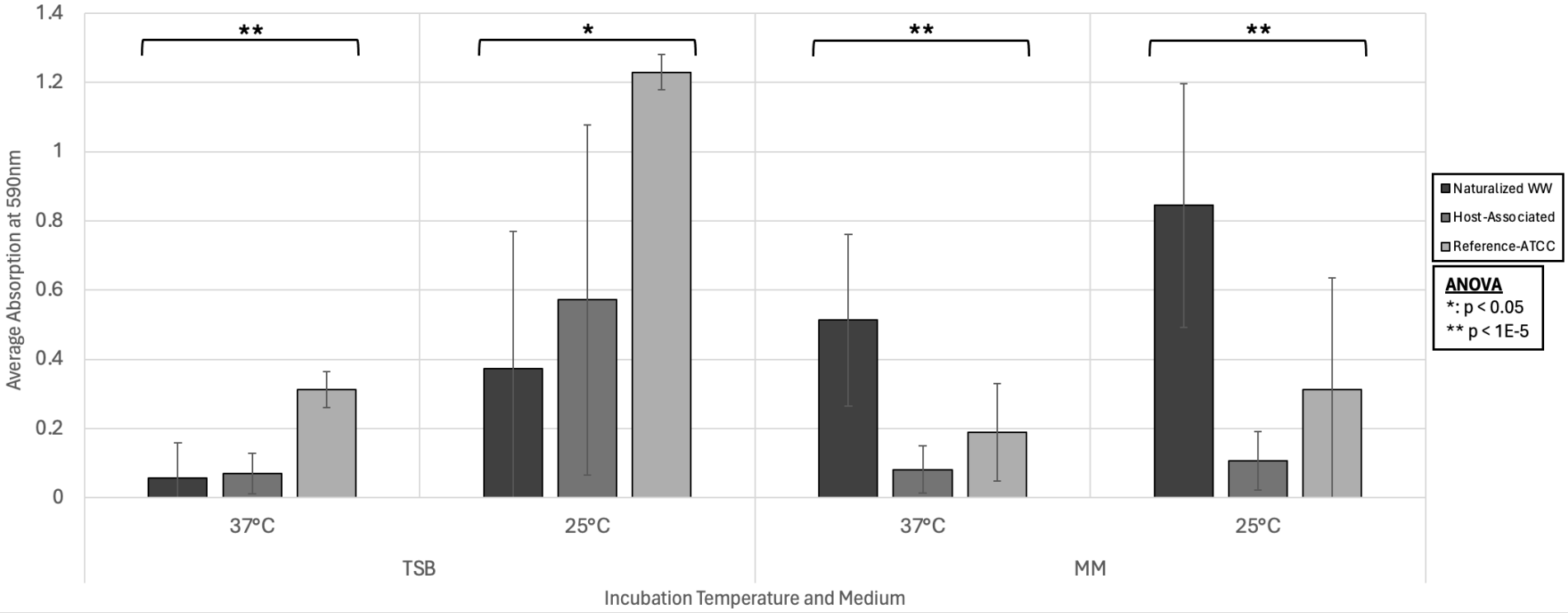



**Figure 5-10. Temperature- and medium-dependent trends in biofilm production between the naturalized wastewater, host-associated, and reference *E. coli* ecotypes.** The individual absorption measurements at 590nm for each strain for each combination of incubation temperatures and media were pooled so that an average absorption value could be calculated for the naturalized wastewater (black bars), host-associated (dark grey bars), and reference (light grey bars) groups. The average absorption at 590nm, correlating with the amount of biofilm produced, for each group in both TSB and MM, and at both 37°C and 25°C, were compared using Single Factor ANOVA tests, any with statistically significant differences between the wastewater, host-associated and reference groups indicated with either a single asterix (*, $p < 0.05$) or double asterixes (**, $p > 1E-5$).

genes that could enhance their survival within wastewater, including against disinfection-related stressors such as DNA-damaging stimuli (i.e., UV radiation), oxidative stress (i.e., advanced oxidants, chlorine), and heat (i.e., sludge digestion, biosolids composting). Importantly, previous studies have recapitulated some of these genotypic findings. For instance, compared to *E. coli* isolates derived from food samples, wastewater strains were found to exhibit enhanced resistance to disinfecting agents including advanced oxidants and chlorine (Wang *et al.* 2020). In the same study, these wastewater strains were also characterized by an 'extreme' resistance phenotype when exposed to 60°C for 5 minutes. The work presented in this chapter provide additional support for these findings, as the WWS-*E. coli* were the only ecotypic group that appeared to survive heat treatment for 5 minutes at temperatures of 60°C and higher. Interestingly, although there was some slight variability in their individual survivability at each temperature, the wastewater strains were found to be even more heat resistant than previously determined (Wang *et al.* 2020). Indeed, all wastewater strains were all able to withstand exposure to 62°C for 5 minutes, while all wastewater strains aside from WW2 were also able to tolerate 64°C without complete inactivation.

The wastewater strains were also found to exhibit higher upper thermal tolerances compared to the host-associated and reference strains. While all non-naturalized strains were completely inactivated during short-term exposure at 62-64°C, the wastewater strains remained culturable following 30 second treatment at 68°C, with select wastewater strains (i.e., WW10 and WW9) surviving temperatures up to 72°C (i.e., food pasteurization). Although this phenotypic trait of enhanced heat resistance may not be directly relevant for survival within the lower ambient temperatures of the wastewater niche, the ability to tolerate elevated temperatures could be important for those naturalized strains that become diverted to biosolids composting during wastewater treatment. Considering that processes such as sludge digestion and composting can

subject biosolids to temperatures ranging from 30–80°C (Gerba and Pepper 2009), the ability to withstand elevated temperatures could be advantageous for their survival during biosolids treatment. It is important to note that, while the naturalized strains could withstand temperatures of up to 64°C for 5 minutes and 72°C for 30 seconds, the duration and modes of exposure used to assess heat resistance in this study may not necessarily reflect the treatment processes used for sludge digestion and composting. Despite this, the enhanced heat resistance exhibited by the naturalized wastewater strains suggests that they may be better equipped to tolerate and survive biosolids treatment compared to their host-associated counterparts.

Although the niche-specificity of the wastewater strains could imply an enhanced ability to potentially grow in the wastewater matrix, the growth patterns of the WWS-*E. coli* strains did not appear to be favored within conditions mimicking a non-host environment. All strains regardless of ecotype were able to grow across different temperatures (25°C and 37°C) and media, aside from the low-nutrient, low-temperature (i.e., MM at 25°C) condition. Interestingly, similar findings have been reported for other non-host, environmental *E. coli* populations. Environmentally-derived *E. coli* strains have been shown to be capable of growing within environmental matrices such as soil (Ishii *et al.* 2006); however, their growth appeared to stagnate once temperatures dropped below the typical 'optimal' growth temperature ranges of 30–37°C (Ishii *et al.* 2010; Petersen and Hubbart 2020; Topp *et al.* 2003). While these findings suggest that the persistence of the WWS-*E. coli* strains within wastewater may not be related to growth, it is important to note that minimal media may not reflect the nutritional content of the wastewater matrix. Indeed, although minimal media was chosen as a proxy for the lower nutrient conditions typically associated with non-host environments, the composition of wastewater (i.e., 'waste' derived from various anthropogenic sources) could provide sufficient nutrition to support

microbial growth. Considering that other studies have been able to demonstrate the growth of environmental *E. coli* isolates at temperatures as low as 15°C (Brennan *et al.* 2010; Brennan *et al.* 2013), the possibility remains that the naturalized wastewater strains might be able to grow within the wastewater environment.

Regardless of the specific growth conditions, the WWS strains were generally found to initiate growth slower compared to the host-associated and reference strains, and also grew to lower peak cell densities overall. Interestingly, 'slow growing' phenotypes have been previously reported to be an adaptive response in *E. coli* to stressful conditions, including exposure to toxic atmospheric pollutants (Zhang *et al.* 2019c), as well as during nutrient starvation (Biselli *et al.* 2020; Nandy 2022). As such, the 'slower' growth kinetics of the WWS strains could reflect their adaptation to a niche that is characterized by high concentrations of toxic contaminants (i.e., antibiotics, pharmaceuticals, detergents, heavy metals, etc.) and low concentrations of available nutrients (i.e., as organics/biosolids are removed from the wastewater matrix during treatment).

While our experiments were unable to demonstrate growth kinetics that strongly reflected a naturalized ecotype, these same environmental conditions (i.e., lower ambient temperature and nutritional availability) appeared to promote biofilm production specifically within the naturalized wastewater strains. Previous work has already characterized the ability of the naturalized strains to produce robust biofilms at 37°C, especially when compared to their enteric counterparts (Zhi *et al.* 2016a); however, the findings presented in this chapter indicate that the naturalized strains appear to form biofilms best at lower temperatures under low-nutrient conditions. Indeed, while biofilm production generally appears to be favored at 25°C as opposed to 37°C (Mathlouthi *et al.* 2018), the naturalized strains were generally better biofilm producers in minimal media whereas the host-associated and reference strains produced more robust biofilms in TSB. Importantly, an

enhanced ability to form biofilms, specifically within nutrient-limiting conditions, could facilitate the survival and persistence of the naturalized strains within the wastewater niche. Considering the roles of biofilms in prolonging environmental persistence on abiotic surfaces (Abdallah *et al.* 2014), and enhancing resistance to various environmental, wastewater treatment-related stressors (i.e., microbial predation, exposure to antibiotics and chemical contaminants, UV radiation, sanitizing disinfectants [McDougald *et al.* 2012; Yang *et al.* 2018]), the production of biofilms under nutrient-depleted conditions could represent a particularly important survival mechanism acquired by the WWS strains within the wastewater environment.

Despite their designation as a single, cohesive naturalized wastewater *E. coli* ecotype, a considerable degree of variability in heat resistance profiles, growth kinetics, and biofilm formation capacities were observed across the naturalized wastewater strains. These phenotypic discrepancies raise an important concept – that ecotypes are not necessarily uniform or clonal, as strains belonging to the same ecotypic group can still exhibit variation in their genotypic and phenotypic properties. Reflecting this, the wastewater strains assessed in this chapter varied both genotypically (i.e., presence of *uspC*–IS30–*flhDC* biomarker and tLST) and phenotypically. Despite this, the naturalized wastewater strains were collectively found to be more heat resistant and better biofilm producers, though potentially slower growers, than their host-associated counterparts. Coupled with previous work assessing disinfection-related resistance in these strains (Wang *et al.* 2020; Zhi *et al.* 2017), this chapter provides additional supporting evidence that the naturalized wastewater strains possess various phenotypic adaptations that may have allowed them to exploit wastewater as a niche – thus reinforcing their distinct ecotypic and genotypic characteristics through a polyphasic lens.

Coupled with the preceding chapters of this thesis, as well as previous work from others (Wang et al., 2020; Zhi et al., 2016a, Zhi *et al.* 2017; Zhi *et al.* 2019), our findings collectively suggest that certain *E. coli* strains have adapted themselves to live and survive within environments that have been specifically engineered to eliminate them (i.e., sewage treatment plants). Interestingly, this phenomenon appears to have occurred worldwide, with the WWS-*E. coli* strains being globally distributed across sewage treatment plants in Canada, the United States, the U.K., Switzerland, and China. The apparent evolution of treatment resistance in these strains, however, is concerning, as sewage treatment and sanitation play a fundamental role in the public health control of infectious diseases in modern society. Indeed, while the naturalized strains do not appear to be host-associating, their characterization raises an important question: could pathogenic microbes be similarly developing resistance to wastewater treatment? Indeed, the selective pressures exerted by wastewater treatment could similarly drive the evolution of other *E. coli* ecotypes towards treatment resistance – a prospect that the next chapter in this thesis will explore in greater detail.

# Chapter Six: The Emergence of Wastewater Treatment Resistance in ExPEC: The Differential Survival of Potentially Septicemic and Meningitic *E. coli* Across the Wastewater Treatment Train[5]

## 6.1 Introduction

The polyphasic characterization of the naturalized, WWS-*E. coli* presented in the preceding chapters of this thesis raises the prospect that microbes may be evolving resistance to wastewater treatment. While the naturalized wastewater strains appear to harbor an abundance of genotypic and phenotypic adaptations that collectively underpin their ability to tolerate the wastewater treatment process, the *E. coli* species as a whole has already been documented to harbor a diverse array of stress response systems. This includes, but is not limited to, the universal stress response (Nachin *et al.* 2005), SOS response (Maslowska *et al.* 2019), oxidant tolerance (ROS) response (Trastoy *et al.* 2018), and general stress (*rpoS*-mediated) response (Landini *et al.* 2014), which all appear to confer resistance to the myriad of environmental stressors that are encountered during wastewater treatment (Gray et al., 2013; Trastoy *et al.* 2018). Additionally, various strains may also employ specific strategies to respond to wastewater treatment-related challenges, such as through encoding several chlorine- (Gray *et al.* 2013), oxidative stress-, and UV-specific (Imlay 2008; Trastoy *et al.* 2018) transcription factors, as well as the production of cross-protective chaperone proteins (Winter *et al.* 2008). As such, beyond the WWS-*E. coli*, other populations

[5] A version of this chapter has been published as: Yu, D., Ryu, K., Otto, S. J. G., Stothard, P., Banting, G., Ruecker, N., Neumann, N. F., and Zhi, S. 2022. Differential survival of potentially pathogenic, septicemia- and meningitis-causing E. coli across the wastewater treatment train. npj Clean Water. 5(1): 1–12. doi:10.1038/s41545-022-00177-y.

within the *E. coli* species may already be genetically equipped to survive wastewater treatment.

Interestingly, genomic analyses of the WWS-*E. coli* strains may indicate alternative *E. coli* ecotypes that are also capable of surviving wastewater treatment. Although the WWS-*E. coli* were characterized by a greatly reduced virulence gene repertoire in Chapter Four of this thesis, previous work by Zhi *et al.* (2019) found that these strains still possessed several UPEC-associated virulence genes, including those involved in iron acquisition (i.e., ferrienterobactin and yersiniabactin biosynthesis and transport), fimbrial adhesion (i.e,, type 1 fimbriae, curli fimbriae), flagellar biosynthesis, and outer membrane transport. The abundance of UPEC virulence genes within the WWS-*E. coli* implies their dual role in mediating UPEC pathogenesis while also enhancing survival within non-host environmental niches like wastewater. The genomic background of UPEC strains, therefore, may be particularly adaptive within wastewater, especially for surviving wastewater treatment.

Reflecting this, evidence suggests that the UPEC pathotype seems particularly adapted to survive the wastewater treatment process. Based on virulence gene screening, Anastasi *et al.* (2010; 2013) found that up to 59.5% of the surviving *E. coli* population following chlorination and UV disinfection represented potential UPEC. This was later corroborated by Adefisoye and Okoh (2016), as they determined that 41.7% of *E. coli* strains within wastewater effluents carried UPEC virulence genes. Similarly, Calhau *et al.* (2015) isolated a potential UPEC strain harboring various UPEC-associated virulence genes and pathogenicity islands that belonged to the major pandemic ExPEC-associated sequence type lineage, ST131. More recently, whole genomic analyses carried out by Zhi *et al.* (2020) demonstrated that a significant proportion of *E. coli* surviving wastewater treatment shared an extreme degree of genomic and phylogenetic similarity with clinically-confirmed UPEC isolates. Interestingly, several of these presumptive wastewater-borne UPEC

strains were found to belong to the emerging community-based, pandemic-associated O25b-ST131 clonal group (Rogers *et al.* 2011) and were characterized as extended spectrum beta-lactamase (ESBL) producing strains. Thus, select *E. coli* strains surviving wastewater treatment may possess the capacity to cause urinary tract infections.

While previous studies primarily focused on UPEC, the same observations appear to apply to the other ExPEC pathotypes, including the septicemic bloodborne *E. coli* (BBEC) and neonatal meningitic *E. coli* (NMEC). For instance, *E. coli* strains isolated from treated wastewater have been shown to cluster within sequence types such as ST95 and ST131 (Calhau *et al.* 2015; Raven *et al.* 2019; Zhi *et al.* 2020), lineages that are typically associated with NMEC (Riley 2014) and BBEC (Riley 2014; Wu *et al.* 2014) outbreaks. Furthermore, while Adefisoye and Okoh (2016) characterized most of their *E. coli* isolates from wastewater effluents as potential UPEC, 14.8% were identified as potential NMEC. Thus, rather than just UPEC alone, all ExPEC pathotypes appear to be capable of differentially surviving wastewater treatment. To interrogate this further, we demonstrate using a comprehensive comparative genomics approach that NMEC and BBEC strains also appear to represent common constituents in full-scale treated wastewater effluents and chlorinated sewage, suggesting these pathotypes are similarly selected for during wastewater treatment. Importantly, our findings contribute to a growing body of evidence that raises a concerning public health prospect – that, beyond naturalized populations, potentially pathogenic *E. coli* strains capable of causing extraintestinal diseases may also be evolving resistance to wastewater treatment.

## 6.2 Materials and Methods

### 6.2.1 Screening of ExPEC-Related Virulence Genes and Molecular Markers

A total of 376 chlorine-tolerant and 261 wastewater treatment-resistant *E. coli* isolates, previous collected by Zhi *et al.* (2020), were assessed in this analysis. All *E. coli* isolates were first grown in TSB overnight at 37°C, following which genomic DNA (gDNA) extractions were performed on the bacterial cultures using DNeasy Blood & Tissue kits (QIAGEN, Toronto, Canada) according to the manufacturer's instructions. These isolates were then screened for the *uspC-IS30-flhDC* locus via PCR to eliminate naturalized WWS-*E. coli* isolates from the library. The remaining isolates were subsequently screened for the *uidA* gene for further species confirmation, following which all confirmed *E. coli* isolates were then screened against a panel of ExPEC-associated virulence genes and molecular markers (Table 6-1). The PCR reactions for the *uidA* and *uspC*–IS30–*flhDC* markers were carried out according to protocols previously described by Taskin *et al.* (2011) and Zhi *et al*. (2016), respectively. For the other molecular markers, the reaction mixtures consisted of 20-40 ng of gDNA template, 12.5 µL of 1X GoTaq Hotstart Mastermix (Promega), and 500 nM of each primer. Cycling conditions varied based on the specific molecular marker. For *papC*, *sfa-foc, iroN* and *ibeA*, cycling conditions were as follows: 95°C for 2 min, followed by 33 cycles of 30 s at 95°C, 30 s at 63°C and 45 s at 72°C, and a final extension for 7 min at 72°C. For *fyuA* and *chuA*: 95°C for 2 min, followed by 33 cycles of 30 s at 95°C, 30 s at 63°C and 1 min at 72°C, and a final extension for 7 min at 72°C. For *kpsM*: 95°C for 2 min, followed by 35 cycles of 20 s at 95°C, 20 s at 62°C and 45 s at 72°C, and a final extension for 7 min at 72°C. For the ST131 marker: 95°C for 2 min, followed by 35 cycles of 20 s at 95°C, 20 s at 57°C and 45 s at 72°C, and a final extension for 7 min at 72°C. For the O25b– ST131 marker: 95°C for 4 min, followed by 30 cycles of 5 s at 94°C and 10 s at 65°C, and a final extension for 5 min at 72°C. All PCR reactions were performed on an ABI 2720 thermocycler (Applied

**Table 6-1.** Overview of target genes and genetic markers included in the ExPEC screening PCR panel

| ExPEC Gene/Marker | Description | Forward Primer (5' – 3') | Reverse Primer (5' – 3') | Amplicon Size (bp) | Anneal Temp (°C) | Reference |
|---|---|---|---|---|---|---|
| ***papC*** | ExPEC virulence gene for an outer membrane usher protein for P fimbriae | GTGGCAGTATGAGTA ATGACCGTTA | ATATCCTTTCTGCAGG GATGCAATA | 202 | 63 | White *et al.* 2011 |
| ***sfa-foc*** | ExPEC virulence gene encoding S/F1C fimbriae | CTCCGGAGAACTGGG TGCATCTTAC | CGGAGGAGTAATTAC AAACCTGGCA | 407 | 63 | White *et al.* 2011 |
| ***chuA*** | ExPEC virulence gene encoding an outer membrane heme receptor | CTGAAACCATGACCGT TACG | TTGTAGTAACGCACTA AACC | 652 | 63 | Spurbeck *et al.* 2012 |
| ***iroN*** | ExPEC virulence gene encoding a siderophore receptor | AAGTCAAAGCAGGGG TTGCCCG | GACGCCGACATTAAG ACGCAG | 665 | 63 | White *et al.* 2011 |
| ***fyuA*** | ExPEC virulence gene encoding an outer membrane iron receptor | GTAAACAATCTTCCCG CTCGGCAT | TGACGATTAACGAAC CGGAAGGGA | 850 | 63 | Spurbeck *et al.* 2012 |
| ***kpsM*** | ExPEC virulence gene encoding the K1 capsule antigen; commonly associated with NMEC strains | CCATCGATACGATCAT TGCACG | ATTGCAAGGTAGTTCA GACTCA | 400 | 60 | Takahashi *et al.* 2006 |
| ***ibeA*** | ExPEC virulence gene encoding an invasin gene required for crossing the blood-brain barrier in NMEC | AGGCAGGTGTGCGCC GCGTAC | TGGTGCTCCGGCAAAC CATGC | 170 | 63 | White *et al.* 2011 |
| ***ST131*** | Genetic marker targeting the predominant ExPEC-associated pandemic lineage | AGCAACGATATTTGCC CATT | GGCGATAACAGTACG CCATT | 580 | 57 | Matsumura *et al.* 2017 |
| ***O25b-ST131*** | Genetic marker targeting a dominant clone of the pandemic ST131 lineage | TCCAGCAGGTGCTGG ATCGT | GCGAAATTTTTCGCCG TACTGT | 347 | 65 | Clermont *et al.* 2009 |
| ***uspC–IS30–flhDC*** | Gene marker associated with naturalized wastewater *E. coli* | CGGGGAACAAATGAG AACAC | TGGAGAAACGACGCA ATC | 386 | 60 | Zhi *et al.* 2016a |

Biosystems) and the resultant products were run on 1.5% agarose gels and photographed on an ImageQuant LAS 4000 (GE Healthcare Life Sciences). PCR screening results for the ExPEC-associated virulence genes were then confirmed following whole genome sequencing for each wastewater ExPEC isolate (see below).

### 6.2.2 Whole Genome Sequencing and Assembly of Presumptive Wastewater ExPEC Strain Genomes

Chlorine-tolerant and wastewater treatment-resistant *E. coli* isolates harboring at least 3 ExPEC virulence genes, as well as any isolates positive for the NMEC-associated *ibeA* gene or ExPEC-associated ST131 genetic marker (regardless of the presence of other virulence genes) were designated as presumptive wastewater ExPEC (W-ExPEC) and selected for whole genome sequencing. Genomic DNA from the presumptive W-ExPEC isolates (n = 86) was sent to Genome Quebec (Montreal, Canada) for sequencing using an Illumina HiSeq X platform (Illumina) with paired-end 150 nucleotide reads. Trimmomatic Version 0.39 (Bolger *et al.* 2014) was used to trim the low-quality reads with the following parameters: SLIDINGWINDOW = 4:15, LEADING = 3, TRAILING = 3, MINLEN = 36. *De novo* genomic assembly was then performed using SPAdes Version 3.11.1 (Bankevich *et al.* 2012) with the '—careful' and '-k 21,33,55,77' options. Any SPAdes-assembled contigs shorter than 1000 bp were excluded from downstream analyses.

### 6.2.3 Core Genome SNP Analysis of W-ExPEC and C-ExPEC Isolates

To evaluate the pathogenic capacity of W-ExPEC isolates, their whole genome sequences were compared against a genomic library of clinical ExPEC (C-ExPEC) strains isolated from cases of bloodborne bacteremia and meningitis. The genome sequences of 320 clinical NMEC and

BBEC isolates were downloaded from the NCBI Assembly database to construct a local repository of clinical ExPECs (C-ExPECs) (Supplementary Table 6-S1). Core genome SNP comparisons were performed for each of the W-ExPEC (n = 86) genomes against the local repository of 320 C-ExPEC strains in a pairwise manner using REALPHY v1.13 (Bertels *et al.* 2014). Briefly, the whole genome assemblies of all W-ExPEC and C-ExPEC isolates (n = 406) were collected into a local folder to be used as the input for REALPHY. One of the *E. coli* assemblies in the input REALPHY folder was randomly selected to be the reference sequence, against which the genome sequences of all other isolates were mapped to produce a core genome alignment for SNP assessment. MEGA-X (Kumar *et al.* 2018) was then used to enumerate the SNPs that differed between all W-ExPEC and C-ExPEC strains.

### 6.2.4 Pairwise Whole-Genome Comparisons of W-ExPEC and C-ExPEC Isolates

W-ExPEC strains that were identified as sharing a high core genome similarity (i.e., differing by fewer than 250 SNPs in a ~417 kb core genome backbone) with at least one clinical BBEC or NMEC counterpart were retained for further comparative genetic analyses (n = 37). Specifically, whole genome approaches were used to evaluate the overall genetic similarity between W-ExPEC strains and their closest clinical counterpart. First, pairwise whole genome comparisons were performed between the 37 remaining W-ExPEC strains and the local repository of 320 C-ExPEC strains using REALPHY v1.13 (Bertels *et al.* 2014) with default parameters. The whole genome similarity for each pairwise comparison was estimated by REALPHY based on the percentage of each C-ExPEC genome that mapped onto each W-ExPEC genome as a reference. Based on previous work, a lower limit of 96.00% for whole genome similarity was set as a lower limit threshold for identifying close genetic matches between W-ExPEC and C-ExPEC strains (Zhi

*et al.* 2020). This lower limit threshold was calculated based on the upper median whole genome similarity shared between STEC O157:H7 *E. coli* isolates, and served as an indication that two strains may share similar pathogenic capabilities.

Additional pairwise whole genome comparisons were also performed for each identified W-ExPEC (n = 37) against a local repository of 46 representative intestinal pathogenic and naturalized wastewater *E. coli* strains downloaded from NCBI (Supplementary Table 6-S1). The W-ExPEC strains were compared to intestinal pathogenic strains (i.e., enterohaemorrhagic *E. coli* [EHEC], enteropathogenic *E. coli* [EPEC], enterotoxigenic *E. coli* [ETEC], etc.) to rule out the possibility that any significant similarities observed between the wastewater and clinical ExPEC strains could be explained by a shared propensity for pathogenicity in general rather than extraintestinal pathogenicity specifically, especially since certain intestinal pathotypes such as ETEC have been recovered from wastewater matrices following primary and secondary treatment (Omar and Bernard 2010). Naturalized wastewater strains were also included to rule out any significant genomic similarity that might be driven through sharing a similar ecological niche (i.e., sewage/wastewater).

### 6.2.5 Core Genome Phylogenetics, Phylogrouping and Multilocus Sequence Typing of W-ExPEC and C-ExPEC Isolates

W-ExPEC strains that shared ≥ 96.00% whole genome similarity with at least one C-ExPEC strain were retained for further analysis. A maximum likelihood phylogenetic tree was generated using RAxML 8.2.12 (Stamatakis 2014) based on the core genome alignments produced by REALPHY, and included all W-ExPEC strains (n = 37) and their closest clinical counterparts (n = 38), as well as reference *E. coli* strains of known phylogroups (Tenaillon *et al.* 2010; Touchon

*et al.* 2009; Zhang and Lin 2012). The phylogroups of the W-ExPEC and C-ExPEC strains were predicted with the ClermonTyping method, using the ClermonTyper v23.06 (Beghain *et al.* 2018) webserver. Multilocus sequence typing (MLST) was also performed on all presumptive W-ExPEC and C-ExPEC strains using mlst v2.22.0 (https://github.com/tseemann/mlst) with the *Escherichia coli* #1 scheme and then cross-referenced against the PubMLST database with the Achtman scheme (Jolley *et al.* 2018). All information pertaining to the bacterial strains included in the phylogenetic analysis can be found in Supplementary Table 6-S1. The phylogenetic tree was then visualized and annotated using the R packages ggplot2 v3.4.2 (Wickham 2016), ape v5.4.1 (Paradis and Schliep, 2019) and ggtree v2.2.4 (Yu et al., 2017).

### 6.2.7 Accessory Genome Clustering and Virulence and Antibiotic Resistance Genes Screening of W-ExPEC and C-ExPEC Isolates

A pan-genome for all presumptive W-ExPEC strains (n=86), their closest clinical NMEC and BBEC counterparts, reference UPEC strains, naturalized wastewater strains, and laboratory reference strains was estimated using Roary v3.13.0 (Page et al., 2015). Reference UPEC strains and naturalized wastewater strains were included in the analysis as previous studies have demonstrated that these strains can be isolated from chlorinated sewage and finished wastewater effluents (Zhi *et al.* 2016; Zhi *et al.* 2019; Zhi *et al.* 2020). Laboratory reference *E. coli* strains (i.e., *E. coli* K12 MG1655, *E. coli* K12 W3110, *E. coli* K12 BW2592, and *E. coli* ATCC 25922) were included in the analysis to represent a non-pathogenic reference subgroup. All original presumptive W-ExPEC isolates identified were included, including any W-ExPEC strains without clinical NMEC and BBEC matches as they could still represent strains with extraintestinal pathogenic potential that lacked a suitable clinical counterpart in the local repository used for this

analysis. An accessory genome clustering analysis was then conducted based on the binary presence and absence of accessory genes within the calculated pan-genome, as determined by Roary. The accessory genome clustering tree was then visualized and annotated using the R packages ggplot2 v3.4.2 (Wickham 2016), ape v5.4.1 (Paradis and Schliep 2019) and ggtree v2.2.4 (Yu *et al.* 2017).

The genomes of the selected W-ExPEC strains and their closest clinical counterparts were also screened for virulence genes and antibiotic resistance genes using ABRicate v1.0.1 (https://github.com/tseemann/abricate), with a minimum query coverage of 90% and a minimum percent identity of ≥ 80%. For virulence gene identification, W-ExPEC and C-ExPEC genomes were screened against the VFDB 2016 database (Chen *et al.* 2016) and *Escherichia coli* Virulence Factor Database (Ecoli_vf) (https://github.com/phac-nml/ecoli_vf). For antibiotic resistance gene identification, *E. coli* genomes were screened against the CARD 2017 database (Jia et al., 2017). The number of virulence and antibiotic resistance genes were then visualized and appended to the accessory genome tree using ggplot2.

## 6.3 Results

### 6.3.1 Identification of Presumptive W-ExPEC

Among 637 wastewater *E. coli* isolates collected in this study, 247 possessed at least one ExPEC-associated virulence gene, while 7 isolates harbored all seven virulence genes screened for (Figure 6-1A). Of the seven ExPEC virulence genes screened for, *fyuA* and *chuA* were the most prevalent, while *sfa/foc* and *ibeA* were the least common (Figure 6-1B). The 637 wastewater *E. coli* isolates were also screened for the major ExPEC pandemic lineage-associated ST131 marker, of which 22 were positive. As such, based on the screening criteria provided in *Materials and*

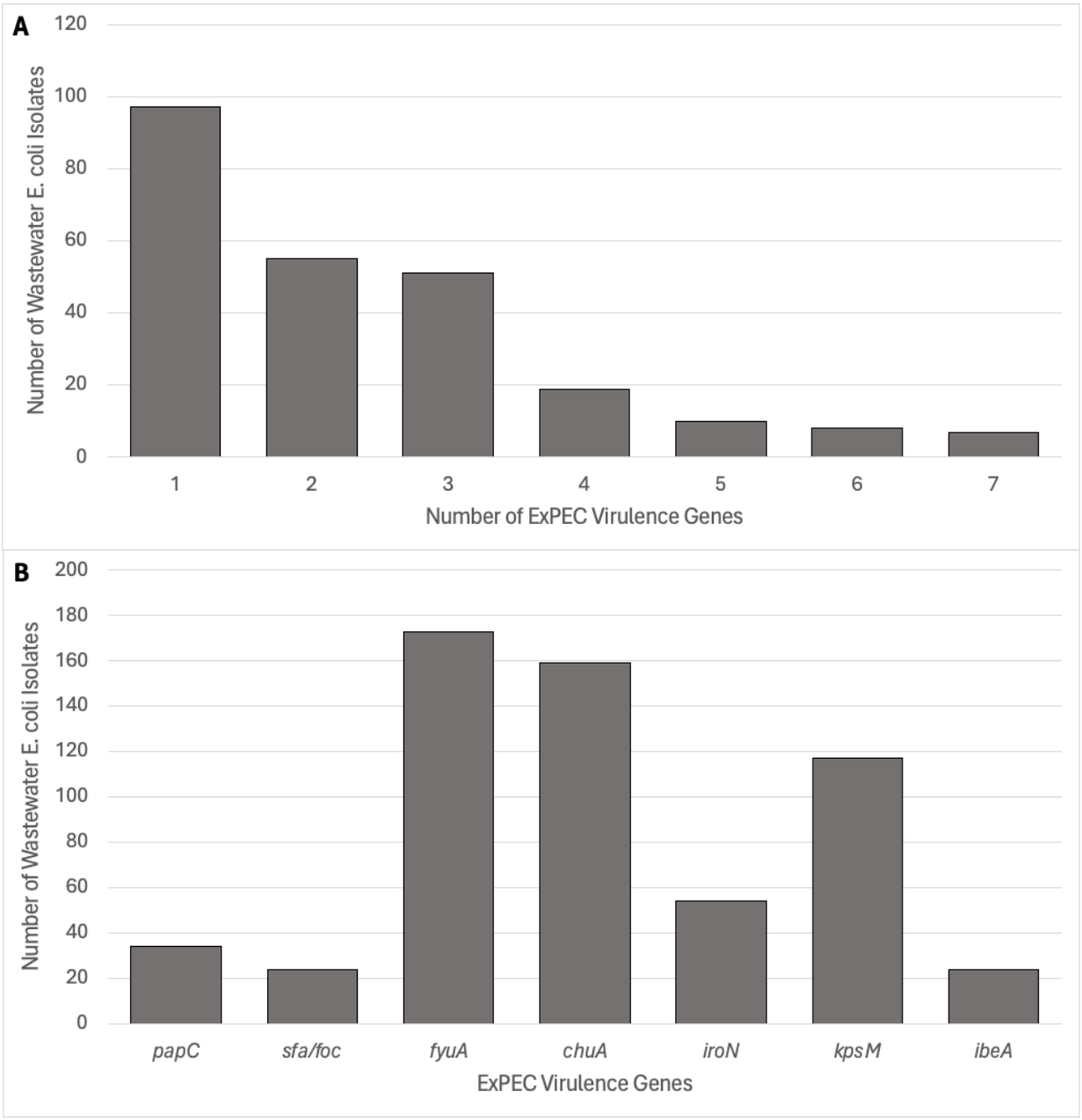



**Figure 6-1. ExPEC virulence gene screening of chlorine- and wastewater-treatment resistant *E. coli* isolates.** Number **(A)** of ExPEC virulence genes harbored by each wastewater isolate and **(B)** the frequency of each virulence gene across the screened isolates. A total of 376 presumed chlorine-tolerant isolates and 261 water treatment-resistant were screened against a PCR panel of ExPEC virulence gene markers. For isolates determined to be presumptive W-ExPEC, the presence of these virulence genes were confirmed with whole genome sequencing.

*Methods Section 6.2.3*, 86 isolates were identified as presumptive W-ExPEC and were selected for whole genome sequencing. The general characteristics of the 86 sequenced W-ExPEC genomes are summarized in Supplementary Table 6-S2.

### 6.3.2 Core Genome Similarity Between W-ExPEC and C-ExPEC Strains

As a first evaluation of the genomic similarity between W-ExPEC and C-ExPEC strains, the core genome SNP distance between all presumptive W-ExPEC (n = 86) and C-ExPEC (n = 320) strains was determined. Thirty-seven W-ExPEC strains were identified to share close core genome similarity to a clinical counterpart, based on an upper maximum difference of 250 SNPs across a ~417 kbp core genome backbone. Based on this criterion, select W-ExPEC strains displayed close similarity with anywhere from 1 to 48 clinical BBEC or NMEC strains (Supplementary Table 6-S3). In particular, select W-ExPEC isolates were found to differ by as few as only 3–6 core SNPs from a clinical BBEC (Table 6-2) and as few as 7 core SNPs from a clinical NMEC (Table 6-3). Interestingly, several wastewater strains (indicated with the 'WU' prefix) that were previously identified to be genetically similar to clinical UPEC strains (Zhi *et al.* 2020) were also found to share high core genome similarity with a clinical NMEC or BBEC strain.

### 6.3.3 Pairwise whole-genome similarity between W-ExPEC and C-ExPEC strains

Remarkably, all 37 W-ExPEC strains that shared close core genome similarity with a clinical counterpart also exhibited greater than 96.00% whole genome similarity with at least 1 clinical BBEC or NMEC strain. Indeed, identified W-ExPEC and C-ExPEC matches exhibited whole genome similarities ranging from 96.04% to 99.74% whole genome similarity, with some W-ExPEC strains exhibiting this degree of similarity with up to 48 C-ExPEC counterparts

**Table 6-2.** Pairwise whole genome similarity and core genome SNP distances between W-ExPEC strains and their two closest clinical BBEC counterparts

| Strain Name (Sequence Type) | Closest Clinical BBEC | Core genome SNP Difference[c] | Whole Genome Similarity | Strain Name (Sequence type) | Closest Clinical BBEC | Core genome SNP Difference[c] | Whole Genome Similarity |
|---|---|---|---|---|---|---|---|
| 1F2A (ST131) | BBEC_268 (ST131) | 6 | 97.80% | 1G6 (ST95)[a, b] | BBEC_41 (ST95) | 30 | 97.38% |
| | BBEC_267 (ST131) | 5 | 97.37% | | | | |
| 1G10A (ST131)[a] | BBEC_259 (ST131) | 12 | 97.50% | 3B9 (ST95)[a, b] | BBEC_38 (ST95) | 18 | 99.62% |
| | BBEC_210 (ST131) | 9 | 97.43% | | BBEC_35 (ST95) | 24 | 98.48% |
| 2B4 (ST131) | BBEC_268 (ST131) | 4 | 98.82% | 4G1 (ST95)[a, b] | BBEC_38 (ST95) | 20 | 99.58% |
| | BBEC_267 (ST131) | 5 | 98.46% | | BBEC_35 (ST95) | 26 | 98.44% |
| 2F5/2F6 (ST131)[a] | BBEC_156 (ST131) | 5 | 98.82% | 4G9 (ST95)[b] | BBEC_41 (ST95) | 15 | 97.00% |
| | BBEC_54 (ST131) | 7 | 98.54% | | | | |
| 3E4 (ST131)[a] | BBEC_267 (ST131) | 47 | 98.03% | WU1033 (ST95)[b] | BBEC_41 (ST95) | 23 | 96.51% |
| | BBEC_280 (ST131) | 51 | 97.48% | | | | |
| 3G8 (ST131) | BBEC_267 (ST131) | 6 | 97.41% | WU1151 (ST95)[b] | BBEC_38 (ST95) | 33 | 97.56% |
| | BBEC_268 (ST131) | 7 | 97.29% | | BBEC_162 (ST95) | 21 | 97.49% |
| 3G9 (ST131)[a] | BBEC_211 (ST131) | 3 | 98.80% | WU1274 (ST95)[b] | BBEC_38 (ST95) | 20 | 97.07% |
| | | | | | BBEC_162 (ST95) | 8 | 96.08% |
| 3G11 (ST131) | BBEC_211 (ST131) | 9 | 96.89% | WU1752 (unnamed)[b] | BBEC_273 (ST95) | 29 | 96.80% |
| | | | | | BBEC_283 (ST95) | 22 | 96.47% |
| 4C1 (ST131)[a] | BBEC_202 (ST131) | 8 | 98.96% | WU3707 (ST95)[b] | BBEC_281 (ST95) | 40 | 96.08% |
| | BBEC_54 (ST131) | 5 | 98.35% | | | | |
| 4C7 (ST131)[a] | BBEC_267 (ST131) | 49 | 98.71% | 3H3 (ST73) | BBEC_171 (ST73) | 38 | 96.45% |
| | BBEC_280 (ST131) | 53 | 98.70% | | BBEC_151 (ST73) | 34 | 96.35% |
| 4D7 (ST131)[a] | BBEC_202 (ST131) | 5 | 98.22% | 4F9 (ST73) | BBEC_171 (ST73) | 47 | 96.98% |
| | BBEC_9 (ST131) | 13 | 97.54% | | BBEC_151 (ST73) | 43 | 96.72% |
| 5A5 (ST131) | BBEC_280 (ST131) | 12 | 98.04% | WU965 (ST73) | BBEC_5 (ST73) | 15 | 97.80% |
| | BBEC_267 (ST131) | 8 | 96.85% | | | | |
| WU1030 (ST131) | BBEC_265 (ST131) | 5 | 99.74% | 1F4 (ST10) | BBEC_285 (ST8453) | 239 | 96.85% |
| | BBEC_158 (ST131) | 6 | 99.55% | | | | |
| WU1036 (ST131) | BBEC_265 (ST131) | 4 | 99.12% | 2F12 (ST10) | BBEC_29 (ST10) | 12 | 96.60% |
| | BBEC_158 (ST131) | 5 | 98.93% | | BBEC_225 (ST10) | 11 | 96.54% |
| WU1155 (ST131) | BBEC_265 (ST131) | 5 | 99.66% | WU1025 (ST127) | BBEC_69 (ST127) | 10 | 97.57% |
| | BBEC_158 (ST131) | 6 | 99.47% | | BBEC_71 (ST127) | 23 | 96.57% |
| WU1265 (ST131) | BBEC_265 (ST131) | 4 | 99.72% | WU1157 (ST127) | BBEC_71 (ST127) | 7 | 98.51% |
| | BBEC_158 (ST131) | 5 | 99.53% | | | | |
| WU1266 (ST131) | BBEC_265 (ST131) | 5 | 99.74% | WU1149 (ST357) | BBEC_141 (ST357) | 11 | 96.29% |
| | BBEC_158 (ST131) | 6 | 99.55% | | BBEC_131 (ST357) | 14 | 96.10% |
| 2H7 (ST44) | BBEC_88 (ST44) | 18 | 97.49% | WU664 (ST538) | BBEC_28 (ST538) | 34 | 98.90% |

[a] Wastewater strains with additional clinical BBEC matches than described in the table, based on a lower whole genome similarity threshold of 96.00% for pathotype similarity.
[b] Wastewater strains with clinical matches to both BBEC and NMEC strains.
[c] Calculated against a core genome backbone size of ~417 kbp.

**Table 6-3.** Pairwise whole genome similarity and core genome SNP distances between W-ExPEC strains and their two closest clinical NMEC counterparts

| Strain Name (Sequence Type) | Closest Clinical BBEC | Core genome SNP Difference[c] | Whole Genome Similarity |
|---|---|---|---|
| 1G6 (ST95)[a, b] | NMEC_9 (ST95) | 7 | 99.72% |
| | NMEC_24 (ST95) | 36 | 98.51% |
| 3B9 (ST95)[b] | NMEC_4 (ST95) | 13 | 99.59% |
| 4G1 (ST95)[b] | NMEC_4 (ST95) | 15 | 99.55% |
| 4G9 (ST95)[a, b] | NMEC_28 (ST95) | 24 | 97.37% |
| | NMEC_10 (ST95) | 38 | 97.35% |
| WU1033 (ST95)[a, b] | NMEC_10 (ST95) | 28 | 98.20% |
| | NMEC_24 (ST95) | 29 | 97.31% |
| WU1151 (ST95)[b] | NMEC_4 (ST95) | 28 | 97.18% |
| WU1274 (ST95)[b] | NMEC_4 (ST95) | 15 | 97.01% |
| WU1752 (ST95)[b] | NMEC_12 (ST95) | 20 | 97.75% |
| | NMEC_26 (ST95) | 16 | 97.06% |

[a] Wastewater strains with additional clinical NMEC matches than described in the table, based on a lower whole genome similarity threshold of 96.00% for pathotype similarity.
[b] Wastewater strains with clinical matches to both BBEC and NMEC strains.
[c] Calculated against a core genome backbone size of ~417 kbp.

(Supplementary Table 6-S3). Remarkably, several of these whole genome matches were found to be extremely similar, as select W-ExPEC strains were found to share high genomic similarity of up to 99.12–99.74% with a clinical BBEC (Table 6-2.13), and 99.55–99.72% with a clinical NMEC (Table 6). Interestingly, and as observed during the core genome comparisons, some wastewater strains (indicated with the 'WU' prefix) that were previously identified to share close genetic similarity to clinical UPEC strains described by Zhi et al. (2020) were again found to be extremely similar to clinical BBEC strains at the whole genome level. In contrast, the W-ExPEC isolates exhibited significantly lower whole genome similarity with intestinal pathogenic and naturalized wastewater *E. coli* strains, with similarity values ranging from as low as 65.9% to only as high as 95.3% (Supplementary Table 6-S4).

### 6.3.4 Core Genome Phylogenetics and Sequence Typing of W-ExPEC and C-ExPEC Strains

To understand the evolutionary relationships between the W-ExPEC and C-ExPEC strains, a maximum-likelihood core genome phylogenetic analysis was performed. Phylogenetically, the W-ExPEC strains were found to cluster amongst the clinical strains, and exclusively within phylogroups A and B2 (Figure 6-2). Thirty-four W-ExPEC strains were distributed amongst clinical NMEC and BBEC strains throughout phylogroup B2, a major phylogroup known to harbor ExPEC pathotypes (Johnson and Russo 2002; Picard *et al.* 1999), whereas the remaining 3 W-ExPEC strains clustered with their closest clinical match in phylogroup A. Although phylogroup A is classically considered to be non-pathogenic (Escobar-Páramo *et al.* 2006; Li *et al.* 2010) and negatively associated with the UPEC pathotype (Hutton *et al.* 2018), some studies have reported clinical ExPEC strains belonging to this phylogroup (Micenková *et al.* 2016). Reflecting this,

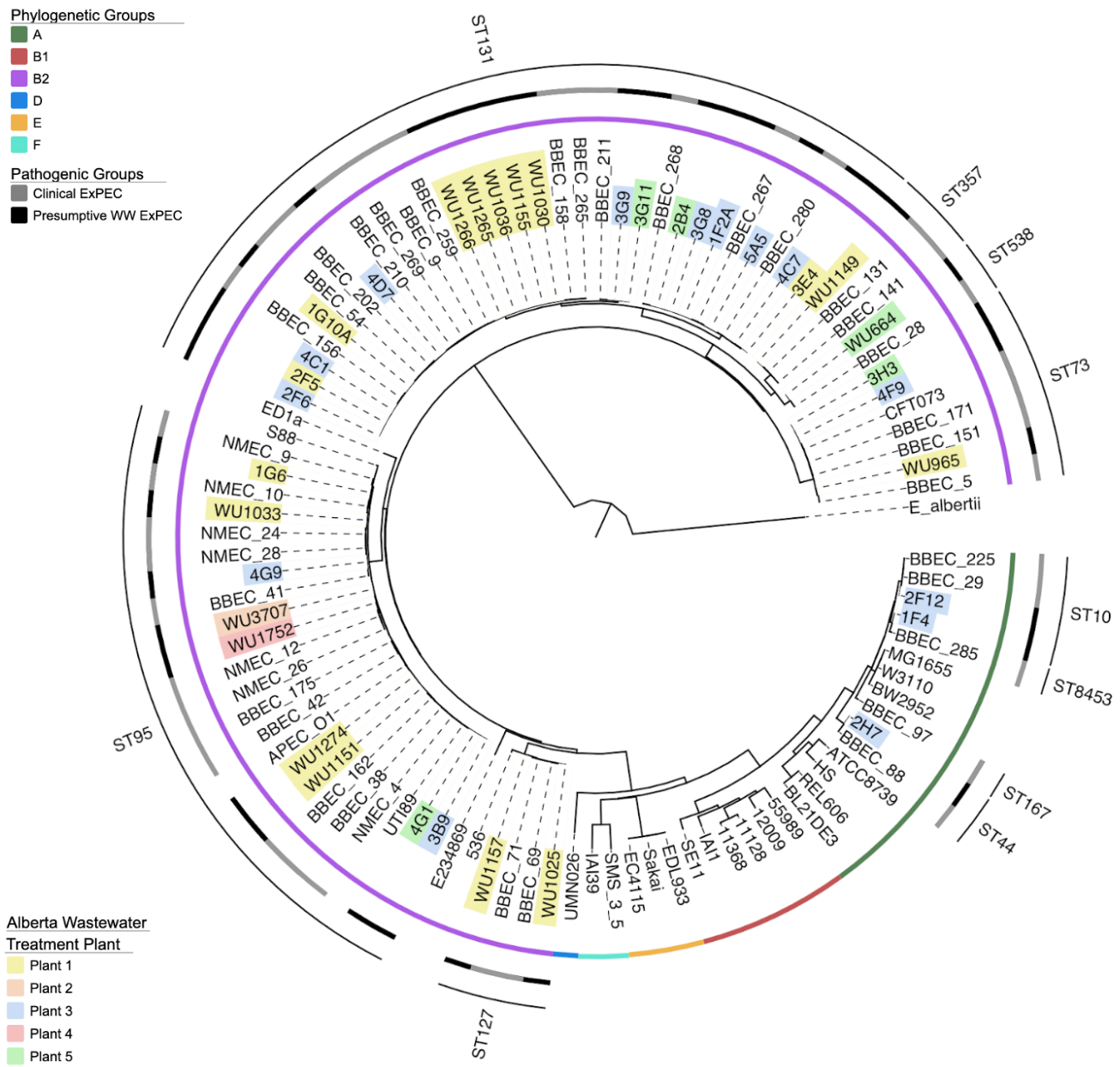



**Figure 6-2. Core genome maximum likelihood phylogenetic tree of wastewater ExPEC strains, their closest NMEC and BBEC counterparts and *E. coli* strains of known phylogroups**. Presumptive wastewater ExPEC strains were collected from samples of chlorinated sewage or treated wastewater effluent from 5 wastewater treatment plants across Alberta, Canada (highlighted according to the lower legend). The core genome phylogenetic similarity between the wastewater ExPEC strains (colored black in the middle ring), clinical NMEC and BBEC strains (colored grey in the middle ring), and various *E. coli* strains of known phylogroups (inner colored ring according to the upper legend) were compared. The main sequence type lineages of the wastewater and clinical ExPEC strains are indicated in the outermost ring. The tree is rooted against an *Escherichia albertii* strain as the outgroup.

within phylogroup A the three non-B2 W-ExPEC isolates still clustered separately from commensal strains into two separate sub-clusters (ST10 and ST44) alongside their clinical BBEC counterparts (Figure 6-2). Interestingly, although phylogroup D has been positively associated with ExPEC pathotypes other than UPEC (Hutton *et al.* 2018), none of the wastewater ExPEC or clinical BBEC and NMEC strains clustered within this phylogroup.

MLST analysis of the 37 W-ExPEC and 38 C-ExPEC strains revealed extensive sub-structuring by sequence type (Figure 6-2). All 37 W-ExPEC strains could be distributed across 8 STs, including those of clinical importance. The major ExPEC lineage-associated sequence type ST131 (Nicolas-Chanoine *et al.* 2014) was the most represented amongst the W-ExPEC strains, with 18 of 37 (48.65%) isolates clustering within this sequence type, which was also confirmed by the PCR screening panel. The NMEC-associated sequence type ST95 (Riley 2014) was also well represented, characterizing 9 out of 37 (24.32%) wastewater isolates. Of the remaining W-ExPEC strains, 3 isolates were designated as ST73 (8.11%), followed by 2 isolates each that were designated as ST10 and ST127 (5.41% each), and 1 isolate each designated as ST357, ST538, and ST44 (2.70% each). The 38 C-ExPEC strains were distributed in a similar manner across the 8 STs represented, as each W-ExPEC strain that was assigned a sequence type belonged to the same cluster as their closest clinical match.

Interestingly, the W-ExPEC strains were isolated across 5 different WWTPs sampled across Alberta, Canada (Figure 6-2). Considering that all 8 STs identified represented clinically important sequence types (Matsumura *et al.* 2017; Riley 2014), and multiple STs were often detected in the effluents of a single treatment plant, chlorine-tolerance and wastewater-treatment resistance appear to be prevalent phenotypes across multiple W-ExPEC-associated lineages.

### 6.3.5 Pan-Genomic Similarity and Accessory Genome Clustering of W-ExPEC and C-ExPEC strains

At the core- and whole-genome level, the W-ExPEC and C-ExPEC strains appear to be highly similar; however, measures of genetic similarity alone may not accurately reflect whether the W-ExPEC strains may share a similar pathogenic potential with their closest clinical counterparts. In contrast to the core genome, which encodes for genes mediating essential housekeeping functions, the accessory genome encodes for genes linked to adaptation, virulence, and antibiotic resistance, and are thus likely reflective of the predominant lifestyle of a given strain (Mira *et al.* 2010). As such, a comparative pan-genomic analysis was performed to determine whether the W-ExPEC strains could share a similar accessory gene profile, and thus similar ecotypic characteristics (i.e., pathogenic potential), with a clinical BBEC or NMEC counterpart.

A pan-genome was calculated for all 86 original presumptive W-ExPEC strains, the 38 closest clinical NMEC/BBEC matches, 9 representative UPEC reference strains, 5 representative naturalized wastewater strains (Zhi *et al.* 2019; Zurfluh *et al.* 2017), and 4 laboratory reference strains (Supplementary Table 6-S1). The pan-genome was estimated to consist of 26865 genes, of which 2133 were considered core, indicating a high level of pan-genomic diversity amongst all the strains analyzed. Indeed, the core genes comprised only ~8% of all pan-genome content, whereas 77% of the pan-genome consisted of genes that were shared by fewer than 15% of the strains analyzed (Figure 6-3). As the W-ExPEC strains possessed an average of 4729 genes in total, the core genes comprised roughly 45% of the W-ExPEC genome. Although this percentage is higher than typical estimates for *E. coli* core genomes (Lukjancenko *et al.* 2010), the majority of the strains evaluated in this analysis either shared a similar ecological niche (i.e., treated wastewater matrices) or a similar extraintestinal pathogenic potential (i.e., wastewater and clinical

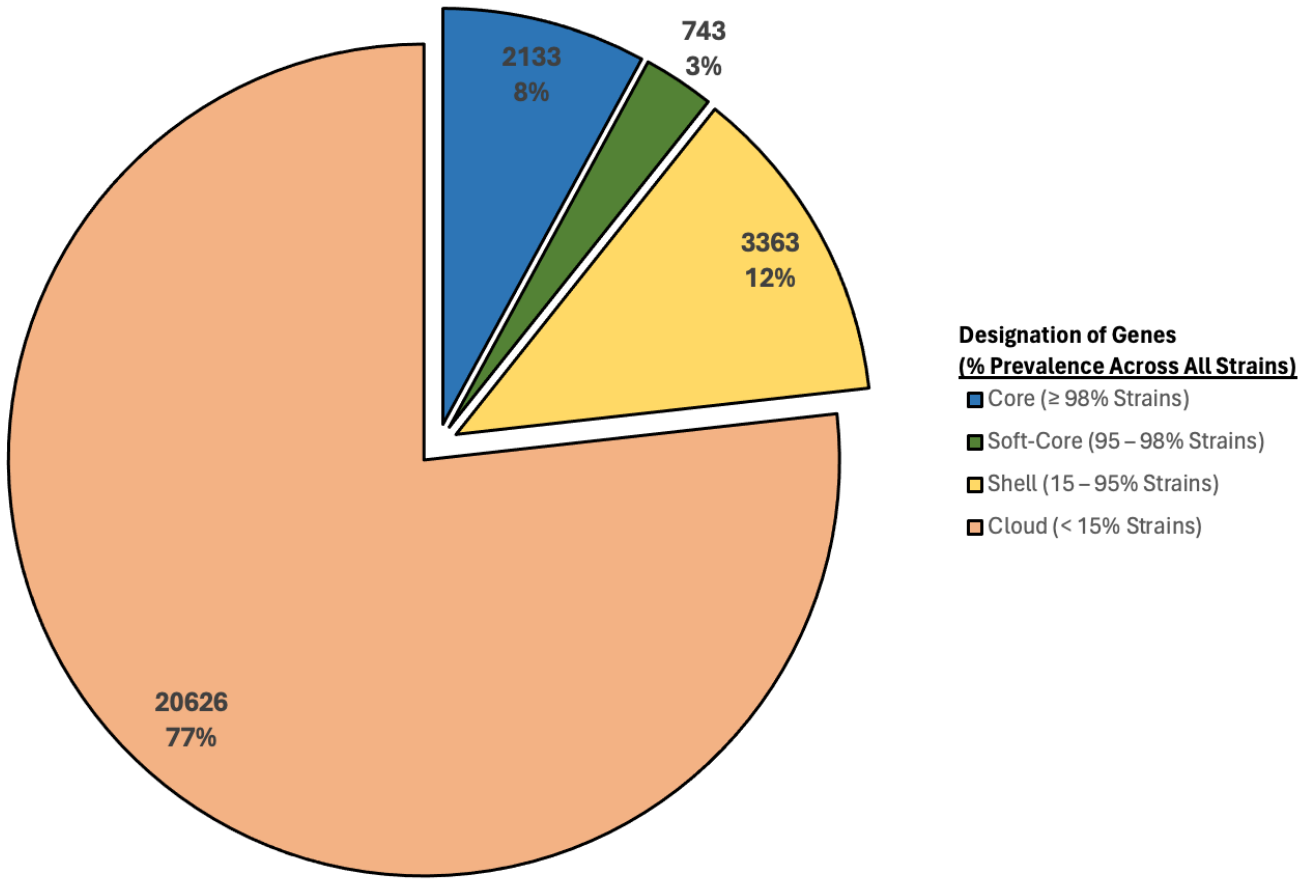



**Figure 6-3. Pan-genome dynamics of W-ExPEC strains, their closest clinical counterparts, and reference UPEC, naturalized wastewater and laboratory *E. coli* strains.** Roary was used to estimate a pan-genome to determine the distribution of genes across all strains analysed. Strains were provided different designations based on their prevalence across the strains included in the pan-genome, as: 'core' if the gene was found in ≥ 98% of all strains (blue); 'soft-core' if the gene was found in 95-98% of all strains (green); 'shell' if the gene was in 15-95% of all strains (yellow); and 'cloud' if the gene was found in less than 15% of all strains (orange).

ExPEC strains). The remaining 55% of genes in each genome would thus presumably consist of accessory genes that could play adaptive roles in the biology of the W-ExPEC strains, including those that may produce an extraintestinal pathogenic phenotype similar to clinical BBEC (i.e., septicemia) or NMEC (i.e., meningitis) strains.

To assess the degree of similarity of the wastewater and clinical strains at the accessory genome level, a clustering analysis was performed based on the binary presence and absence of accessory genes within the pan-genome. According to the generated clustering tree, the isolates were grouped into three main clusters (Figure 6-4A). The most basal cluster (Cluster 1) consisted of *E. coli* isolates that did not share any significant similarities at the whole or core genome level with any clinical ExPEC, wastewater ExPEC, laboratory reference or naturalized wastewater *E. coli* strains. Clusters 2 and 3, on the other hand, included the naturalized wastewater (Cluster 2a, including all ST635 strains) as well as all wastewater ExPEC and clinical ExPEC (Cluster 2b and Cluster 3) strains. The Cluster 2 W-ExPEC isolates were distributed across two main subclusters designated 2a and 2b, with 2 of the 86 sequenced wastewater *E. coli* isolates (2F11 and 1H6) clustering with previously characterized naturalized wastewater *E. coli* strains including WW223, WW10 (Zhi *et al.* 2019) and ABWA45 (Zurfluh *et al.* 2017). Conversely, Cluster 2b consisted of a small group of W-ExPEC strains (4H1, 2C8, 2F12, 2H7 and 1F4) that grouped with their corresponding clinical ExPEC strains (BBEC_29, BBEC_88, BBEC_285 and the UPEC strain *E. coli* 219). These clinical ExPEC strains were designated as ST10 and ST44 strains, representing relatively minor ExPEC lineages. Some W-ExPEC strains in this cluster, however, did not have a close clinical match (4E10 and 4G8), potentially due to the under-representation of these minor ExPEC ST lineages in the local repository. The last and largest cluster included most of the W-ExPEC isolates and their closest clinical NMEC and BBEC matches, distributed across the major

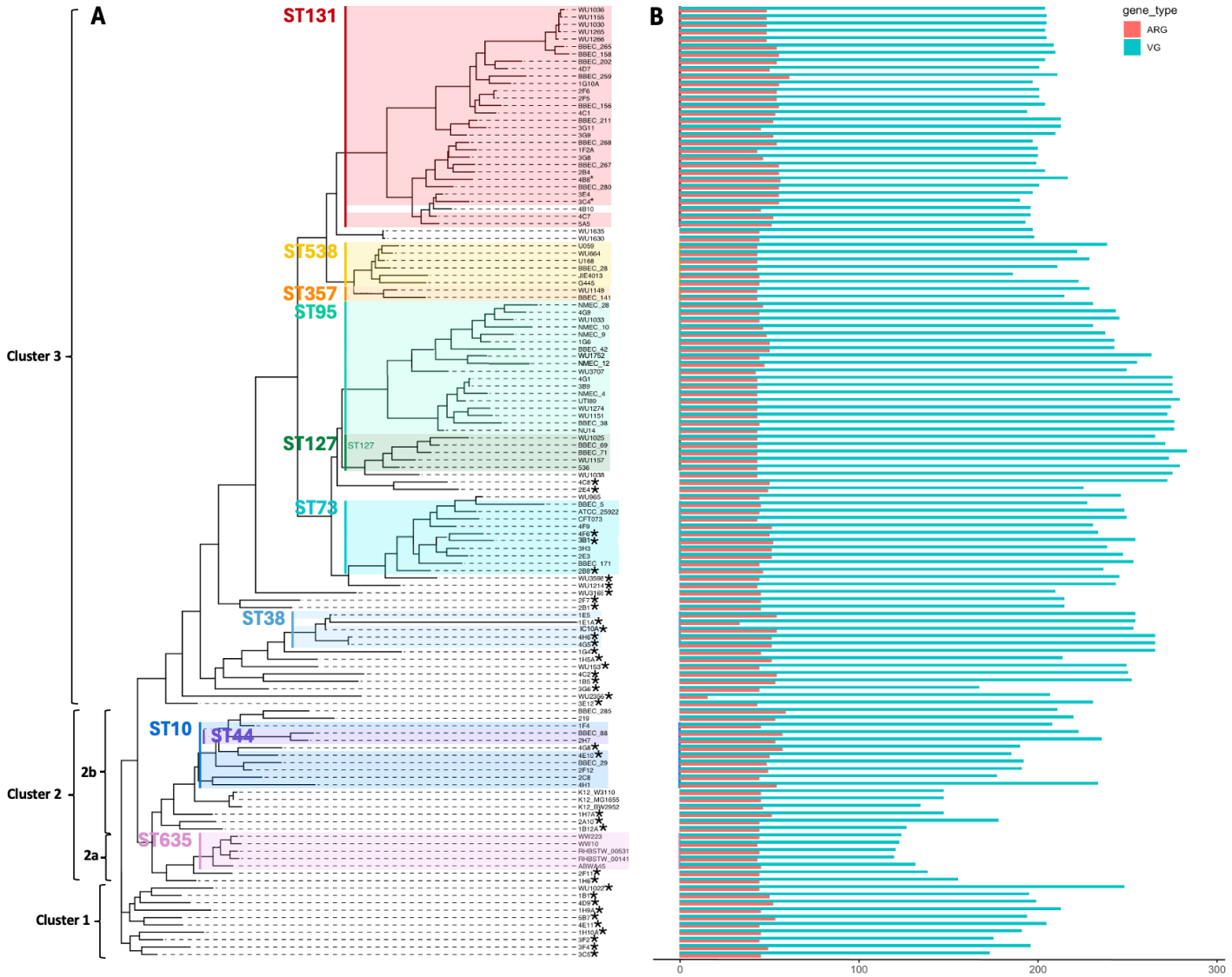



**Figure 6-4. Accessory genome clustering and virulence and antibiotic resistance gene screening of W-ExPEC and C-ExPEC strains.** Binary accessory gene presence absence clustering tree **(A)** of all original presumptive wastewater ExPEC strains (n=86), as well as clinical ExPEC strains with close W-ExPEC matches, naturalized wastewater *E. coli* strains, and reference laboratory reference strains. Appended to the clustering tree is a bar chart **(B)** depicting the number of antibiotic resistance genes (ARG; red bars) and virulence genes (VG; teal bars) present in each isolate. The major sequence types represented are indicated in highlighted boxes in the clustering tree. Any original presumptive wastewater ExPEC strains that did not exhibit high genetic similarity with a clinical ExPEC strain at the whole and core genome levels are indicated in the clustering tree with an asterisk (*).

ExPEC lineages, including ST131, ST95, ST73, and ST127 (Riley 2014), as well as other important ExPEC lineages such as ST538 and ST357. While most W-ExPEC strains clustered closely with their closest clinical BBEC or NMEC counterpart, some wastewater ExPEC strains grouped closer to a reference UPEC strain, including WU664 (W-ExPEC) with U059 (C-ExPEC) and 4F9 (W-ExPEC) with CFT073 (C-ExPEC). Although several W-ExPEC strains within Cluster 3 did not have a direct clinical match (i.e., > 96.00% whole genome similarity to a clinical ExPEC in the local repository) they still grouped within accessory genome-based clusters dominated by clinical ExPEC strains, suggesting that they may possess the accessory genes necessary for an extraintestinal pathogenic phenotype. Indeed, the finding that W-ExPEC strains clustered closely alongside clinical ExPEC strains based on accessory gene composition suggests that they may exhibit a similar pathogenic presentation to their clinical counterparts.

### 6.3.6 Virulence and Antibiotic Resistance Gene Screening of W-ExPEC and C-ExPEC strains

Given the close genetic similarity between select W-ExPEC and C-ExPEC strains across the core, whole and accessory genome, their virulence gene (VG) and antibiotic resistance gene (ARG) repertoires were examined to better clarify the pathogenic potential of the W-ExPEC strains. Virulence gene composition was found to be roughly bimodally distributed across the accessory genome analysis (Figure 6-4B), with the 2a Cluster of naturalized wastewater strains having the lowest number of virulence genes compared to the flanking strains in Cluster 1 and Cluster 3 (W-ExPEC and C-ExPEC strains). In contrast, W-ExPEC strains sharing close similarity to at least one clinical ExPEC strain all had extensive virulence repertoires, ranging from 195 to 278 virulence genes depending on the strain (Supplemental Table 6-S5). Although the number of

virulence genes were consistently highest among the ST95, ST127, and ST73 strains, the W-ExPECs appeared to share a similar complement of VGs with a clinical NMEC or BBEC. Indeed, when compared to their closest clinical match, select W-ExPECs were found to differ by as few as only 1 or 2 VGs in their entire virulence repertoires (Supplementary Table 6-S5).

In terms of antibiotic resistance, the W-ExPEC strains possessed anywhere from 43 to 56 ARGs, though generally their clinical counterparts harbored more (Supplementary Table 6-S6). In particular, the ST131 W-ExPEC strains appeared to possess the highest number of ARGs, followed by W-ExPECs belonging to the ST44 and ST73 sequence types. Interestingly, several ARGs of major clinical importance were identified in the W-ExPEC strains, including various aminoglycoside modification enzyme (AME) genes (*AAC(3)-IId*, *AAC(3)-IIe*, *AAC(6')-Ib-cr*, *APH(3")-Ib*, *APH(6)-Id*), sulfonamide resistance genes (*sul1*, *sul2*, *sul3*), and tetracycline efflux genes (*tet(A)*, *tet(B)*). Additionally, several strains were found to carry various beta-lactamases ($bla_{\text{OXA-1}}$, $bla_{CTX\text{-}M\text{-}14}$, $bla_{CTX\text{-}M\text{-}15}$, $bla_{CTX\text{-}M\text{-}27}$, $bla_{TEM\text{-}181}$, *ampC*, *ampH*), such that they could represent multi-drug resistant, ESBL-producing strains. Overall, the W-ExPEC strains shared similar ARG repertoires with NMEC or BBEC strains, with select W-ExPECs either differing from their closest clinical match by only 1 ARG or sharing the same complement of ARGs with a clinical counterpart (Supplementary Table 6-S6).

## 6.4 Discussion

Whether for drinking, agricultural, or recreational purposes, access to safe, clean water is of paramount importance to public health. While this is most apparent with drinking water (Cutler and Miller, 2005; Hunter *et al.* 2010), effective wastewater sanitation also represents one of the most important interventions for public health. Across 39 nations, inadequate wastewater treatment

was found to correlate with increased disease mortality, irrespective of national income, development, and overall sanitation (Naik and Stenstrom 2012). As such, the treatment and sanitation of all urban water supplies serves as a key preventative measure for the control and mitigation of infectious diseases. What if, however, microbes could evolve resistance against wastewater treatment and disinfection? The importance of wastewater treatment in the maintenance of a hygienic urban water infrastructure rests on the fundamental role of sanitary management in eliminating microbes (i.e., pathogens) from the water supply. Concerningly, accruing evidence suggests that microbes could be evolving resistance to wastewater treatment.

The polyphasic characterization of the WWS-*E. coli* outlined throughout this thesis strongly illustrate this prospect. Chapters Two to Five demonstrated that these naturalized wastewater strains appear to be genotypically, phenotypically, and ecotypically adapted to wastewater, having acquired the necessary cellular mechanisms to resist wastewater treatment and exploit this anthropogenic, engineered environment as a niche. Interestingly, the evidence presented in this chapter suggests that the distribution of genetic determinants underlying the phenotype of wastewater treatment resistance may extend beyond these WWS-*E. coli* strains alone. Indeed, the relatively close clustering of the naturalized wastewater strains (Cluster 2a) with wastewater-derived and clinical ExPECs (particularly those within Cluster 2b) based on accessory genome content suggests that potentially extraintestinal pathogenic *E. coli* strains may also be genetically equipped to survive across the wastewater treatment train.

Indeed, the remarkable degree of genomic similarity identified between W-ExPEC and C-ExPEC strains suggests that a significant proportion of the *E. coli* population surviving wastewater treatment could represent highly pathogenic strains capable of causing extraintestinal diseases such as septicemia, meningitis and, based on the work conducted by Zhi *et al.* (2020), urinary tract

infections. Reflecting this, previous studies have found that a significant proportion of the *E. coli* present within finished effluents may represent ExPEC strains (Adefisoye and Okoh 2016; Anastasi *et al.* 2010; Anastasi *et al.* 2013; Calhau *et al.* 2015; Mahfouz *et al.* 2018). A major limitation of these studies, however, was a common reliance on virulence gene screening as the primary means of identifying potential ExPEC in wastewater samples. Unfortunately, virulence gene screening alone provides limited information on a given strain's pathogenic potential since no single VG or set of VGs can differentiate the ExPEC pathotypes (Sarowska et al., 2019), especially given their presence within commensal strains as common host-adaptive genes (Bok *et al.* 2018; Qin *et al.* 2013).

Instead, taking a more comprehensive, whole genome-based approach, as conducted in this chapter as well as in more recent studies (Paulshus *et al.* 2019; Zhi *et al.* 2020), may provide better clarity on the pathogenic potential of *E. coli* surviving wastewater treatment. In the present analysis, multiple measures of genomic similarity were integrated to evaluate the pathogenic capacity of *E. coli* strains surviving wastewater treatment. Essentially, the higher the similarity between wastewater and clinical strains across all genomic levels, the more likely these wastewater-derived strains could represent potential ExPEC. Indeed, of the 637 *E. coli* isolates that were initially recovered from chlorinated sewage and treated wastewater effluent samples, a select number were found to be extremely similar to a clinical BBEC or NMEC counterpart across all comparative genomic methods used in this study (i.e., core, whole, and accessory genome similarity). For instance, the wastewater strains 2F5 and 2F6 were found to share 98.82% whole genome similarity with the septicemia-causing strain 'BBEC_156' and differed by only 5 SNPs across a ~417 kbp core genome backbone. In terms of accessory gene content, these wastewater strains also shared 203 common VGs and harbored an ARG repertoire that differed by only 1

resistance gene (i.e., lacking the beta-lactamase *bla*$_{\text{CTX-M-9}}$) when compared to their BBEC counterpart. Similarly, the wastewater strain 3G9 shared 98.80% whole genome similarity with the bloodborne strain 'BBEC_211', while differing by only 3 core SNPs and sharing 212 common VGs and an identical ARG repertoire. Additionally, compared to the clinical strain 'NMEC_4', the wastewater strains 3B9 and 4G1 each shared over 99.50% whole genome similarity their clinical counterpart, differed by only 13 and 15 core genome SNPs respectively, and shared an identical ARG profile and over 276 common VGs.

It is important to note that while the wastewater strains described above appear to represent the strongest ExPEC candidates in this analysis, the other wastewater strains may still possess extraintestinal pathogenic potential even if they did not exhibit the same degree or consistency of genomic similarity with a clinical counterpart. Indeed, this analysis primarily focused on NMEC and BBEC strains belonging to the major ExPEC lineages (i.e., ST131, ST95). Thus, W-ExPEC isolates sharing high genomic similarity with clinical UPEC strains (Zhi *et al.* 2020), or those belonging to more minor ExPEC lineages, may have lacked proper clinical representatives within the local repository in this analysis. As such, broadening the representation of clinical ExPECs in the repository, such as by including additional reference strains belonging to the other extraintestinal pathotypes (i.e., UPEC, extraintestinal EAEC, etc.) and minor ExPEC ST lineages (i.e., ST10, ST44, ST38, etc.), could improve upon the number of wastewater isolates identified as potential ExPEC.

This becomes especially important when considering the potential role of the naturalized wastewater strains in the emergence of wastewater treatment resistance in the W-ExPECs. Based on the evidence presented in Chapter Four of this thesis, the naturalized wastewater strains may be actively involved in the dissemination of treatment resistance determinants to other microbes

that are present within the wastewater matrix. Indeed, the findings in this chapter, as well as previous work conducted by Zhi *et al.* (2019), indicate that the naturalized WWS-*E. coli* and ExPEC populations may share commonalities in accessory gene content. Interestingly, a subset of the W-ExPECs, specifically those belonging to the minor ExPEC lineages including ST10 and ST44 (Riley 2014), were found to group especially close to the naturalized wastewater strains (i.e., Cluster 2a and Cluster 2b) based on accessory genome similarity. These ST10 and ST44 W-ExPEC groups were also found to cluster within the same phylogroup (i.e., phylogroup A) as the naturalized WWS-*E. coli* strains. As such, given their similar accessory genomic and phylogenetic backgrounds, the naturalized wastewater strains may play an important role in the dissemination of wastewater treatment resistance determinants to ExPEC strains that specifically belong to these emerging ExPEC lineages (Riley 2014).

The apparent evolutionary emergence of wastewater treatment resistance in ExPEC carries important implications for public health. A dual phenotype of wastewater treatment resistance and extraintestinal pathogenesis suggests that, at some minimal level, sustained waterborne transmission of these W-ExPECs must occur – particularly from wastewater-contaminated environments. ExPEC are currently recognized as a leading, and growing, cause of extraintestinal diseases including urinary tract infections, bacteraemia, and neonatal meningitis (Ouchenir *et al.* 2017; Poolman and Wacker 2016); however, the number of reported ExPEC infections likely underestimates their true scope and burden, such that the community infection rates are likely higher than expected. These elevated infection rates not only provide the sustained evolutionary pressure for the maintenance of pathogenesis within the host population, but also drives the continual shedding of ExPEC from the population back into wastewater treatment plants, which in turn leads to the evolution of wastewater treatment resistance.

The ability to maintain the phenotypes of wastewater treatment resistance and extraintestinal pathogenesis then requires that the W-ExPEC surviving disinfection must cycle back into the human population. Wastewater has been well-documented to impact (i.e., contaminate) both environmental and urban water supplies (Rice and Westerhoff 2015; Rice and Westerhoff 2017), and exposure to these contaminated waters, particularly through recreational use, has already been epidemiologically linked to increased rates of urinary tract disease due to UPEC (Søraas *et al.* 2013). Given that ESBL-producing *E. coli* isolated from regional clinical, wastewater, and recreational water samples have been found to be clonal (Jørgensen *et al.* 2017), a considerable degree of microbial cycling appears to occur between the human population, wastewater, and recreational or drinking water supplies. With ESBL-producing *E. coli* appearing to be common constituents within wastewater (Paulshus *et al.* 2019), wastewater contamination may introduce potentially extraintestinal pathogenic, multi-drug resistant *E. coli* into drinking (Tanner *et al.* 2019) and recreational water supplies (Blaak *et al.* 2014). While typically associated with diarrheal diseases, this evidence suggests that water may also represent an important, yet overlooked, source of transmission for extraintestinal pathogens. Indeed, it appears that the cautionary tale posed by the naturalized wastewater *E. coli* may already be a reality – that the same evolutionary forces underlying the niche-specificity of the WWS strains may be similarly driving pathogenic microbes to evolving resistance to wastewater sanitation and disinfection practices.

# Chapter Seven: General Discussion

Contrary to its traditional view as a host- and niche-generalist, a growing body of evidence suggests that the *Escherichia coli* species exhibits a significant degree of niche-specificity. As reviewed in Chapter One of this thesis, various genotypic and phenotypic markers of *E. coli* niche-specificity have been documented to date, indicating that the species may be more aptly described using a species-complex model (Yu *et al.* 2021). The evidence presented in the preceding chapters of this thesis reinforce this concept, particularly through the polyphasic characterization of a unique, WWS-*E. coli* ecotype that appears to have evolved to exploit wastewater as a primary niche.

Through the work presented in Chapters Two to Five, we were able to address the three main objectives outlined for this thesis, which included:

i) Validating the utility of logic regression as an ecotype discovery and attribution tool;

ii) Demonstrating how genotype, phenotype, and ecotype can recapitulate the evolution of niche-specificity in *E. coli*, through the polyphasic characterization of naturalized *E. coli* ecotypes that have emerged within man-made, engineered environments; and

iii) Evaluating how the processes of niche-adaptation can lead to the phenotypic convergence of different ecotypes, through the characterization of wastewater treatment resistance in WWS-*E. coli* and wastewater-borne ExPEC strains.

The main findings of this thesis, particularly as they relate to the major research questions above, are discussed in greater detail in the following section.

## 7.1 Major Findings and Contributions

The major findings, as presented in this thesis, are as follows:

i) Logic regression analysis of ITGR sequence data represents an effective approach for the discovery and source attribution of putative ecotypes within the *E. coli* species;

ii) Distinct *E. coli* ecotypes appear to have evolved to become genotypically, phenotypically, and ecotypically adapted to man-made, engineered environments as primary niches;

iii) Naturalized-engineered *E. coli* strains specifically residing in wastewater may be involved in the dissemination and evolution of antibiotic and potentially treatment resistance in the wastewater environment; and

iv) The conditions and challenges of the wastewater environment appear to have similarly driven the evolution of ExPEC towards wastewater treatment resistance.

### 7.1.1 Logic Regression as an Ecotype Discovery and Attribution Tool

Confirming previous studies conducted by Zhi *et al.* (2015; 2016a; 2016b), we were able to utilize logic regression to identify sensitive and highly ecotype-specific SNP-SNP biomarkers across a wide range of host- and environmentally-derived (i.e., wastewater) *E. coli* strains. Remarkably, even when applied against an expanded repository of strains, the ability of logic regression in identifying ecotype-informative biomarkers did not appear to be significantly diminished. Furthermore, several improvements were made to the logic regression workflow as previously established by Zhi *et al.* (2015; 2016a; 2016b), including adopting a more iterative approach to the building and statistical validation (i.e., 10-fold cross validation versus 5-fold cross validation) of the logic models. While previous work demonstrated that biomarker performance could be improved by concatenating several ITGR sequences (Zhi *et al.* 2015; Zhi *et al.* 2016a), we also found that an iterative approach to sequence concatenation (i.e., by evaluating different

permutations of the ITGR sequences being concatenated) resulted in further improvements to the sensitivities and specificities of biomarkers that were generated for several source categories. Moreover, alongside other ITGRs that were previously found to encode source-informative information (*uspC–flhDC*, *csgDEFG–csgBAC*, *asnS–ompF*, and *yedS–yedR*) (Zhi *et al.* 2015; Zhi *et al.* 2016b), we were able to identify additional ITGR targets (*emrKY–evgAS*, *yedS–yedR*, *ompC–rcsDB*, and *ibsB–[mdtABCD-baeSR]*) that appeared to be particularly ecotype informative.

Reinforcing the ecotypic relevance of our biomarker discovery approach, we also extended the use of logic regression for ecotype attribution purposes. In Chapter Two of this thesis, logic regression was successfully used to predict the original host source of 48 out of 113 environmental *E. coli* isolates collected in Sweden with a high degree of consensus. While other groups have utilized alternative supervised learning methods for bacterial source attribution analyses (Lupolova *et al.* 2017; Lupolova *et al.* 2019; Zhang *et al.* 2019b), our study represents the first to apply such algorithms for the source attribution of bacterial isolates from unknown sources. The utility of logic regression for ecotype attribution purposes was further reinforced in Chapter Three, as it was found to outperform other ecotype prediction methods in correctly clustering meat plant- and wastewater-derived *E. coli* isolates into naturalized-engineered ecotypic groups that were separate from their host-associated and environmental counterparts. As such, our results highlight logic regression as a particularly effective approach for the discovery and source attribution of putative *E. coli* ecotypes.

### 7.1.2 Polyphasic Characterization of Naturalized-Engineered *E. coli* Ecotypes

The central focus of this thesis involved the polyphasic characterization of *E. coli* strains that appear to have evolved to exploit engineered environments as primary niches. In Chapter

Three, *E. coli* strains collected from wastewater and meat plants were found to group into distinct, naturalized-engineered-associated phylogenetic and ecotypic clusters. Despite a relatively wide geographical distribution (particularly for the wastewater isolates), the majority of the naturalized-engineered strains clustered within two sequence types, ST635 and ST399. Interestingly, these sequence types have been previously identified in *E. coli* populations derived from other engineered environments, including septic tanks (Behruznia *et al.* 2022a; Behruznia *et al.* 2022b) and sink drainage systems (Constantinides *et al.* 2020), suggesting they may represent specific *E. coli* lineages that are predisposed to becoming naturalized within built environments. Furthermore, considering that multiple serotypes were also found to be represented across the wastewater and meat plant strains, several *E. coli* lineages appear to have independently emerged across various food- and water-associated engineered environments, collectively constituting distinct naturalized-engineered *E. coli* ecotypes.

Reinforcing these findings, the insertion element marker screening analyses in Chapter Three and pan-genome-wide-association studies in Chapter Four revealed that the WWS- and MPS-*E. coli* strains harbored various genetic adaptations that appeared to reflect their naturalization towards, and niche-specificity to, their respective engineered environments. Specifically, these strains were characterized by an over-representation of genes and intergenic insertion markers associated with functions important for surviving outside of the host gastrointestinal environment, including biofilm formation, microbial defense (i.e., inter-microbial competition, protection from phages, etc.), and stress resistance (i.e., DNA-damaging stimuli, oxidative stress, heat, heavy metals, etc.). Interestingly, a subset of these genetic features appeared to be specific to the wastewater (i.e., heavy metal resistance, phage defense) or meat plant strains (i.e., carnitine utilisation, chlorine resistance), suggesting they may represent distinct ecotypic

groups defined by unique niche-adaptive characteristics. Conversely, a multitude of genes associated with colonization (i.e., colonization factors, gut environment-associated stress resistance genes, etc.) and virulence (i.e., iron acquisition, secretion systems, etc.) were either found to be absent or disrupted by insertion elements across the naturalized-engineered strains, indicating their naturalization may have come at a cost of fitness within the host environment.

In Chapter Five, some of these genotypic findings were recapitulated through discernable wastewater-adaptive phenotypic traits within the WWS-*E. coli* strains. Compared to their host-associated counterparts, the wastewater strains were found to be significantly more resistant to heat, and even more resistant than previously found by Wang *et al.* (2020), as they could withstand temperatures of up to 64°C for 5 minutes and 72°C for 30 seconds. Interestingly, the WWS-*E. coli* strains appeared to exhibit slower growth kinetics compared to their host-derived counterparts, which could reflect an adaptation to the contaminant-rich, nutrient-deprived environment of wastewater. The WWS-*E. coli* strains were also found to be robust biofilm producers, mirroring previous findings from Zhi *et al.* (2017), especially under nutrient-deprived and low temperature conditions. Considering that previous studies have also demonstrated these wastewater strains to exhibit enhanced resistance to chlorine and other oxidizing agents (Wang *et al.* 2020), these findings collectively indicate that the WWS-*E. coli* appear to be phenotypically adapted to survive within the wastewater environment.

As such, the data presented in Chapters Three to Five demonstrate that strains belonging to the WWS-*E. coli* ecotype appear to be genotypically, phenotypically, and ecotypically adapted to the wastewater environment. As demonstrated throughout this thesis, these wastewater strains served as an effective model system for our polyphasic workflow in understanding microbial niche-specificity. From an ecotypic perspective, the ecological niche of wastewater presents

unique stressors and challenges that could drive the evolution of select *E. coli* populations to exploit the wastewater environment. In line with this, the wastewater strains harbored unique genetic repertoires that appear to promote their survival against the harsh conditions of wastewater. These genetic features were then recapitulated in phenotype, as the WWS-*E. coli* were characterized by various phenotypic traits reflecting their adaptation to, and survival across, the wastewater treatment train. Although the evidence provided by each perspective alone may be circumstantial, the strength of the polyphasic approach lies in its consideration of all perspectives collectively. Indeed, combining and integrating the ecotypic, genotypic, and phenotypic evidence provides a more comprehensive understanding of the adaptive mechanisms potentially underlying the evolution and niche-specificity of the WWS-*E. coli* ecotype.

### 7.1.3 The Potential Role of WWS-*E. coli* in the Dissemination of Resistance in Wastewater

As they were found to be specifically adapted to the wastewater niche, the WWS-*E. coli* strains may be capable of establishing resident populations within wastewater treatment plants. Given that wastewater is considered to be an environmental hotspot for resistance (Kunhikannan *et al.* 2021; Rizzo *et al.* 2013), coupled with the finding that disinfection-related stressors can promote increased rates of HGT through the SOS response (Aminov 2011; Zhang *et al.* 2021), this WWS ecotype might serve as a reservoir of key resistance determinants within the wastewater environment. Interestingly, in Chapter Four several wastewater strains were found to harbor various antibiotic resistance genes of growing clinical concern, including beta-lactamases, aminoglycoside modification enzymes, and key genes mediating resistance against quinolones, sulfonamides, tetracyclines and trimethoprims. These genes were all predicted to be localized onto

plasmid sequences, indicating that these wastewater strains could potentially transfer these resistance determinants to other microbes present within the wastewater matrix. Remarkably, several microbial defense and disinfection-related stress resistance genes (i.e., against oxidative stress, heat, heavy metals, etc.) were also predicted to be plasmid-localized, suggesting that these wastewater-adaptive genes may also be mobilizable within the wastewater environment. As such, this evidence suggests that the naturalized WWS-*E. coli* strains could play an important role in driving the co-evolution of antibiotic and wastewater treatment resistance in the wastewater microbiome (Yu *et al.* 2022).

### 7.1.4 The Emergence of Treatment Resistance in Wastewater-Borne ExPEC

The existence of a wastewater-specific, potentially disinfection-resistant, *E. coli* ecotype implies that other microbes could be following a similar evolutionary path towards treatment resistance. In particular, studies have demonstrated that *E. coli* strains belonging to the ExPEC pathotypes appear to be differentially surviving across the wastewater treatment train (Paulshus *et al.* 2019). Indeed, previous work by Zhi *et al.* (2020) revealed that a significant proportion of the surviving *E. coli* population following sewage chlorination and tertiary treatment were virtually identical genomically to a clinical UPEC counterpart, with a select number also representing multidrug-resistant, ESBL-producing ST131:O25b strains. Expanding on these studies, the findings presented in Chapter Six of this thesis demonstrate that these treatment-resistant, wastewater-borne *E. coli* strains also shared a similarly high degree of similarity with a clinical NMEC or BBEC counterpart at the core, whole, and accessory genome level. As such, beyond UPEC, clinically-relevant meningitic and septicaemic *E. coli* strains also appear to be differentially surviving wastewater treatment.

Interestingly, the accessory genome analyses performed in Chapter Six revealed several similarities between the WWS and W-ExPEC strains that could underlie their shared persistence across the wastewater treatment train. On the basis of accessory gene content, these two groups (i.e., especially the WWS and ST10 and ST44 ExPEC strains) were found to cluster closely, suggesting they may share similar genetic adaptations promoting their survival within the wastewater environment. Reflecting this, previous work conducted by Zhi *et al.* (2019) revealed that the WWS-*E coli* strains possessed an abundance of UPEC-associated virulence genes, suggesting they may play dual roles in mediating pathogenesis in ExPEC populations while also promoting survival within non-host, environmental niches. Beyond these shared virulence determinants, the close accessory genome clustering of these two groups could also reflect the potential role of the WWS-*E. coli* strains in transferring additional wastewater-adaptive genes to their W-ExPEC counterparts, such as implied by the gene localization analyses performed in Chapter Four. Importantly, although the exact nature of this relationship may be unclear, these findings demonstrate that the selective pressures imposed by the wastewater treatment plant environment can lead to the development of wastewater treatment resistance across several distinct *E. coli* ecotypic groups.

## 7.2 Limitations and Future Research

The findings presented in this thesis have highlighted the advantages of our polyphasic, logic regression-based approach for understanding the genotypic, phenotypic and ecotypic mechanisms underlying the niche-specificity, and potential phenotypic convergence (i.e., wastewater treatment resistance), of *E. coli* ecotypes. Despite the strengths in our approach, we

have also identified several key areas of improvement. These limitations, as well as potential directions for future research, are discussed below.

### 7.2.1 Refining the Logic Regression Workflow

Throughout this thesis we were able to make several improvements to the logic regression workflow for microbial niche-specificity analyses that was first established by Zhi *et al.* (2015; 2016a; 2016b). One of the major advancements involved adopting a more 'iterative' approach to the logic modelling process, which included optimizing the model parameters for each ITGR target and ecotype, evaluating the performance of biomarkers across different permutations of concatenated ITGR sequences, and pooling the results of several classification trials for source attribution analyses. These changes were critical for improving biomarker performance and providing a sufficient degree of consensus for classification purposes; however, they also drastically increased the complexity and workload of the analyses involved, making our approach more time- and labour-intensive and less user-friendly overall. Although this is a common limitation for supervised learning algorithms (Lupolova *et al*. 2019), additional adjustments can be made to simplify our logic regression workflow.

For instance, our custom logic regression script can be further refined and linearized in order to automate all iterative model building and classification steps. Alternatively, the logic regression algorithm itself can be improved, such as through the incorporation of decision trees within the logic regression approach via logicDT (Lau *et al.* 2024). While logicDT retains the use of Boolean combinations of input variables (i.e., biologically plausible SNP-SNP interactions) for model building, it also has additional strengths over logic regression. For instance, while we had to manually perform the iterative model building steps with logic regression, logicDT

automatically attempts to find the best performing model parameters during the model building process. Furthermore, unlike logic regression, logicDT allows the user to estimate the importance of all identified variables and their potential combinations, which can help improve the interpretability of the results through identifying the specific SNPs and interactions that are most important and thus ecotype-informative in each biomarker. Adjusting the methodology around logicDT, therefore, might help to refine the workflow by reducing its complexity, while simultaneously improving the performance and interpretability of the generated results.

### 7.2.2 Limited Number of Naturalized-Engineered *E. coli* Strains

Although we have been able to provide comprehensive evidence for the existence of distinct naturalized-engineered *E. coli* ecotypes, it is important to note the limited number of strains that were analysed in this thesis. Overall, only 20 WWS- and 17 MPS-*E. coli* strains were included in our genotypic analyses in Chapters Three and Four, while only 4 WWS-*E. coli* strains were phenotypically characterized in Chapter Five. Unfortunately, with such small sample sizes it can be difficult to ascertain whether our results (i.e., the adaptive traits identified) might apply universally to all *E. coli* strains belonging to these naturalized-engineered ecotypes, or just to the strains that we were able to evaluate in each chapter. This is an especially important limitation for the meat plant-derived strains as, unlike the wastewater strains which were found to exhibit a worldwide distribution in Chapter Three, all MPS strains analysed in this thesis were collected from a single meat processing plant in Alberta, Canada. The degree of clonal representation was thus likely higher for these meat plant strains, which calls into question their designation as a distinct ecotypic group (i.e., as opposed to several strains of a single clonal lineage) – especially given our finding that strains belonging to the same ecotypic group may still exhibit significant

variability in their individual genotypic and phenotypic traits.

As such, future efforts should be focused on building a more comprehensive library of *E. coli* isolates derived from engineered environments. Although hundreds of thousands of *E. coli* genomes have been sequenced to date on NCBI, an overwhelming majority of these entries are biased towards clinical or human and animal-associated strains. Thus, isolating and sequencing additional ST635 and ST399 *E. coli* strains from wastewater treatment plants (i.e., wastewater effluents, sewage, and sludge) and meat processing facilities, but also other engineered sources such as septic tanks (Behruznia *et al.* 2022a; Behruznia *et al.* 2022b) and drainage systems (Constantinides *et al.* 2020), will help to bolster the representation of these naturalized-engineered *E. coli* strains in public databases. Once collected, these additional naturalized-engineered strains can then be incorporated into additional analyses to validate our findings. In particular, re-analysing these naturalized-engineered strains against their host-associated and environmental counterparts, such as through the typing analyses in Chapter Three, the pan-genome-wide-association studies in Chapter Four, and the phenotypic assessments in Chapter Five, will help to reinforce the conclusions presented in this thesis. Performing these additional analyses will also help answer a central question raised in this thesis – whether *E. coli* strains derived from disparate man-made environments might collectively constitute a single naturalized-engineered ecotype, or several distinct, niche-specific ecotypic groups.

### 7.2.3 Need for Further Phenotypic and Ecotypic Validation of Findings

One of the central tenets of this thesis was the importance of the polyphasic approach to understanding bacterial evolution and diversity. While we were able to apply a polyphasic lens to our analyses by recapitulating some of the ecotypic and genotypic adaptations within the WWS

strains with observable phenotypic traits (i.e., enhanced biofilm formation, heat resistance, etc.), the bulk our analyses were *in-silico* (i.e., bioinformatic) in nature and thus could still benefit from further phenotypic and ecotypic validation. For instance, although the work presented in this thesis and by others (Zhi *et al.* 2016a; Zhi *et al.* 2017; Zhi *et al.* 2019) have provided comprehensive evidence for a WWS-*E. coli* ecotype, the supposed niche-specificity of these strains have yet to be demonstrated ecotypically with a model system. Thus, future studies should seek to validate these hypotheses by comparing the ability of naturalized WWS- and host-derived *E. coli* strains in colonizing and/or surviving within a host system (i.e., *E. coli*-free mouse models as previously described by Ju *et al.* [2017]) and in wastewater.

Similarly, future studies could also aim to further reinforce the major genotypic findings obtained in each chapter of this thesis using phenotypic or ecotypic means. For example, in Chapter Two we were able to identify ecotype-informative biomarkers within ITGRs across the *E. coli* genome for a wide variety of host animal sources. Mechanistically, these SNP-SNP biomarkers were thought to be indicative of a given strain's regulatory capacity in sensing and responding to the specific gastrointestinal conditions of a given host species. Although these host-informative biomarkers were indirectly validated through the source attribution of unknown environmental water *E. coli* isolates, their ecotypic relevance can also be directly assessed through other methods. For instance, as described above, competitive colonization studies can be conducted in animal models (i.e., *E. coli*-free mice and rats) to evaluate whether *E. coli* strains harboring different host-specific biomarkers (i.e., mouse vs. rat) are restricted to, or alternatively preferentially colonize, their corresponding host species. As further evidence for the adaptive value of these ITGR-encoded biomarkers, the ITGR sequences of different host-adapted strains can also be targeted for exchange using similar techniques leveraged for mutagenic and complementation studies (Kobras

*et al.* 2021), and then subsequently validated again using competitive colonization experiments.

In Chapters Three and Four, we were able to identify several genetic determinants (i.e., insertion element markers, genes, etc.) within the WWS- and MPS-*E. coli* strains that appeared to underlie their adaptation to their corresponding engineered niches. While some of these genotypic features were reinforced through observable phenotypic traits in Chapter Five (i.e., heat resistance, biofilm formation capacity), additional studies may be performed to further characterize the phenotypic adaptations harbored by the wastewater and meat plant strains. For instance, comparative gene expression studies via qPCR, such as described by Zhi *et al.* (2017), can be performed to evaluate the functional relevance of intergenic insertions, while targeted phenotypic assays and complementation or gene editing studies, such as described by Kobras *et al.* (2021), can be conducted to assess the impact of intragenic insertions. Furthermore, targeted phenotypic assays can also be performed to validate the pan-genomic findings described in Chapter Four, including heavy metal resistance assays (Azam *et al.* 2018) as well as phage resistance panels (McGee *et al.* 2023) and prediction modelling (Smug *et al.* 2022) for the WWS strains, and chlorine resistance assays (Wang *et al.* 2020) for the MPS strains.

Finally, the evidence presented in Chapter Six indicated that a significant proportion of *E. coli* strains surviving wastewater treatment may represent clinically relevant ExPEC. Although these conclusions were made based on the finding that several W-ExPEC strains were virtually identical to a clinical ExPEC representative using multiple genomic metrics, additional studies can be performed to verify the pathogenic capacity of these wastewater strains. Indeed, infection studies with these candidate W-ExPEC strains can be conducted using mouse models, to assess the ability of these strains to cause extraintestinal diseases such as urinary tract infections (Frick-Cheng *et al.* 2020), meningitis (Su *et al.* 2023), or bacteraemia (Landraud *et al.* 2013). Importantly,

validating the pathogenic capacity of these wastewater-borne strains will provide valuable insight for evaluating the public health risks associated with exposure to environments impacted by wastewater and/or sewage run-off.

### 7.2.4 Relationship Between WWS-*E. coli* and W-ExPEC Strains

Although they may represent distinct ecotypic groups, the WWS- and W-ExPEC strains both appear to be capable of differentially surviving wastewater treatment. Interestingly, the evidence presented in this thesis suggested that the WWS strains may be playing an important role in the emergence of this phenotype within the W-ExPEC strains. Indeed, the plasmid localization of treatment-related resistance genes within the WWS-*E. coli* strains described in Chapter Four, and the close accessory genome clustering of the two groups in Chapter Six, suggest that the WWS-*E. coli* strains may be capable of transferring key treatment resistance-related determinants to the W-ExPEC strains. Unfortunately, these analyses only provide circumstantial evidence for this relationship – thus, additional analyses can be performed to further explore this possibility.

For instance, similar to that performed for the WWS-*E. coli* strains in Chapter Four, plasmid screening and gene localization studies can be conducted for the W-ExPEC strains as an initial assessment into whether strains belonging to these two ecotypic groups might share similar plasmid sequences and plasmid-derived genetic repertoires. Furthermore, to directly assess whether these strains can actively participate in the horizontal exchange of genetic material, co-incubation and plasmid transfer assays can be performed. As described by Darphorn *et al.* (2022), individual WWS- and W-ExPEC strains can be co-incubated to assess the degree of genetic trafficking that can occur between these strains. These strains can then be exposed to treatment-related stressors (i.e., chlorine, oxidative stress) during co-incubation to assess the impact of

wastewater treatment on plasmid transfer dynamics (i.e., whether wastewater treatment promotes increased HGT) and its potential role in the evolution and dissemination of resistance within the wastewater environment.

## 7.3 Implications of the Research

The findings presented in this thesis have important implications for various fields of study, including theoretical microbiology, bacterial systematics, and public health. Conceptually, the demonstration and validation of *E. coli* niche-specificity provides valuable insight clarifying the prevalence of niche-generalism versus niche-specialism within the *E. coli* species. Niche-generalism and specialism represent two distinct evolutionary strategies that can each maximize the evolutionary success of a given microbial species. While *E. coli* is conventionally understood to be a niche-generalist capable of colonizing and transmitting across its various host species and environmental niches, the work presented in this thesis and by others (Zhi *et al.* 2015; Zhi *et al.* 2016b) demonstrate that a significant proportion of the *E. coli* species may be restricted to their present niche. Indeed, according to logic regression modelling, a large percentage of *E. coli* isolates (i.e., up to 70.00% based on our analyses and 92.00% based on previous work [Zhi *et al.* 2016b], depending on the specific source and ITGR sequences) appear to harbor niche-specific biomarkers and thus could represent niche-specialists. Interestingly, based on these results, the number of specialists may even out-number their niche-generalist counterparts within any given niche.

Importantly, while this suggests that niche-specialism may be the dominant evolutionary strategy for most *E. coli* strains, our findings should not discount the importance of niche-generalism within the species. Reflecting this, the logic regression analyses also demonstrated that a proportion of the *E. coli* isolates derived from any given source may not harbor niche-informative

biomarkers, and thus could represent potential niche-generalists. Similarly, when applied for source attribution purposes, logic regression was found to consistently classify select environmental isolates as multi-host (i.e., 'Beaver | Human | Reindeer') strains. Given that *E. coli* is widely recognized as an important zoonotic pathogen (Bélanger *et al.* 2011; Garcia *et al.* 2010), niche-generalism still appears to represent a successful evolutionary strategy within the species. Considering this, the prevailing discourse pitting niche-generalism against niche-specialism may not accurately reflect the evolutionary dynamics of the *E. coli* species. Rather, generalism and specialism likely represent two extremes amongst a wide range of different evolutionary strategies (von Meijenfeldt *et al.* 2023). Individual strains, therefore, may adopt certain generalist or specialist characteristics depending on their particular niche conditions (i.e., degree of trafficking between multiple niches, how distinct the conditions of a given niche are, etc.) and adaptive capabilities (i.e., ability to persist during trafficking across niches, competitiveness against other con-specifics within a given niche, etc.). As such, the labelling of the *E. coli* species simply as niche-generalists or niche-specialists may be inaccurate – rather, the species likely consists of a mosaic of niche-specific ecotypes interspersed with niche-generalist populations that can mediate a sufficient degree of ecological and genetic trafficking to maintain the taxonomic coherence of the species.

The sheer degree of diversity in adaptive paths that can exist within a microbial species (as described above) highlights the importance of expanding current taxonomic schemes beyond just genotypic criteria. In our case, although it appears that *E. coli* can be characterized by a wide range of evolutionary strategies, we have provided strong evidence indicating that a significant proportion of the species can be segregated into distinct niche-defined ecotypes, using the WWS-*E. coli* ecotype as a model system. Currently, the diversification of the species into several niche-

specific ecotypic groups goes unrecognized given that all strains, regardless of their unique ecological characteristics and evolutionary histories, are collectively and simply referred to as '*E. coli*' *sensu lato*. We argue that additional criteria, including phenotype and (importantly) ecotype, should be included for taxonomic considerations as it provides additional information that can be used to resolve the differences that may exist between these ecologically distinct, ecotypic groups. Indeed, it has been argued that current taxonomic schemes that end at the species level may lead to the erroneous assumption that all distinct ecotypic members of a given species taxon are ecologically interchangeable (Cohan 2019). Thus, to address the 'bacterial species problem' (Bobay 2020) appropriately, ecotypic criteria can be incorporated into current taxonomic schemes, or at least recognized in nomenclature such as in the 'species-complex' model for *Borrelia burgforferi* (Rudenko *et al.* 2011) or the 'serovar' model for *Salmonella* (Singh 2013), as this will better reflect the wide ecotypic and evolutionary diversity that can exist between members of a given microbial species.

Beyond the important conceptual and theoretical impacts described above, our findings also raise concerning public health implications regarding the potential emergence of disinfection resistance within the microbial world. Although they may represent distinct ecotypic groups, both WWS-*E. coli* and W-ExPEC strains appear to have evolved a common capacity to tolerate and survive wastewater treatment. Concerningly, these results demonstrate that the phenotype of wastewater treatment resistance may not necessarily be restricted to naturalized (i.e., non-pathogenic), wastewater-specific microbial populations, but may extend to their pathogenic counterparts that are also present within the wastewater matrix. Indeed, the wastewater microbiome represents an incredibly diverse and rich microbial community upon which the same selection pressures can act, and several studies suggest that a common paradigm for the evolution

of treatment resistance may already exist for other pathogenic microbial species. For instance, Kelly *et al.* (2021) determined that potentially pathogenic *Acinetobacter* strains remained abundant in wastewater effluents following treatment. Similarly, using a 16S rRNA sequencing approach, Numberger *et al.* (2019) observed an increase in the relative abundance of *Legionella* in wastewater following treatment, suggesting that *Legionella* species may be particularly resilient against wastewater disinfection processes. A similar finding was also obtained by Cai and Zhang (2013), where *Mycobacterium tuberculosis* was found to be the most prevalent pathogen in sludge and effluents following treatment. Indeed, a wide variety of pathogenic genera, including *Roseomonas, Pseudomonas, Mycobacterium, Aeromonas, Legionella, Yersinia,* and *Escherichia*, have been found to be particularly abundant across all stages of wastewater treatment, even following disinfection (Osunmakinde *et al.* 2019). As such, reflecting the phenotypic convergence between the WWS- and W-ExPEC strains towards wastewater treatment resistance, our findings suggest that other microbes of growing clinical concern may be similarly evolving resistance to the wastewater treatment process. Future work, therefore, is urgently needed to assess the potential public health risks associated with the evolution of wastewater-borne pathogenic microbes against the disinfection and sanitation practices designed to eliminate them – otherwise, neglecting this could be to our peril.

# References

Abbott, S.L., O'Connor, J., Robin, T., Zimmer, B.L., and Janda, J.M. 2003. Biochemical properties of a newly described *Escherichia* species, *Escherichia albertii*. J. Clin. Microbiol. **41**(10): 4852–4854. doi:10.1128/JCM.41.10.4852-4854.2003.

Abdallah, M., Benoliel, C., Drider, D., Dhulster, P., and Chihib, N.E. 2014. Biofilm formation and persistence on abiotic surfaces in the context of food and medical environments. Arch. Microbiol. **196**(7): 453–472. doi:10.1007/s00203-014-0983-1.

Adefisoye, M.A., and Okoh, A.I. 2016. Identification and antimicrobial resistance prevalence of pathogenic *Escherichia coli* strains from treated wastewater effluents in Eastern Cape, South Africa. Microbiologyopen **5**(1): 143–151. doi:10.1002/mbo3.319.

Alcock, B.P., Raphenya, A.R., Lau, T.T.Y., Tsang, K.K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A.L. V., Cheng, A.A., Liu, S., Min, S.Y., Miroshnichenko, A., Tran, H.K., Werfalli, R.E., Nasir, J.A., Oloni, M., Speicher, D.J., Florescu, A., Singh, B., Faltyn, M., Hernandez-Koutoucheva, A., Sharma, A.N., Bordeleau, E., Pawlowski, A.C., Zubyk, H.L., Dooley, D., Griffiths, E., Maguire, F., Winsor, G.L., Beiko, R.G., Brinkman, F.S.L., Hsiao, W.W.L., Domselaar, G. V., and McArthur, A.G. 2020. CARD 2020: Antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res. **48**(D1): D517–D525. doi:10.1093/nar/gkz935.

Allen, K.P., Randolph, M.M., and Fleckenstein, J.M. 2006. Importance of heat-labile enterotoxin in colonization of the adult mouse small intestine by human enterotoxigenic *Escherichia coli* strains. Infect. Immun. **74**(2): 869–875. doi:10.1128/IAI.74.2.869-875.2006.

Alnajar, S., and Gupta, R.S. 2017. Phylogenomics and comparative genomic studies delineate six main clades within the family Enterobacteriaceae and support the reclassification of several polyphyletic members of the family. Infect. Genet. Evol. **54**: 108–127. Elsevier B.V. doi:10.1016/j.meegid.2017.06.024.

Aminov, R. I. 2011. Horizontal gene exchange in environmental microbiota. Front. Microbiol. **2**(July): 1-19. doi:10.3389/fmicb.2011.00158

Anastasi, E.M., Matthews, B., Gundogdu, A., Vollmerhausen, T.L., Ramos, N.L., Stratton, H., Ahmed, W., and Katouli, M. 2010. Prevalence and persistence of *Escherichia coli* strains with uropathogenic virulence characteristics in sewage treatment plants. Appl. Environ. Microbiol. **76**(17): 5882–5886. doi:10.1128/AEM.00141-10.

Anastasi, E.M., Wohlsen, T.D., Stratton, H.M., and Katouli, M. 2013. Survival of *Escherichia coli* in two sewage treatment plants using UV irradiation and chlorination for disinfection. Water Res. **47**(17): 6670–6679. Elsevier Ltd. doi:10.1016/j.watres.2013.09.008.

Azam, M., Jan, A.T., Kumar, A., Siddiqui, K., Mondal, A.H., and Haq, Q.M.R. 2018. Study of pandrug and heavy metal resistance among *E. coli* from anthropogenically influenced Delhi stretch of river Yamuna. Braz. J. Microbiol. **49**(3): 471–480. doi:10.1016/j.bjm.2017.11.001

Baldwin, T., Sakthianandeswaren, A., Curtis, J.M., Kumar, B., Smyth, G.K., Foote, S.J., and Handman, E. 2007. Wound healing response is a major contributor to the severity of cutaneous leishmaniasis in the ear model of infection. Parasite Immunol. **29**(10): 501–513. John Wiley & Sons, Ltd. doi:https://doi.org/10.1111/j.1365-3024.2007.00969.x.

Ballesté, E., Blanch, A.R., Muniesa, M., García-Aljaro, C., Rodríguez-Rubio, L., Martín-Díaz, J., Pascual-Benito, M., and Jofre, J. 2022. Bacteriophages in sewage: abundance, roles, and applications. FEMS Microbes **3**(March): 1–12. doi:10.1093/femsmc/xtac009.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M.A., and Pevzner, P.A. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. **19**(5): 455–477. doi:10.1089/cmb.2012.0021.

Banting, G.S., Braithwaite, S., Scott, C., Kim, J., Jeon, B., Ashbolt, N., Ruecker, N., Tymensen, L., Charest, J., Pintar, K., Checkley, S., and Neumanna, N.F. 2016. Evaluation of various *Campylobacter*-specific quantitative PCR (qPCR) assays for detection and enumeration of *Campylobacteraceae* in irrigation water and wastewater via a miniaturized most-probable-number-qPCR assay. Appl. Environ. Microbiol. **82**(15): 4743–4756. doi:10.1128/AEM.00077-16.

Baranova, N., and Nikaido, H. 2002. The BaeSR two-component regulatory system activates transcription of the *yegMNOB* (*mdtABCD*) transporter gene cluster in *Escherichia coli* and increases its resistance to novobiocin and deoxycholate. J. Bacteriol. **184**(15): 4168–4176. doi:10.1128/JB.184.15.4168-4176.2002.

Barnich, N., and Darfeuille-Michaud, A. 2007, January. Adherent-invasive *Escherichia coli* and Crohn's disease. Curr. Opin. Gastroen. **23**(1): 16–20 doi:10.1097/MOG.0b013e3280105a38.

Bastian, M., Heymann, S., and Jacomy, M. 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks Visualization and Exploration of Large Graphs. Proc., Int. AAAI Conf. Web Soc. Media: 361–362. Available from www.aaai.org.

Bateman, A., Martin, M.J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., Bely, B., Bingley, M., Bonilla, C., Britto, R., Bursteinas, B., Bye-AJee, H., Cowley, A., Da Silva, A., De Giorgi, M., Dogan, T., Fazzini, F., Castro, L.G., Figueira, L., Garmiri, P., Georghiou, G., Gonzalez, D., Hatton-Ellis, E., Li, W., Liu, W., Lopez, R., Luo, J., Lussi, Y., MacDougall, A., Nightingale, A., Palka, B., Pichler, K., Poggioli, D., Pundir, S., Pureza, L., Qi, G., Rosanoff, S., Saidi, R., Sawford, T., Shypitsyna, A., Speretta, E., Turner, E., Tyagi, N., Volynkin, V., Wardell, T., Warner, K., Watkins, X., Zaru, R., Zellner, H., Xenarios, I., Bougueleret, L., Bridge, A., Poux, S., Redaschi, N., Aimo, L., ArgoudPuy, G., Auchincloss, A., Axelsen, K., Bansal, P., Baratin, D., Blatter, M.C., Boeckmann, B., Bolleman, J., Boutet, E., Breuza, L., Casal-Casas, C., De Castro, E., Coudert, E., Cuche, B., Doche, M., Dornevil, D., Duvaud, S., Estreicher, A., Famiglietti, L., Feuermann, M., Gasteiger, E., Gehant, S., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Jungo, F., Keller, G., Lara, V., Lemercier, P., Lieberherr, D., Lombardot, T., Martin, X., Masson, P., Morgat, A., Neto, T., Nouspikel, N., Paesano, S., Pedruzzi, I., Pilbout, S., Pozzato, M., Pruess, M., Rivoire, C., Roechert, B., Schneider, M., Sigrist, C., Sonesson, K., Staehli, S., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Veuthey, A.L., Wu, C.H., Arighi, C.N., Arminski, L., Chen, C., Chen, Y., Garavelli, J.S., Huang, H., Laiho, K., McGarvey, P., Natale, D.A., Ross, K., Vinayaka, C.R., Wang, Q., Wang, Y., Yeh, L.S., and Zhang, J. 2017. UniProt: The universal protein knowledgebase. Nucleic Acids Res. **45**(D1): D158–D169. Oxford University Press. doi:10.1093/nar/gkw1099.

Baumdicker, F., Hess, W.R., and Pfaffelhuber, P. 2012. The infinitely many genes model for the distributed genome of bacteria. Genome Biol. Evol. **4**(4): 443–456. doi:10.1093/gbe/evs016.

Bäumler, A., and Fang, F.C. 2013. Host specificity of bacterial pathogens. Cold Spring Harb. Perspect. Med. **3**(12): 1–19. doi:10.1101/cshperspect.a010041.

Beghain, J., Bridier-Nahmias, A., Nagard, H. Le, Denamur, E., and Clermont, O. 2018. ClermonTyping: An easy-to-use and accurate in silico method for *Escherichia* genus strain phylotyping. Microb. Genom. **4**(7): 1–8. doi:10.1099/mgen.0.000192.

Behruznia, M., O'Brien, C.L., and Gordon, D.M. 2022a. Prevalence, diversity and genetic structure of *Escherichia coli* isolates from septic tanks. Environ. Microbiol. Rep. **14**(1): 138–146. United States. doi:10.1111/1758-2229.13035.

Behruznia, M., and Gordon, D.M. 2022b. Molecular and metabolic characteristics of wastewater associated Escherichia coli strains. Environ. Microbiol. Rep. **14**(4): 646–654. doi:10.1111/1758-2229.13076.

Bekal, S., Brousseau, R., Masson, L., Prefontaine, G., Fairbrother, J., and Harel, J. 2003. Rapid identification of *Escherichia coli* pathotypes by virulence gene detection with DNA microarrays. J. Clin. Microbiol. **41**(5): 2113–2125. doi:10.1128/JCM.41.5.2113-2125.2003.

Bélanger, L., Garenaux, A., Harel, J., Boulianne, M., Nadeau, E., and Dozois, C.M. 2011. *Escherichia coli* from animal reservoirs as a potential source of human extraintestinal

pathogenic *E. coli*. FEMS Immunol. Med. Microbiol. **62**(1): 1–10. doi:10.1111/j.1574-695X.2011.00797.x

Berglund, F., Ebmeyer, S., Kristiansson, E., and Larsson, D.G.J. 2023. Evidence for wastewaters as environments where mobile antibiotic resistance genes emerge. Commun. Biol. **6**(1): 1–11. Springer US. doi:10.1038/s42003-023-04676-7.

Berkvens, A., Chauhan, P., and Bruggeman, F.J. 2022. Integrative biology of persister cell formation: molecular circuitry, phenotypic diversification and fitness effects. J. R. Soc. Interface **19**(194). doi:10.1098/rsif.2022.0129.

Bertels, F., Silander, O.K., Pachkov, M., Rainey, P.B., and Van Nimwegen, E. 2014. Automated reconstruction of whole-genome phylogenies from short-sequence reads. Mol. Biol. Evol. **31**(5): 1077–1088. doi:10.1093/molbev/msu088.

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

Biselli, E., Schink, S.J., and Gerland, U. 2020. Slower growth of *Escherichia coli* leads to longer survival in carbon starvation due to a decrease in the maintenance rate . Mol. Syst. Biol. **16**(6): 1–13. doi:10.15252/msb.20209478.

Bisercic, M., Feutrier, J.Y., and Reeves, P.R. 1991. Nucleotide sequences of the gnd genes from nine natural isolates of *Escherichia coli*: Evidence of intragenic recombination as a

contributing factor in the evolution of the polymorphic *gnd* locus. J. Bacteriol. **173**(12): 3894–3900. doi:10.1128/jb.173.12.3894-3900.1991.

Blaak, H., De Kruijf, P., Hamidjaja, R.A., Van Hoek, A.H.A.M., De Roda Husman, A.M., and Schets, F.M. 2014. Prevalence and characteristics of ESBL-producing *E. coli* in Dutch recreational waters influenced by wastewater treatment plants. Vet. Microbiol. **171**(3–4): 448–459. Elsevier B.V. doi:10.1016/j.vetmic.2014.03.007.

Blyton, M.D.J., and Gordon, D.M. 2017. Genetic attributes of *E. coli* isolates from chlorinated drinking water. PLoS One, **12**(1): e0169445. doi:10.1371/journal.pone.0169445.

Blyton, M.D.J., Pi, H., Vangchhia, B., Abraham, S., Trott, D.J., Johnson, J.R., and Gordon, D.M. 2015. Genetic structure and antimicrobial resistance of *Escherichia coli* and cryptic clades in birds with diverse human associations. Appl. Environ. Microbiol. **81**(15): 5123–5133. doi:10.1128/AEM.00861-15.

Bobay, L.M. 2020. The prokaryotic species concept and challenges. *In* The Pangenome: Diversity, Dynamics and Evolution of Genomes. pp. 21–49. doi:doi.org/10.1007/978-3-030-38281-0_2

Bok, E., Mazurek, J., Myc, A., Stosik, M., Wojciech, M., and Baldy-Chudzik, K. 2018. Comparison of commensal *Escherichia coli* isolates from adults and young children in Lubuskie province, Poland: Virulence potential, phylogeny and antimicrobial resistance. Int. J. Environ. Res. Public Health **15**(4): 1–19. doi:10.3390/ijerph15040617.

Bolger, A.M., Lohse, M., and Usadel, B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics **30**(15): 2114–2120. doi:10.1093/bioinformatics/btu170.

Brennan, F.P., Abram, F., Chinalia, F.A., Richards, K.G., and O'Flaherty, V. 2010. Characterization of environmentally persistent *Escherichia coli* isolates leached from an Irish soil. Appl. Environ. Microbiol. **76**(7): 2175–2180. doi:10.1128/AEM.01944-09.

Brennan, F.P., Grant, J., Botting, C.H., O'Flaherty, V., Richards, K.G., and Abram, F. 2013. Insights into the low-temperature adaptation and nutritional flexibility of a soil-persistent *Escherichia coli*. FEMS Microbiol. Ecol. **84**(1): 75–85. doi:10.1111/1574-6941.12038.

Brenner, D.J., Davis, B.R., Steigerwalt, A.G., Riddle, C.F., McWhorter, A.C., Allen, S.D., Farmer, J.J., Saitoh, Y., and Fanning, G.R. 1982b. Atypical biogroups of *Escherichia coli* found in clinical specimens and description of *Escherichia hermannii* sp. nov. J. Clin. Microbiol. **15**(4): 703–713. doi:10.1128/jcm.15.4.703-713.1982.

Brenner, D.J., McWhorter, A.C., Knutson, K.L., and Steigerwalt, A.G. 1982a. *Escherichia vulneris*: A new species of Enterobacteriacea associated with human wounds. Journal of Clinical Microbiology, **15**(6): 1133–1140. doi:10.1128/jcm.15.6.1133-1140.1982.

Brown, S.P., Cornforth, D.M., and Mideo, N. 2012. Evolution of virulence in opportunistic pathogens: Generalism, plasticity, and control. Trends Microbiol. **20**(7): 336–342. Elsevier Ltd. doi:10.1016/j.tim.2012.04.005.

Brynildsrud, O., Bohlin, J., Scheffer, L., and Eldholm, V. 2016. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol. **17**(1): 1–9. Genome Biology. doi:10.1186/s13059-016-1108-8.

Burgess, N.R.H., McDermott, S.N., and Whiting, J. 1973. Aerobic bacteria occurring in the hind-gut of the cockroach, *Blatta orientalis*. J. Hyg. **71**(1): 1–8. doi:10.1017/S0022172400046155.

Byappanahalli, M.N., Yan, T., Hamilton, M.J., Ishii, S., Fujioka, R.S., Whitman, R.L., and Sadowsky, M.J. 2012. The population structure of *Escherichia coli* isolated from subtropical and temperate soils. Sci. Total Environ. **417–418**: 273–279. Elsevier B.V. doi:10.1016/j.scitotenv.2011.12.041.

Cai, L., and Zhang, T. 2013. Detecting human bacterial pathogens in wastewater treatment plants by a high-throughput shotgun sequencing technique. Environ. Sci. Technol. **47**(10): 5433–5441. doi:10.1021/es400275r.

Calhau, V., Mendes, C., Pena, A., Mendonça, N., and Da Silva, G.J. 2015. Virulence and plasmidic resistance determinants of *Escherichia coli* isolated from municipal and hospital wastewater treatment plants. J. Water Health **13**(2): 311–318. doi:10.2166/wh.2014.327.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. 2009. BLAST+: Architecture and applications. BMC Bioinformatics **10**: 1–9. doi:10.1186/1471-2105-10-421.

Carattoli, A., and Hasman, H. 2020. PlasmidFinder and In Silico pMLST: Identification and Typing of Plasmid Replicons in Whole-Genome Sequencing (WGS). *In* Horizontal Gene Transfer: Methods and Protocols. pp. 285–294. doi:10.5040/9781474265126.0025.

Carlos, C., Pires, M.M., Stoppe, N.C., Hachich, E.M., Sato, M.I.Z., Gomes, T.A.T., Amaral, L.A., and 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 Microbiol. **10**: 161 doi:10.1186/1471-2180-10-161.

Chahal, C., van den Akker, B., Young, F., Franco, C., Blackbeard, J., and Monis, P. 2016. Pathogen and Particle Associations in Wastewater: Significance and Implications for Treatment and Disinfection Processes. *In* Advances in Applied Microbiology. Elsevier Ltd. doi:10.1016/bs.aambs.2016.08.001.

Chattopadhyay, I., J, R.B., Usman, T.M.M., and Varjani, S. 2022. Exploring the role of microbial biofilm for industrial effluents treatment. Bioengineered **13**(3): 6420–6440. Taylor & Francis. doi:10.1080/21655979.2022.2044250.

Chaudhuri, R.R., and Henderson, I.R. 2012. The evolution of the *Escherichia coli* phylogeny. Infect. Genet. Evol. **12**(2): 214–226. Elsevier B.V. doi:10.1016/j.meegid.2012.01.005.

Chen, L., Zheng, D., Liu, B., Yang, J., and Jin, Q. 2016. VFDB 2016: Hierarchical and refined dataset for big data analysis - 10 years on. Nucleic Acids Res. **44**(D1): D694–D697. doi:10.1093/nar/gkv1239.

Chérifi, A., Contrepois, M., Picard, B., Goullet, P., Ørskov, I., and Orskov, F. 1994. Clonal relationships among *Escherichia coli* serogroup O78 isolates from human and animal infections. J. Clin. Microbiol. **32**(5): 1197–1202. doi:10.1128/jcm.32.5.1197-1202.1994.

Choi, U., and Lee, C.R. 2019. Distinct Roles of Outer Membrane Porins in Antibiotic Resistance and Membrane Integrity in *Escherichia coli*. Front. Microbiol. **10**(April): 1–9. doi:10.3389/fmicb.2019.00953.

Cleaveland, S., Laurenson, M.K., and Taylor, L.H. 2001. Diseases of humans and their domestic mammals: Pathogen characteristics, host range and the risk of emergence. Philos. Trans. R Soc. Lond. B Biol. Sci., **356**(1411): 991–999. doi:10.1098/rstb.2001.0889.

Clermont, O., Lescat, M., O'Brien, C.L., Gordon, D.M., Tenaillon, O., and Denamur, E. 2008. Evidence for a human-specific *Escherichia coli* clone. Environ. Microbiol. **10**(4): 1000–1006. doi:10.1111/j.1462-2920.2007.01520.x.

Clermont, O., Dhanji, H., Upton, M., Gibreel, T., Fox, A., Boyd, D., Mulvey, M.R., Nordmann, P., Ruppé, E., Sarthou, J.L., Frank, T., Vimont, S., Arlet, G., Branger, C., Woodford, N., and Denamur, E. 2009. Rapid detection of the O25b-ST131 clone of *Escherichia coli* encompassing the CTX-M-15-producing strains. J. Antimicrob. Chemother. **64**(2): 274–277. doi:10.1093/jac/dkp194.

Clermont, O., Gordon, D.M., Brisse, S., Walk, S.T., and Denamur, E. 2011a. Characterization of the cryptic *Escherichia* lineages: Rapid identification and prevalence. Environ. Microbiol. **13**(9): 2468–2477. doi:10.1111/j.1462-2920.2011.02519.x.

Clermont, O., Olier, M., Hoede, C., Diancourt, L., Brisse, S., Keroudean, M., Glodt, J., Picard, B., Oswald, E., and Denamur, E. 2011b. Animal and human pathogenic *Escherichia coli* strains share common genetic backgrounds. Infect. Genet. Evol. **11**(3): 654–662. doi:10.1016/j.meegid.2011.02.005.

Clermont, O., Christenson, J.K., Denamur, E., and Gordon, D.M. 2013. The Clermont *Escherichia coli* phylo-typing method revisited: Improvement of specificity and detection of new phylo-groups. Environ. Microbiol. Rep. **5**(1): 58–65. doi:10.1111/1758-2229.12019.

Clermont, O., Dixit, O.V.A., Vangchhia, B., Condamine, B., Dion, S., Bridier-nahmias, A., Denamur, E., and Gordon, D. 2019. Characterization and rapid identification of phylogroup G in *Escherichia coli*, a lineage with high virulence and antibiotic resistance potential. Environ. Microbiol. **21**: 3107–3117. doi:10.1111/1462-2920.14713.

Cohan, F.M. 2019. Systematics: The Cohesive Nature of Bacterial Species Taxa. Curr. Biol. 29(5): 169–172. Cell Press. doi:10.1016/j.cub.2019.01.033

Cohan, F.M., and Kopac, S.M. 2011. Microbial genomics: *E. coli* relatives out of doors and out of body. Curr. Biol. **21**(15): R587–R589. Elsevier Ltd. doi:10.1016/j.cub.2011.06.011.

Cohan, F.M., and Perry, E.B. 2007. A Systematics for Discovering the Fundamental Units of Bacterial Diversity. Curr. Biol. **17**(10): 373–386. doi:10.1016/j.cub.2007.03.032.

Collins, R.E., and Higgs, P.G. 2012. Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. Mol. Biol. Evol. **29**(11): 3413–3425. doi:10.1093/molbev/mss163.

Constantinides, B., Chau, K.K., Phuong Quan, T., Rodger, G., Andersson, M.I., Jeffery, K., Lipworth, S., Gweon, H.S., Peniket, A., Pike, G., Millo, J., Byukusenge, M., Holdaway, M., Gibbons, C., Mathers, A.J., Crook, D.W., Peto, T.E.A., Sarah Walker, A., and Stoesser, N. 2020. Genomic surveillance of *Escherichia coli* and *Klebsiella* spp. in hospital sink drains and patients. Microb. Genomics **6**(7): 4–16. doi:10.1099/mgen.0.000391.

Consuegra, J., Gaffé, J., Lenski, R.E., Hindré, T., Barrick, J.E., Tenaillon, O., and Schneider, D. 2021. Insertion-sequence-mediated mutations both promote and constrain evolvability during a long-term experiment with bacteria. Nat. Commun. **12**(1). Springer US. doi:10.1038/s41467-021-21210-7.

Conway, T., and Cohen, P.S. 2015. Commensal and Pathogenic *Escherichia coli* Metabolism in the Gut. Microbiol. Spectr. **3**(3). doi:10.1128/microbiolspec.mbp-0006-2014.

Cueva, C., Moreno-Arribas, M.V., Martín-Álvarez, P.J., Bills, G., Vicente, M.F., Basilio, A., Rivas, C.L., Requena, T., Rodríguez, J.M., and Bartolomé, B. 2010. Antimicrobial activity of phenolic acids against commensal, probiotic and pathogenic bacteria. Res. Microbiol. **161**(5): 372–382. doi:10.1016/j.resmic.2010.04.006.

Cutler, D., and Miller, G. 2005. The role of public health improvements in health advances: The twentieth-century United States. Demography **42**(1): 1–22. doi:10.1353/dem.2005.0002.

Cydzik-Kwiatkowska, A., and Zielińska, M. 2016. Bacterial communities in full-scale wastewater treatment systems. World J. Microbiol. Biotechnol. **32**(4): 1–8. doi:10.1007/s11274-016-2012-9.

Daer, S., Goodwill, J.E., and Ikuma, K. 2021. Effect of ferrate and monochloramine disinfection on the physiological and transcriptomic response of *Escherichia coli* at late stationary phase. Water Res. **189**: 116580. Elsevier Ltd. doi:10.1016/j.watres.2020.116580.

Daer, S., Rehmann, E., Rehmann, J., and Ikuma, K. 2022. Development of Resistance in *Escherichia coli* Against Repeated Water Disinfection. Front. Environ. Sci. **10**(March): 1–12. doi:10.3389/fenvs.2022.855224.

Darphorn, T.S., Koenders-Van Sintanneland, B.B., Grootemaat, A.E., van der Wel, N.N., Brul, S., and ter Kuile, B.H. 2022. Transfer dynamics of multi-resistance plasmids in *Escherichia coli i*solated from meat. PLoS ONE. **17**(7): 1–14. doi:10.1371/journal.pone.0270205

Denamur, E., Clermont, O., Bonacorsi, S., and Gordon, D. 2021. The population genetics of pathogenic *Escherichia coli*. Nat. Rev. Microbiol. **19**(1): 37–54. Springer US. doi:10.1038/s41579-020-0416-x.

Deschamps, C., Clermont, O., Hipeaux, M.C., Arlet, G., Denamur, E., and Branger, C. 2009. Multiple acquisitions of CTX-M plasmids in the rare D2 genotype of *Escherichia coli* provide evidence for convergent evolution. Microbiology, **155**(5): 1656–1668. Microbiology Society. doi:https://doi.org/10.1099/mic.0.023234-0.

Desjardins, P., Picard, B., Kaltenböck, B., Elion, J., and Denamurl, E. 1995. Sex in *Escherichia coli* does not disrupt the clonal structure of the population: evidence from random amplified polymorphic DNA and restriction-fragment-length polymorphism. J. Mol. Evol. **41**(4): 440–448. doi:10.1007/BF00160315.

Devane, M.L., Moriarty, E., Weaver, L., Cookson, A., and Gilpin, B. 2020. Fecal indicator bacteria from environmental sources; strategies for identification to improve water quality monitoring. Water Res. **185**: 116204. Elsevier Ltd. doi:10.1016/j.watres.2020.116204.

Devanga Ragupathi, N.K., Muthuirulandi Sethuvel, D.P., Inbanathan, F.Y., and Veeraraghavan, B. 2018. Accurate differentiation of *Escherichia coli* and *Shigella* serogroups: challenges and strategies. New Microbes New Infect. **21**: 58–62. Elsevier Ltd. doi:10.1016/j.nmni.2017.09.003.

Dillon, M.M., Thakur, S., Almeida, R.N.D., and Guttman, D.S. 2019. Recombination of ecologically and evolutionarily significant loci maintains genetic cohesion in the *Pseudomonas syringae* species complex. Genome Biol. **20**(3): 1–28. Genome Biology. doi:10.1186/s13059-018-1606-y.

Doranga, S., and Conway, T. 2023. *OmpC*-Dependent Bile Tolerance Contributes to *E. coli* Colonization of the Mammalian Intestine. Microbiol. Spectr. **11**(3): 1–19. American Society for Microbiology. doi:10.1128/spectrum.05241-22.

Dubreuil, J.D., Isaacson, R.E., and Schifferli, D.M. 2016. Animal enterotoxigenic *Escherichia coli*. EcoSal Plus, **7**(1). doi:10.1128/ecosalplus.ESP-0006-2016.ANIMAL.

Durrant, M.G., Li, M.M., Siranosian, B.A., Montgomery, S.B., and Bhatt, A.S. 2020. A Bioinformatic Analysis of Integrative Mobile Genetic Elements Highlights Their Role in Bacterial Adaptation. Cell Host Microbe **27**(1): 140-153.e9. Elsevier Inc. doi:10.1016/j.chom.2019.10.022.

Dykhuizen, D.E., and Green, L. 1991. Recombination in *Escherichia coli* and the definition of biological species. J. Bacteriol. **173**(22): 7257 LP – 7268. doi:10.1128/jb.173.22.7257-7268.1991.

Elso, C.M., Roberts, L.J., Smyth, G.K., Thomson, R.J., Baldwin, T.M., Foote, S.J., and Handman, E. 2004. Leishmaniasis host response loci (lmr1-3) modify disease severity through a Th1/Th2-independent pathway. Genes Immun. **5**(2): 93–100. doi:10.1038/sj.gene.6364042.

Eppinger, M., Mammel, M.K., LeClerc, J.E., Ravel, J., and Cebula, T.A. 2011. Genome signatures of *Escherichia coli* O157:H7 isolates from the bovine host reservoir. Appl. Environ. Microbiol. **77**(9): 2916–2925. doi:10.1128/AEM.02554-10.

Escherich, T. 1988. The Intestinal Bacteria of the Neonate and Breast-Fed Infant. Rev. Infect. Dis. **10**(6): 1220–1225. doi:10.1093/clinids/10.6.1220.

Escobar-Páramo, P., Clermont, O., Blanc-Potard, A.-B., Bui, H., Bouguénec, C., and Denamur, E. 2004. A Specific Genetic Background Is Required for Acquisition and Expression of Virulence Factors in *Escherichia coli*. Mol. Biol. Evol. **21**: 1085–1094. doi:10.1093/molbev/msh118.

Escobar-Páramo, P., Le Menac'h, A., Le Gall, T., Amorin, C., Gouriou, S., Picard, B., Skurnik, D., and Denamur, E. 2006. Identification of forces shaping the commensal *Escherichia coli* genetic structure by comparing animal and human isolates. Environ. Microbiol. **8**(11): 1975–1984. doi:10.1111/j.1462-2920.2006.01077.x.

Eswarappa, S.M., Janice, J., Nagarajan, A.G., Balasundaram, S. V., Karnam, G., Dixit, N.M., and Chakravortty, D. 2008. Differentially evolved genes of Salmonella pathogenicity islands: Insights into the mechanism of host specificity in *Salmonella*. PLoS One, **3**(12). doi:10.1371/journal.pone.0003829.

Ewing, W.H. 1949. *SHIGELLA* NOMENCLATURE. J. Bacteriol. **57**(6): 633–638.doi:10.1128/JB.57.6.633-638.1949.

Farmer, J.J., Fanning, G.R., Davis, B.R., O'Hara, C.M., Riddle, C., Hickman-Brenner, F.W., Asbury, M.A., Lowery, V.A., and Brenner, D.J. 1985. *Escherichia fergusonii* and *Enterobacter taylorae*, two new species of Enterobacteriaceae isolated from clinical specimens. J. Clin. Microbiol. **21**(1): 77–81. doi:10.1128/jcm.21.1.77-81.1985.

Fernández-Gómez, P., Trigal, E., Alegría, Á., Santos, J.A., López, M., Prieto, M., and Alvarez-Ordóñez, A. 2022. Biofilm formation ability and tolerance to food-associated stresses among ESBL-producing *Escherichia coli* strains from foods of animal origin and human patients. LWT - Food Sci. Technol. **168**. doi:10.1016/j.lwt.2022.113961.

Foley, S.L., Lynne, A.M., and Nayak, R. 2009. Molecular typing methodologies for microbial source tracking and epidemiological investigations of Gram-negative bacterial foodborne pathogens. Infect. Genet. Evol. **9**(4): 430–440. doi:https://doi.org/10.1016/j.meegid.2009.03.004.

Foster, D.M., and Smith, G.W. 2009. Pathophysiology of Diarrhea in Calves. Vet. Clin. North Am. Food Anim. Pract. **25**(1): 13–36. Elsevier Ltd. doi:10.1016/j.cvfa.2008.10.013.

Franz, E., Van Hoek, A.H.A.M., Van Der Wal, F.J., De Boer, A., Zwartkruis-Nahuis, A., Van Der Zwaluw, K., Aarts, H.J.M., and Heuvelinkd, A.E. 2012. Genetic features differentiating bovine, food, and human isolates of Shiga toxin-producing *Escherichia coli* O157 in The Netherlands. J. Clin. Microbiol. **50**(3): 772–780. doi:10.1128/JCM.05964-11.

Frick-Cheng, A.E., Sintsova, A., Smith, S.N., Krauthammer, M., Eaton, K.A., and Mobley, H.L.T. (2020). The gene expression profile of uropathogenic *Escherichia coli* in women with uncomplicated urinary tract infections is recapitulated in the mouse model. mBio. **11**(4): 1–16. doi:10.1128/mBio.01412-20

Fry, N.K., Savelkoul, P.H.M., and Visca, P. 2009. Amplified fragment-length polymorphism analysis. Methods Mol. Biol. **551**: 89–104. Respiratory and Systemic Infection Laboratory, Health Protection Agency, Centre for Infections, London, UK. doi:10.1007/978-1-60327-999-4_8.

Fu, L.L., and Li, J.R. 2014. Microbial Source Tracking: A Tool for Identifying Sources of Microbial Contamination in the Food Chain. Crit. Rev. Food Sci. Nutr. **54**(6): 699–707. doi:10.1080/10408398.2011.605231.

Garcia, A., Fox, J.G., and Besser, T.E. 2010. Zoonotic Enterohemorrhagic *Escherichia coli*: A One Health Perspective. ILAR J. **51**(3): 221–232. doi:10.1093/ilar.51.3.221

Garcia, E.C., Brumbaugh, A.R., and Mobley, H.L.T. 2011. Redundancy and specificity of *Escherichia coli* iron acquisition systems during urinary tract infection. Infect. Immun. **79**(3): 1225–1235. doi:10.1128/IAI.01222-10.

Garnett, J.A., Martínez-Santos, V.I., Saldaña, Z., Pape, T., Hawthorne, W., Chan, J., Simpson, P.J., Cota, E., Puente, J.L., Girón, J.A., and Matthews, S. 2012. Structural insights into the biogenesis and biofilm formation by the *Escherichia coli* common pilus. Proc. Natl. Acad. Sci. U. S. A. **109**(10): 3950–3955. doi:10.1073/pnas.1106733109.

Georgiades, K., and Raoult, D. 2011. Defining pathogenic bacterial species in the genomic era. Front. Microbiol. **1**(January): 1–13. doi:10.3389/fmicb.2010.00151.

Gerba, C.P., and Pepper, I.L. 2009. Wastewater Treatment and Biosolids Reuse. *In* Environmental Microbiology, Second Edition. Elsevier Inc. doi:10.1016/B978-0-12-370519-8.00024-9.

Geurtsen, J., de Been, M., Weerdenburg, E., Zomer, A., McNally, A., and Poolman, J. 2022. Genomics and pathotypes of the many faces of *Escherichia coli*. FEMS Microbiol. Rev. **46**(6): 1–30. doi:10.1093/femsre/fuac031.

Glaeser, S.P., and Kämpfer, P. 2015. Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. Syst. Appl. Microbiol. **38**(4): 237–245. Elsevier GmbH. doi:10.1016/j.syapm.2015.03.007.

Gomi, R., Matsuda, T., Matsui, Y., and Yoneda, M. 2014. Fecal source tracking in water by next-generation sequencing technologies using host-specific *Escherichia coli* genetic markers. Environ. Sci. Technol. **48**(16): 9616–9623. doi:10.1021/es501944c.

Goodswen, S.J., Barratt, J.L.N., Kennedy, P.J., Kaufer, A., Calarco, L., and Ellis, J.T. 2021. Machine learning and applications in microbiology. FEMS Microbiol. Rev. **45**(5): 1–19. Oxford University Press. doi:10.1093/femsre/fuab015.

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

Gordon, D.M., and Cowling, A. 2003. The distribution and genetic structure of *Escherichia coli* in Australian vertebrates: Host and geographic effects. Microbiology, **149**(12): 3575–3586. doi:10.1099/mic.0.26486-0.

van der Graaf-Van Bloois, L., Wagenaar, J.A., and Zomer, A.L. 2021. RFPlasmid: predicting plasmid sequences from short-read assembly data using machine learning. Microb. Genomics **7**(11): 1–11. doi:10.1099/mgen.0.000683.

Grant, J.R., Enns, E., Marinier, E., Mandal, A., Herman, E.K., Chen, C., Graham, M., Domselaar, G. Van, and Stothard, P. 2023. Proksee: in-depth characterization and visualization of bacterial genomes. Nucleic Acids Res.: 1–9. Oxford University Press.

Gray, M.J., Wholey, W.Y., and Jakob, U. 2013. Bacterial Responses to Reactive Chlorine Species. Annu. Rev. Microbiol. **67**(1): 141–160. doi:10.1146/annurev-micro-102912-142520.Bacterial.

Harwood, V.J., Staley, C., Badgley, B.D., Borges, K., and Korajkic, A. 2014. Microbial source tracking markers for detection of fecal contamination in environmental waters: Relationships

between pathogens and human health outcomes. FEMS Microbiol. Rev. **38**(1): 1–40. doi:10.1111/1574-6976.12031.

Hata, H., Natori, T., Mizuno, T., Kanazawa, I., Eldesouky, I., Hayashi, M., Miyata, M., Fukunaga, H., Ohji, S., Hosoyama, A., Aono, E., Yamazoe, A., Tsuchikane, K., Fujita, N., and Ezaki, T. 2016. Phylogenetics of family Enterobacteriaceae and proposal to reclassify *Escherichia hermannii* and *Salmonella subterranea* as *Atlantibacter hermannii* and *Atlantibacter subterranea* gen. nov., comb. nov. Microbiol. Immunol. **60**(5): 303–311. doi:10.1111/1348-0421.12374.

Hayashi, T., Makino, K., Ohnishi, M., Kurokawa, K., Ishii, K., Yokoyama, K., Han, C.-G., Ohtsubo, E., Nakayama, K., Murata, T., Tanaka, M., Tobe, T., Iida, T., Takami, H., Honda, T., Sasakawa, C., Ogasawara, N., Yasunaga, T., Kuhara, S., Shiba, T., Hattori, M., and Shinagawa, H. 2001. Complete Genome Sequence of Enterohemorrhagic *Escherichia coli* O157:H7 and Genomic Comparison with a Laboratory Strain K-12. DNA Res. **8**(1): 11–22. doi:10.1093/dnares/8.1.11.

Hendrickson, H. 2009. Order and disorder during *Escherichia coli* divergence. PLoS Genet. **5**(1): 1–2. doi:10.1371/journal.pgen.1000335.

Henry, R. 2015. Etymologia: *Escherichia coli*. Emerg. Infect. Dis. **21**(8): 1310. Centers for Disease Control and Prevention. doi:10.3201/eid2108.ET2108.

Ho, T.D., Davis, B.M., Ritchie, J.M., and Waldor, M.K. 2008. Type 2 secretion promotes enterohemorrhagic *Escherichia coli* adherence and intestinal colonization. Infect. Immun. **76**(5): 1858–1865. doi:10.1128/IAI.01688-07.

Hunter, P.R., MacDonald, A.M., and Carter, R.C. 2010. Water Supply and Health. PLoS Med. **7**(11): 1–9. doi:10.1371/journal.pmed.1000361.

Hutinel, M., Fick, J., Larsson, D.G.J., and Flach, C.F. 2021. Investigating the effects of municipal and hospital wastewaters on horizontal gene transfer. Environ. Pollut. **276**: 116733. Elsevier Ltd. doi:10.1016/j.envpol.2021.116733.

Hutton, T.A., Innes, G.K., Harel, J., Garneau, P., Cucchiara, A., Schifferli, D.M., and Rankin, S.C. 2018. Phylogroup and virulence gene association with clinical characteristics of *Escherichia coli* urinary tract infections from dogs and cats. J. Vet. Diagn. Invest. **30**(1): 64–70. doi:10.1177/1040638717729395.

Huys, G., Cnockaert, M., Janda, J.M., and Swings, J. 2003. *Escherichia albertii* sp. nov., a diarrhoeagenic species isolated from stool specimens of Bangladeshi children. Int. J. Syst. Evol. Microbiol. **53**(3): 807–810. doi:10.1099/ijs.0.02475-0.

Imlay, J.A. 2008. Cellular defenses against superoxide and hydrogen peroxide. Annu. Rev. Biochem. **77**(1): 755–776. doi:10.1146/annurev.biochem.77.061606.161055.Cellular.

Ingle, D.J., Clermont, O., Skurnik, D., Denamur, E., Walk, S.T., and Gordon, D.M. 2011. Biofilm formation by and thermal niche and virulence characteristics of *Escherichia* spp. Appl. Environ. Microbiol. **77**(8): 2695–2700. doi:10.1128/AEM.02401-10.

Ishii, S., Ksoll, W.B., Hicks, R.E., and Sadowsky, M.J. 2006. Presence and Growth of Naturalized *Escherichia coli* in Temperate Soils from Lake Superior Watersheds. Applied and Environ. Microbiol. **72**(1): 612 LP – 621. doi:10.1128/AEM.72.1.612-621.2006.

Ishii, S., Yan, T., Vu, H., Hansen, D.L., Hicks, R.E., and Sadowsky, M.J. 2010. Factors controlling long-term survival and growth of naturalized *Escherichia coli* populations in temperate field soils. Microbes Environ. **25**(1): 8–14. doi:10.1264/jsme2.ME09172.

Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., and Aluru, S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat. Commun. **9**(1): 1–8. doi:10.1038/s41467-018-07641-9.

Janda, J.M., and Abbott, S.L. 2002. Bacterial Identification for Publication: When Is Enough Enough? J. Clin. Microbiol. **40**(6): 1887–1891. doi:10.1128/JCM.40.6.1887.

Janda, J.M., Abbott, S.L., and Albert, M.J. 1999. Prototypal diarrheagenic strains of *Hafnia alvei* are actually members of the genus *Escherichia*. J. Clin. Microbiol. **37**(8): 2399–2401. doi:10.1128/jcm.37.8.2399-2401.1999.

Jang, J., Unno, T., Lee, S.W., Cho, K.H., Sadowsky, M.J., Ko, G., Kim, J.H., and Hur, H.G. 2011. Prevalence of season-specific *Escherichia coli* strains in the Yeongsan River Basin of South Korea. Environ. Microbiol. **13**(12): 3103–3113. doi:10.1111/j.1462-2920.2011.02541.x.

Jang, J., Di, D.Y.W., Han, D., Unno, T., Lee, J.H., Sadowsky, M.J., and Hur, H.G. 2015. Dynamic changes in the population structure of *Escherichia coli* in the Yeongsan River basin of South Korea. FEMS Microbiol. Ecol. **91**(11): 1–9. doi:10.1093/femsec/fiv127.

Jang, J., Hur, H.G., Sadowsky, M.J., Byappanahalli, M.N., Yan, T., and Ishii, S. 2017. Environmental *Escherichia coli*: ecology and public health implications—a review. J. Appl. Microbiol. **123**(3): 570–581. doi:10.1111/jam.13468.

Johnson, J.R., Ørskov, I., Ørskov, F., Goullet, P., Picard, B., Moseley, S.L., Roberts, P.L., and Stamm, W.E. 1994. O, K, and H Antigens Predict Virulence Factors, Carboxylesterase B Pattern, Antimicrobial Resistance, and Host Compromise among *Escherichia coli* Strains Causing Urosepsis. J. Infect. Dis. **169**(1): 119–126. doi:10.1093/infdis/169.1.119.

Johnson, J.R., and Stell, A.L. 2000. Extended Virulence Genotypes of *Escherichia coli* Strains from Patients with Urosepsis in Relation to Phylogeny and Host Compromise. J. Infect. Dis. **181**(1): 261–272. doi:10.1086/315217.

Johnson, J.R., and Russo, T.A. 2002. Extraintestinal pathogenic *Escherichia coli*: "The other bad *E. coli*." J. Lab. Clin. Med. **139**(3): 155–162. doi:10.1067/mlc.2002.121550.

Johnson, J.R., Clabots, C., and Kuskowski, M.A. 2008. Multiple-host sharing, long-term persistence, and virulence of *Escherichia coli* clones from human and animal household members. J. Clin. Microbiol. **46**(12): 4078–4082. doi:10.1128/JCM.00980-08.

Jolley, K.A., Bray, J.E., and Maiden, M.C.J. 2018. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res. **3**(0): 1–20. doi:10.12688/wellcomeopenres.14826.1.

Jørgensen, S.B., Søraas, A. V., Arnesen, L.S., Leegaard, T.M., Sundsfjord, A., and Jenum, P.A. 2017. A comparison of extended spectrum β-lactamase producing *Escherichia coli* from

clinical, recreational water and wastewater samples associated in time and location. PLoS One **12**(10): 1–15. doi:10.1371/journal.pone.0186576.

Ju, T., Shoblak, Y., Gao, Y., Yang, K., Fouhse, J., Finlay, B. B., So, Y. W., Stothard, P., and Willing, B.P. 2017. Initial gut microbial composition as a key factor driving host response to antibiotic treatment, as exemplified by the presence or absence of commensal *Escherichia coli*. Appl. Environ. Microbiol. **83**(17): 1–15. doi:10.1128/AEM.01107-17.

Kai, A., Konishi, N., and Obata, H. 2010. [Diarrheagenic *Escherichia coli*]. Nippon rinsho. Japanese Clinical Medicine, 68 Suppl **6**(1): 203–207. doi:10.1016/b978-012677530-3/50289-0.

Kämpfer, P. 2014. Continuing importance of the "Phenotype" in the genomic era. Methods in Microbiology, **41**: 307–320. doi:10.1016/bs.mim.2014.07.005.

Kämpfer, P., and Glaeser, S.P. 2012. Prokaryotic taxonomy in the sequencing era - the polyphasic approach revisited. Environ. Microbiol. **14**(2): 291–317. doi:10.1111/j.1462-2920.2011.02615.x.

Karkman, A., Do, T.T., Walsh, F., and Virta, M.P.J. 2018. Antibiotic-Resistance Genes in Waste Water. Trends Microbiol. **26**(3): 220–228. Elsevier Ltd. doi:10.1016/j.tim.2017.09.005.

Karp, P.D., Ong, W.K., Paley, S., Billington, R., Caspi, R., Fulcher, C., Kothari, A., Krummenacker, M., Latendresse, M., Midford, P.E., Subhraveti, P., Gama-Castro, S., Muñiz-Rascado, L., Bonavides-Martinez, C., Santos-Zavaleta, A., Mackie, A., Collado-Vides, J.,

Keseler, I.M., and Paulsen, I. 2018. The EcoCyc Database. EcoSal Plus **8**(1): 1–34. doi:10.1128/ecosalplus.esp-0006-2018.

Kassen, R. 2002. The experimental evolution of specialists, generalists, and the maintenance of diversity. J. Evol. Biol. **15**(2): 173–190. John Wiley & Sons, Ltd. doi:10.1046/j.1420-9101.2002.00377.x.

Kelly, J.J., London, M.G., McCormick, A.R., Rojas, M., Scott, J.W., and Hoellein, T.J. 2021. Wastewater treatment alters microbial colonization of microplastics. PLoS One **16**(1 January): 1–19. doi:10.1371/journal.pone.0244443.

Khatib, L.A., Tsai, Y.L., and Olson, B.H. 2003. A biomarker for the identification of swine fecal pollution in water, using the STII toxin gene from enterotoxigenic *Escherichia coli*. Appl. Microbiol. Biotechnol. **63**(2): 231–238. doi:10.1007/s00253-003-1373-9.

Khademi, S.M.H., Sazinas, P., and Jelsbak, L. 2019. Within-host adaptation mediated by intergenic evolution in *Pseudomonas aeruginosa*. Genome Biol. Evol. **11**(5): 1385–1397. doi:10.1093/gbe/evz083

Kim, J., Nietfeldt, J., and Benson, A.K. 1999. Octamer-based genome scanning distinguishes a unique subpopulation of *Escherichia coli* O157:H7 strains in cattle. Proc. Natl. Acad. Sci. U. S. A. **96**(23): 13288–13293. doi:10.1073/pnas.96.23.13288.

Kirzinger, M.W.B., and Stavrinides, J. 2012. Host specificity determinants as a genetic continuum. Trends Microbiol. **20**(2): 88–93. Elsevier Ltd. doi:10.1016/j.tim.2011.11.006.

Kobras, C.M., Fenton, A.K., and Sheppard, S.K. 2021. Next-generation microbiology: from comparative genomics to gene function. Genome Biol. **22**(1): 1–16. BioMed Central Ltd. doi:10.1186/s13059-021-02344-9

Koeppel, A., Perry, E.B., Sikorski, J., Krizanc, D., Warner, A., Ward, D.M., Rooney, A.P., Brambilla, E., Connor, N., Ratcliff, R.M., Nevo, E., and Cohan, F.M. 2008. Identifying the fundamental units of bacterial diversity: A paradigm shift to incorporate ecology into bacterial systematics. Proc. Natl. Acad. Sci. U. S. A. **105**(7): 2504–2509. doi:10.1073/pnas.0712205105.

Kondrashov, F.A. 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc. R. Soc. B Biol. Sci. **279**(1749): 5048–5057. doi:10.1098/rspb.2012.1108.

Konstantinidis, K.T., Ramette, A., and Tiedje, J.M. 2006. The bacterial species definition in the genomic era. Philos. Trans. R. Soc. Lond. B. Biol. Sci. **361**(1475): 1929–1940. doi:10.1098/rstb.2006.1920.

Korea, C.G., Badouraly, R., Prevost, M.C., Ghigo, J.M., and Beloin, C. 2010. *Escherichia coli* K-12 possesses multiple cryptic but functional chaperone-usher fimbriae with distinct surface specificities. Environ. Microbiol. **12**(7): 1957–1977. doi:10.1111/j.1462-2920.2010.02202.x.

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. 2018. MEGA X: Molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. **35**(6): 1547–1549. doi:10.1093/molbev/msy096.

Kunhikannan, S., Thomas, C.J., Franks, A.E., Mahadevaiah, S., Kumar, S., and Petrovski, S. 2021. Environmental hotspots for antibiotic resistance genes. Microbiologyopen **10**(3): 1–11. doi:10.1002/mbo3.1197.

Landini, P., Egli, T., Wolf, J., and Lacour, S. 2014. sigmaS, a major player in the response to environmental stresses in *Escherichia coli*: Role, regulation and mechanisms of promoter recognition. Environ. Microbiol. Rep. **6**(1): 1–13. doi:10.1111/1758-2229.12112.

Landraud, L., Jauréguy, F., Frapy, E., Guigon, G., Gouriou, S., Carbonnelle, E., Clermont, O., Denamur, E., Picard, B., Lemichez, E., Brisse, S., and Nassif, X. 2013. Severity of *Escherichia coli* bacteraemia is independent of the intrinsic virulence of the strains assessed in a mouse model. Clin. Microbiol. Infect. **19**(1): 85–90. doi:10.1111/j.1469-0691.2011.03750.x

Lau, M., Schikowski, T., and Schwender, H. 2024. logicDT: a procedure for identifying response-associated interactions between binary predictors. Mach. Learn, **113**(2): 933–992. Doi:10.1007/s10994-023-06488-6

Lawrence, J.G., and Ochman, H. 1997. Amelioration of Bacterial Genomes: Rates of Change and Exchange. J. Mol. Evol. **44**(4): 383–397. doi:10.1007/PL00006158.

Lawrence, J.G., Ochman, H., and Ragan, M.A. 2002. Reconciling the many faces of lateral gene transfer. Trends Microbiol. **10**(1): 1–4. doi:10.1016/S0966-842X(01)02282-X.

Le Gall, T., Clermont, O., Gouriou, S., Picard, B., Nassif, X., Denamur, E., and Tenaillon, O. 2007. Extraintestinal virulence is a coincidental by-product of commensalism in b2

phylogenetic group *Escherichia coli* strains. Mol. Biol. Evol. **24**(11): 2373–2384. doi:10.1093/molbev/msm172.

Leclerc, H. 1962. Biochemical study of pigmented Enterobacteriaceae. Annales de L'institut Pasteur (Paris). **102**: 726–741. Available from http://europepmc.org/abstract/MED/14463377.

Leggett, H.C., Buckling, A., Long, G.H., and Boots, M. 2013. Generalism and the evolution of parasite virulence. Trends Ecol. Evol. **28**(10): 592–596. doi:10.1016/j.tree.2013.07.002.

Li, B., Sun, J.Y., Han, L.Z., Huang, X.H., Fu, Q., and Ni, Y.X. 2010. Phylogenetic groups and pathogenicity island markers in fecal *Escherichia coli* isolates from asymptomatic humans in China. Appl. Environ. Microbiol. **76**(19): 6698–6700. doi:10.1128/AEM.00707-10.

Li, H., and Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**(14): 1754–1760. doi:10.1093/bioinformatics/btp324.

Liu, S., Jin, D., Lan, R., Wang, Y., Meng, Q., Dai, H., Lu, S., Hu, S., and Xu, J. 2015. *Escherichia marmotae* sp. nov., isolated from faeces of *Marmota himalayana*. Int. J. Syst. Evol. Microbiol. **65**(7): 2130–2134. doi:10.1099/ijs.0.000228.

Liu, S., Feng, J., Pu, J., Xu, X., Lu, S., Yang, J., Wang, Y., Jin, D., Du, X., Meng, X., Luo, X., Sun, H., Xiong, Y., Ye, C., Lan, R., and Xu, J. 2019. Genomic and molecular characterisation of *Escherichia marmotae* from wild rodents in Qinghai-Tibet plateau as a potential pathogen. Sci. Rep. **9**(1): 1–9. doi:10.1038/s41598-019-46831-3.

Llor, C., and Bjerrum, L. 2014. Antimicrobial resistance: Risk associated with antibiotic overuse and initiatives to reduce the problem. Ther. Adv. Drug Saf. **5**(6): 229–241. doi:10.1177/2042098614554919.

Lu, S., Jin, D., Wu, S., Yang, J., Lan, R., Bai, X., Liu, S., Meng, Q., Yuan, X., Zhou, J., Pu, J., Chen, Q., Dai, H., Hu, Y., Xiong, Y., Ye, C., and Xu, J. 2016. Insights into the evolution of pathogenicity of *Escherichia coli* from genomic analysis of intestinal *E. coli* of *Marmota himalayana* in Qinghai-Tibet plateau of China. Emerg. Microbes Infect. **5**(12): 1–9. doi:10.1038/emi.2016.122.

Lukjancenko, O., Wassenaar, T.M., and Ussery, D.W. 2010. Comparison of 61 Sequenced *Escherichia coli* Genomes. Microb. Ecol. **60**(4): 708–720. doi:10.1007/s00248-010-9717-3.

Luo, C., Walk, S.T., Gordon, D.M., Feldgarden, M., Tiedje, J.M., and Konstantinidis, K.T. 2011. Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. Proc. Natl. Acad. Sci. U. S. A. **108**(17): 7200–7205. doi:10.1073/pnas.1015622108.

Lupolova, N., Dallman, T.J., Matthews, L., Bono, J.L., and Gally, D.L. 2016. Support vector machine applied to predict the zoonotic potential of *E . coli* O157 cattle isolates. Proc. Natl. Acad. Sci. U. S. A. **113**(40): 11312–11317. doi:10.1073/pnas.1606567113.

Lupolova, N., Dallman, T.J., Holden, N.J., and Gally, D.L. 2017. Patchy promiscuity: Machine learning applied to predict the host specificity of *Salmonella enterica* and *Escherichia coli*. Microb. Genomics **3**(10): 1–10. doi:10.1099/mgen.0.000135.

Lupolova, N., Lycett, S.J., and Gally, D.L. 2019. A guide to machine learning for bacterial host attribution using genome sequence data. Microb. Genomics **5**(12): 1–14. doi:10.1099/mgen.0.000317.

Luppi, A. 2017. Swine enteric colibacillosis: Diagnosis, therapy and antimicrobial resistance. Porcine Health Manag. **3**: 1–18. Porcine Health Management. doi:10.1186/s40813-017-0063-4.

Maes, S., Odlare, M., and Jonsson, A. 2022. Fecal indicator organisms in northern oligotrophic rivers: An explorative study on *Escherichia coli* prevalence in a mountain region with intense tourism and reindeer herding. Environ. Monit. Assess. **194**(4). Springer International Publishing. doi:10.1007/s10661-022-09865-1.

Mahfouz, N., Caucci, S., Achatz, E., Semmler, T., Guenther, S., Berendonk, T.U., and Schroeder, M. 2018. High genomic diversity of multi-drug resistant wastewater *Escherichia coli.* Sci. Rep. **8**(1): 1–12. Springer US. doi:10.1038/s41598-018-27292-6.

Maiden, M.C.J., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M., and Spratt, B.G. 1998. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. U. S. A. **95**(6): 3140–3145. doi:10.1073/pnas.95.6.3140.

Mammeri, H., Galleni, M., and Nordmann, P. 2009. Role of the Ser-287-Asn replacement in the hydrolysis spectrum extension of AmpC β-lactamases in *Escherichia coli*. Antimicrob. Agents Chemother. **53**(1): 323–326. doi:10.1128/AAC.00608-08.

Manyi-Loh, C., Mamphweli, S., Meyer, E., and Okoh, A. 2018. Antibiotic use in agriculture and its consequential resistance in environmental sources: Potential public health implications. Molecules **23**(795): 1–48. doi:10.3390/molecules23040795.

Marcobal, A., Southwick, A.M., Earle, K.A., and Sonnenburg, J.L. 2013. A refined palate: Bacterial consumption of host glycans in the gut. Glycobiology, **23**(9): 1038–1046. doi:10.1093/glycob/cwt040.

Maslowska, K.H., Makiela-Dzbenska, K., and Fijalkowska, I.J. 2019. The SOS system: A complex and tightly regulated response to DNA damage. Environ. Mol. Mutagen. **60**(4): 368–384. doi:10.1002/em.22267.

Mathlouthi, A., Pennacchietti, E., and De Biase, D. 2018. Effect of temperature, pH and plasmids on in vitro biofilm formation in *Escherichia coli*. Acta Naturae **10**(4): 129–132. doi:10.32607/20758251-2018-10-4-129-132.

Matsumura, Y., Pitout, J.D.D., Peirano, G., Devinney, R., Noguchi, T., Yamamoto, M., Gomi, R., Matsuda, T., Nakano, S., Nagao, M., Tanaka, M., and Ichiyama, S. 2017. Rapid Identification of Different *Escherichia coli* Sequence Type 131 Clades. Antimicrob. Agents Chemother. **61**(8): e00179-17. doi:10.1128/AAC.00179-17.

Mau, B., Glasner, J.D., Darling, A.E., and Perna, N.T. 2006. Genome-wide detection and analysis of homologous recombination among sequenced strains of *Escherichia coli*. Genome Biol. **7**(5): 1–12. doi:10.1186/gb-2006-7-5-r44.

McDougald, D., Rice, S.A., Barraud, N., Steinberg, P.D., and Kjelleberg, S. 2012. Should we stay or should we go: Mechanisms and ecological consequences for biofilm dispersal. Nat. Rev. Microbiol. **10**(1): 39–50. Nature Publishing Group. doi:10.1038/nrmicro2695.

McGee, L.W., Barhoush, Y., Shima, R., and Hennessy, M. 2023. Phage-resistant mutations impact bacteria susceptibility to future phage infections and antibiotic response. Ecol. Evol. **13**(1): 1–7. doi:10.1002/ece3.9712

Meadows, J.A., and Wargo, M.J. 2015. Carnitine in bacterial physiology and metabolism. Microbiology **161**: 1161–1174. doi:10.1099/mic.0.000080.

von Meijenfeldt, F.A.B., Hogeweg, P., and Dutilh, B.E. 2023. A social niche breadth score reveals niche range strategies of generalists and specialists. Nat. Ecol. Evol. **7**(5): 768–781. doi:10.1038/s41559-023-02027-7

von Mentzer, A., and Svennerholm, A.M. 2023. Colonization factors of human and animal-specific enterotoxigenic *Escherichia coli* (ETEC). Trends Microbiol.: 1–17. doi:10.1016/j.tim.2023.11.001.

Mercer, R.G., Zheng, J., Garcia-Hernandez, R., Ruan, L., Gänzle, M.G., and McMullen, L.M. 2015. Genetic determinants of heat resistance in *Escherichia coli*. Front. Microbiol. **6**: 932. doi:10.3389/fmicb.2015.00932.

Mercer, R., Nguyen, O., Ou, Q., McMullen, L., and Gänzle, M.G. 2017. Functional analysis of genes comprising the locus of heat resistance in *Escherichia coli*. Appl. Environ. Microbiol. **83**(20): 1–13. doi:10.1128/AEM.01400-17.

Micenková, L., Bosák, J., Vrba, M., Ševčíková, A., and Šmajs, D. 2016. Human extraintestinal pathogenic *Escherichia coli* strains differ in prevalence of virulence factors, phylogroups, and bacteriocin determinants. BMC Microbiol. **16**(1): 1–8. doi:10.1186/s12866-016-0835-z.

Milkman, R. 1973. Electrophoretic variation in *Escherichia coli* from natural sources. Science, **182**(4116): 1024–1026. doi:10.1126/science.182.4116.1024.

Milkman, R., and Bridges, M.M. 1990. Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames. Genetics, **126**(3): 505–517.

Mira, A., Martín-Cuadrado, A.B., D'Auria, G., and Rodríguez-Valera, F. 2010. The bacterial pan-genome: A new paradigm in microbiology. Int. Microbiol. **13**(2): 45–57. doi:10.2436/20.1501.01.110.

Mokady, D., Gophna, U., and Ron, E.Z. 2005. Virulence factors of septicemic *Escherichia coli* strains. Int. J. Med. Microbiol. **295**(6–7): 455–462. doi:10.1016/j.ijmm.2005.07.007.

Moldovan, M.A., and Gelfand, M.S. 2018. Pangenomic definition of prokaryotic species and the phylogenetic structure of *Prochlorococcus* spp. Front. Microbiol. **9**(March): 1–11. doi:10.3389/fmicb.2018.00428.

Moulin-Schouleur, M., Répérant, M., Laurent, S., Brée, A., Mignon-Grasteau, S., Germon, P., Rasschaert, D., and Schouler, C. 2007. Extraintestinal pathogenic *Escherichia coli* strains of avian and human origin: Link between phylogenetic relationships and common virulence patterns. J. Clin. Microbiol. **45**(10): 3366–3376. doi:10.1128/JCM.00037-07.

Mousavi, S.A., and Khodadoost, F. 2019. Effects of detergents on natural ecosystems and wastewater treatment processes: a review. Environ. Sci. Pollut. Res. **26**(26): 26439–26448. Environ. Sci. Pollut. Res. doi:10.1007/s11356-019-05802-x.

Mutuku, C., Gazdag, Z., and Melegh, S. 2022. Occurrence of antibiotics and bacterial resistance genes in wastewater: resistance mechanisms and antimicrobial resistance control approaches. World J. Microbiol. Biotechnol. **38**(9): 1–27. Springer Netherlands. doi:10.1007/s11274-022-03334-0.

Nachin, L., Nannmark, U., and Nyström, T. 2005. Differential roles of the universal stress proteins of *Escherichia coli* in oxidative stress resistance, adhesion, and motility. J. Bacteriol. **187**(18): 6265–6272. doi:10.1128/JB.187.18.6265-6272.2005.

Naik, K.S., and Stenstrom, M.K. 2012. Evidence of the influence of wastewater treatment on improved public health. Water Sci. Technol. **66**(3): 644–652. doi:10.2166/wst.2012.144.

Nandakafle, G., Huegen, T., Potgieter, S.C., Steenkamp, E., Venter, S.N., and Brözel, V.S. 2021. Niche preference of *Escherichia coli* in a peri-urban pond ecosystem. Life **11**(10). doi:10.3390/life11101020.

Nandy, P. 2022. The role of sigma factor competition in bacterial adaptation under prolonged starvation. Microbiology **168**(5): 1–11. doi:10.1099/mic.0.001195.

Nicolas-Chanoine, M.H., Bertrand, X., and Madec, J.Y. 2014. *Escherichia coli* ST131, an Intriguing Clonal Group. Clin. Microbiol. Rev. **27**(3): 543–574. doi:10.1128/CMR.00125-13.

Nishino, K., Inazumi, Y., and Yamaguchi, A. 2003. Global analysis of genes regulated by *EvgA* of the two-component regulatory system in *Escherichia coli*. J. Bacteriol. **185**(8): 2667–2672. doi:10.1128/JB.185.8.2667-2672.2003.

Nowrouzian, F.L., Adlerberth, I., and Wold, A.E. 2006. Enhanced persistence in the colonic microbiota of *Escherichia coli* strains belonging to phylogenetic group B2: role of virulence factors and adherence to colonic cells. Microbes Infect. **8**(3): 834—840. doi:10.1016/j.micinf.2005.10.011.

Numberger, D., Ganzert, L., Zoccarato, L., Mühldorfer, K., Sauer, S., Grossart, H.P., and Greenwood, A.D. 2019. Characterization of bacterial communities in wastewater with enhanced taxonomic resolution by full-length 16S rRNA sequencing. Sci. Rep. **9**(1): 1–14. doi:10.1038/s41598-019-46015-z.

Ochman, H., Whittam, T.S., Caugant, D.A., and Selander, R.K. 1983. Enzyme polymorphism and genetic population structure in *Escherichia coli* and *Shigella*. The Journal of General Microbiology, **129**(9): 2715–2726. doi:10.1099/00221287-129-9-2715.

Ochman, H., and Selander, R.K. 1984a. Evidence for clonal population structure in *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. **81**(1 I): 198–201. doi:10.1073/pnas.81.1.198.

Ochman, H., and Selander, R.K. 1984b. Standard reference strains of *Escherichia coli* from natural populations. J. Bacteriol. **157**(2): 690–693.

Omar, K.B., and Barnard, T.G. 2010. The occurrence of pathogenic *Escherichia coli* in South African wastewater treatment plants as detected by multiplex PCR. Water SA **36**(2): 172–176. doi:10.4314/wsa.v36i2.183725.

Ooka, T., Ogura, Y., Katsura, K., Seto, K., Kobayashi, H., Kawano, K., Tokuoka, E., Furukawa, M., Harada, S., Yoshino, S., Seto, J., Ikeda, T., Yamaguchi, K., Murase, K., Gotoh, Y., Imuta, N., Nishi, J., Gomes, T.A., Beutin, L., and Hayashi, T. 2015. Defining the genome features of *Escherichia albertii*, an emerging enteropathogen closely related to *Escherichia coli*. Genome Biol. Evol. **7**(12): 3170–3179. doi:10.1093/gbe/evv211.

Ørskov, F., Ørskov, I., Evans, D.J., Sack, R.B., Sack, D.A., and Wadström, T. 1976. Special *Escherichia coli* serotypes among enterotoxigenic strains from diarrhoea in adults and children. Med. Microbiol. Immunol. **162**(2): 73–80. doi:10.1007/BF02121318.

Osunmakinde, C.O., Selvarajan, R., Mamba, B.B., and Msagati, T.A.M. 2019. Profiling bacterial diversity and potential pathogens in wastewater treatment plants using high-throughput sequencing analysis. Microorganisms **7**(11): 1–18. doi:10.3390/microorganisms7110506.

Ouchenir, L., Renaud, C., Khan, S., Bitnun, A., Boisvert, A.A., McDonald, J., Bowes, J., Brophy, J., Barton, M., Ting, J., Roberts, A., Hawkes, M., and Robinson, J.L. 2017. The epidemiology, management, and outcomes of bacterial meningitis in infants. Pediatrics **140**(1): 1–8. doi:10.1542/peds.2017-0476.

Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.G., Fookes, M., Falush, D., Keane, J.A., and Parkhill, J. 2015. Roary: Rapid large-scale prokaryote pan genome analysis. Bioinformatics **31**(22): 3691–3693. doi:10.1093/bioinformatics/btv421.

Paradis, E., and Schliep, K. 2019. Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics **35**(3): 526–528. doi:10.1093/bioinformatics/bty633.

Paruch, L., and Paruch, A.M. 2022. An Overview of Microbial Source Tracking Using Host-Specific Genetic Markers to Identify Origins of Fecal Contamination in Different Water Environments. Water **14**(11): 1–18. doi:10.3390/w14111809.

Paulshus, E., Thorell, K., Guzman-Otazo, J., Joffre, E., Colque, P., Kühn, I., Möllby, R., Sørum, H., and Sjöling, Å. 2019. Repeated Isolation of Extended-Spectrum-Lactamase-Positive *Escherichia coli* Sequence Types 648 and 131 from Community Wastewater Indicates that Sewage Systems Are Important Sources of Emerging Clones of Antibiotic-Resistant Bacteria. Antimicrob. Agents Chemother. **63**(9). doi:10.1128/AAC.00823-19.

Pearce, M.E., Langridge, G.C., Lauer, A.C., Grant, K., Maiden, M.C.J., and Chattaway, M.A. 2021. An evaluation of the species and subspecies of the genus *Salmonella* with whole genome sequence data: Proposal of type strains and epithets for novel *S. enterica* subspecies VII, VIII, IX, X and XI. Genomics **113**(5): 3152–3162. Elsevier Inc. doi:10.1016/j.ygeno.2021.07.003.

Penders, J., Thijs, C., Vink, C., Stelma, F.F., Snijders, B., Kummeling, I., van den Brandt, P.A., and Stobberingh, E.E. 2006. Factors Influencing the Composition of the Intestinal Microbiota in Early Infancy. Pediatrics, **118**(2): 511–521. doi:10.1542/peds.2005-2824.

Perna, N.T., Plunkett, G., Burland, V., Mau, B., Glasner, J.D., Rose, D.J., Mayhew, G.F., Evans, P.S., Gregor, J., Kirkpatrick, H.A., Pósfai, G., Hackett, J., Klink, S., Boutin, A., Shao, Y., Miller, L., Grotbeck, E.J., Davis, N.W., Lim, A., Dimalanta, E.T., Potamousis, K.D., Apodaca, J., Anantharaman, T.S., Lin, J., Yen, G., Schwartz, D.C., Welch, R.A., and Blattner, F.R. 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7 . Nature, **409**(6819): 529–533. doi:10.1038/35054089.

Petersen, F., and Hubbart, J.A. 2020. Physical factors impacting the survival and occurrence of *Escherichia coli* in secondary habitats. Water **12**(6): 1–15. doi:10.3390/w12061796.

Petersen, K.R., Streett, D.A., Gerritsen, A.T., Hunter, S.S., and Settles, M.L. 2015. Super Deduper, fast PCR duplicate detection in fastq files. BCB 2015 - 6th ACM Conf. Bioinformatics, Comput. Biol. Heal. Informatics: 491–492. doi:10.1145/2808719.2811568.

Pfennig, K.S. 2001. Evolution of pathogen virulence: The role of variation in host phenotype. Proc. R. Soc. B. **268**(1468): 755–760. doi:10.1098/rspb.2000.1582.

Picard, B., Garcia, J.S., Gouriou, S., Duriez, P., Brahimi, N., Bingen, E., Elion, J., and Denamur, E. 1999. The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection? Infect. Immun. **67**(2): 546–553.

Poolman, J.T., and Wacker, M. 2016. Extraintestinal pathogenic *Escherichia coli*, a common human pathogen: challenges for vaccine development and progress in the field. J. Infect. Dis. **213**(1): 6–13. doi:10.1093/infdis/jiv429.

Power, M.L., Littlefield-Wyer, J., Gordon, D.M., Veal, D.A., and Slade, M.B. 2005. Phenotypic and genotypic characterization of encapsulated *Escherichia coli* isolated from blooms in two Australian lakes. Environ. Microbiol. **7**(5): 631–640. doi:10.1111/j.1462-2920.2005.00729.x.

Price, M.N., Dehal, P.S., and Arkin, A.P. 2010. FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS One **5**(3). doi:10.1371/journal.pone.0009490.

Priest, F.G., and Barker, M. 2010. Gram-negative bacteria associated with brewery yeasts: Reclassification of *Obesumbacterium proteus* biogroup 2 as *Shimwellia pseudoproteus* gen. nov., sp. nov., and transfer of *Escherichia blattae* to *Shimwellia blattae* comb. nov. Int. J. Syst. Evol. Microbiol. **60**(4): 828–833. doi:10.1099/ijs.0.013458-0.

Qasem, N.A.A., Mohammed, R.H., and Lawal, D.U. 2021. Removal of heavy metal ions from wastewater: a comprehensive and critical review. npj Clean Water **4**(1). Springer US. doi:10.1038/s41545-021-00127-0.

Qin, X., Hu, F., Wu, S., Ye, X., Zhu, D., Zhang, Y., and Wang, M. 2013. Comparison of Adhesin Genes and Antimicrobial Susceptibilities between Uropathogenic and Intestinal Commensal *Escherichia coli* Strains. PLoS One **8**(4): 1–7. doi:10.1371/journal.pone.0061169.

Rasko, D.A., Rosovitz, M.J., Myers, G.S.A., Mongodin, E.F., Fricke, W.F., Gajer, P., Crabtree, J., Sebaihia, M., Thomson, N.R., Chaudhuri, R., Henderson, I.R., Sperandio, V., and Ravel, J. 2008. The pangenome structure of *Escherichia coli*: Comparative genomic analysis of *E. coli* commensal and pathogenic isolates. J. Bacteriol. **190**(20): 6881–6893. doi:10.1128/JB.00619-08.

Raven, K.E., Ludden, C., Gouliouris, T., Blane, B., Naydenova, P., Brown, N.M., Parkhill, J., and Peacock, S.J. 2019. Genomic surveillance of *Escherichia coli* in municipal wastewater treatment plants as an indicator of clinically relevant pathogens and their resistance genes. Microb. Genomics **5**(5). doi:10.1099/mgen.0.000267.

Rice, J., and Westerhoff, P. 2015. Spatial and temporal variation in de facto wastewater reuse in drinking water systems across the U.S.A. Environ. Sci. Technol. **49**(2): 982–989. doi:10.1021/es5048057.

Rice, J., and Westerhoff, P. 2017. High levels of endocrine pollutants in US streams during low flow due to insufficient wastewater dilution. Nat. Geosci. **10**(8): 587–591. doi:10.1038/NGEO2984.

Riley, L.W. 2014. Pandemic lineages of extraintestinal pathogenic *Escherichia coli*. Clin. Microbiol. Infect. **20**(5): 380–390. European Society of Clinical Infectious Diseases. doi:10.1111/1469-0691.12646.

Rizzo, L., Manaia, C., Merlin, C., Schwartz, T., Dagot, C., Ploy, M.C., Michael, I., and Fatta-Kassinos, D. 2013. Urban wastewater treatment plants as hotspots for antibiotic resistant bacteria and genes spread into the environment: A review. Sci. Total Environ. **447**: 345–360. Elsevier B.V. doi:10.1016/j.scitotenv.2013.01.032.

Rodriguez-Mozaz, S., Vaz-Moreira, I., Varela Della Giustina, S., Llorca, M., Barceló, D., Schubert, S., Berendonk, T.U., Michael-Kordatou, I., Fatta-Kassinos, D., Martinez, J.L., Elpers, C., Henriques, I., Jaeger, T., Schwartz, T., Paulshus, E., O'Sullivan, K., Pärnänen, K.M.M., Virta, M., Do, T.T., Walsh, F., and Manaia, C.M. 2020. Antibiotic residues in final

effluents of European wastewater treatment plants and their impact on the aquatic environment. Environ. Int. **140**(March): 105733. Elsevier. doi:10.1016/j.envint.2020.105733.

Rodriguez-Siek, K.E., Giddings, C.W., Doetkott, C., Johnson, T.J., Fakhr, M.K., and Nolan, L.K. 2005. Comparison of *Escherichia coli* isolates implicated in human urinary tract infection and avian colibacillosis. Microbiology, **151**(6): 2097–2110. doi:10.1099/mic.0.27499-0.

Rogers, B.A., Sidjabat, H.E., and Paterson, D.L. 2011. *Escherichia coli* O25b-ST131: A pandemic, multiresistant, community-associated strain. J. Antimicrob. Chemother. **66**(1): 1–14. doi:10.1093/jac/dkq415.

Ron, E.Z. 2006. Host specificity of septicemic *Escherichia coli*: human and avian pathogens. Curr. Opin. Microbiol. **9**(1): 28–32. doi:10.1016/j.mib.2005.12.001.

Ruczinski, I., Kooperberg, C., and Leblanc, M. 2003. Logic Regression. J. Comput. Graph. Stat. **12**(3): 475–511. doi:10.1198/1061860032238.

Rudenko, N., Golovchenko, M., Grubhoffer, L., and Oliver, J.H. 2011. Updates on *borrelia burgdorferi* sensu lato complex with respect to public health. Ticks Tick Borne. Dis. **2**(3): 123–128. Elsevier GmbH. doi:10.1016/j.ttbdis.2011.04.002.

Runa, V., Wenk, J., Bengtsson, S., Jones, B. V., and Lanham, A.B. 2021. Bacteriophages in Biological Wastewater Treatment Systems: Occurrence, Characterization, and Function. Front. Microbiol. **12**(October). doi:10.3389/fmicb.2021.730071.

Samrot, A. V., Wilson, S., Sanjay Preeth, R.S., Prakash, P., Sathiyasree, M., Saigeetha, S., Shobana, N., Pachiyappan, S., and Rajesh, V.V. 2023. Sources of Antibiotic Contamination in

Wastewater and Approaches to Their Removal—An Overview. Sustainability **15**(16): 1–25. doi:10.3390/su151612639.

Sarowska, J., Futoma-Koloch, B., Jama-Kmiecik, A., Frej-Madrzak, M., Ksiazczyk, M., Bugla-Ploskonska, G., and Choroszy-Krol, I. 2019. Virulence factors, prevalence and potential transmission of extraintestinal pathogenic *Escherichia coli* isolated from different sources: Recent reports. Gut Pathog. **11**(1): 1–16. BioMed Central. doi:10.1186/s13099-019-0290-0.

Scaletsky, I.C.A., Fabbricotti, S.H., Carvalho, R.L.B., Nunes, C.R., Maranhão, H.S., Morais, M.B., and Fagundes-Neto, U. 2002. Diffusely adherent *Escherichia coli* as a cause of acute diarrhea in young children in northeast Brazil: A case-control study. J. Clin. Microbiol. **40**(2): 645–648. doi:10.1128/JCM.40.2.645-648.2002.

Schierack, P., Walk, N., Ewers, C., Wilking, H., Steinrück, H., Filter, M., and Wieler, L.H. 2008. ExPEC-typical virulence-associated genes correlate with successful colonization by intestinal *E. coli* in a small piglet group. Environ. Microbiol. **10**(7): 1742–1751. doi:10.1111/j.1462-2920.2008.01595.x.

Scott, T.M., Rose, J.B., Jenkins, T.M., Farrah, S.R., and Lukasik, J. 2002. Microbial source tracking: Current methodology and future directions. Appl. Environ. Microbiol. **68**(12): 5796–5803. doi:10.1128/AEM.68.12.5796-5803.2002.

Seemann, T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics **30**(14): 2068–2069. doi:10.1093/bioinformatics/btu153.

Selander, R.K., and Levin, B.R. 1980. Genetic diversity and structure in *Escherichia coli* populations. Science, **210**(4469): 545–547. doi:10.1126/science.6999623.

Serapio-Palacios, A., Woodward, S.E., Vogt, S.L., Deng, W., Creus-Cuadros, A., Huus, K.E., Cirstea, M., Gerrie, M., Barcik, W., Yu, H., and Finlay, B.B. 2022. Type VI secretion systems of pathogenic and commensal bacteria mediate niche occupancy in the gut. Cell Rep. **39**(4): 1-16. doi:10.1016/j.celrep.2022.110731.

Sheludchenko, M.S., Huygens, F., and Hargreaves, M.H. 2010. Highly discriminatory single-nucleotide polymorphism interrogation of *Escherichia coli* by use of allele-specific real-time PCR and eBURST analysis. Appl. Environ. Microbiol. **76**(13): 4337–4345. doi:10.1128/AEM.00128-10.

Sheng, H., Lim, J.Y., Knecht, H.J., Li, J., and Hovde, C.J. 2006. Role of *Escherichia coli* O157:H7 virulence factors in colonization at the bovine terminal rectal mucosa. Infect. Immun. **74**(8): 4685–4693. doi:10.1128/IAI.00406-06.

Sicard, J.F., Bihan, G. Le, Vogeleer, P., Jacques, M., and Harel, J. 2017. Interactions of intestinal bacteria with components of the intestinal mucus. Front. Cell. Infect. Microbiol. **7**: 387. doi:10.3389/fcimb.2017.00387.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J.D., and Higgins, D.G. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. **7**(539). doi:10.1038/msb.2011.75.

Singh, V. 2013. *Salmonella* Serovars and Their Host Specificity. J. Vet. Sci. Anim. Husb. **1**(3): 1–4. doi:10.15744/2348-9790.1.301

Šlapeta, J. 2013. Ten simple rules for describing a new (parasite) species. Int. J. Parasitol. Parasites Wildl. **2**(1): 152–154. Australian Society for Parasitology. doi:10.1016/j.ijppaw.2013.03.005.

Slater, S.L., Sågfors, A.M., Pollard, D.J., Ruano-Gallego, D., and Frankel, G. 2018. The Type III Secretion System of Pathogenic *Escherichia coli*. *In* Escherichia coli, a Versatile Pathogen. pp. 51–72. doi:10.1007/82.

Smati, M., Clermont, O., Bleibtreu, A., Fourreau, F., David, A., Daubié, A.S., Hignard, C., Loison, O., Picard, B., and Denamur, E. 2015. Quantitative analysis of commensal *Escherichia coli* populations reveals host-specific enterotypes at the intra-species level. Microbiologyopen **4**(4): 604–615. doi:10.1002/mbo3.266.

Smith, J.M., Smith, N.H., O'Rourke, M., and Spratt, B.G. 1993. How clonal are bacteria? Proc. Natl. Acad. Sci. U. S. A. **90**(10): 4384–4388. doi:10.1073/pnas.90.10.4384.

Smug, B.J., Majkowska-Skrobek, G., and Drulis-Kawa, Z. 2022. PhREEPred: Phage Resistance Emergence Prediction Web Tool to Foresee Encapsulated Bacterial Escape from Phage Cocktail Treatment. J. Mol. Biol. **434**(14): 1–17. doi:10.1016/j.jmb.2022.167670

Søraas, A., Sundsfjord, A., Sandven, I., Brunborg, C., and Jenum, P.A. 2013. Risk Factors for Community-Acquired Urinary Tract Infections Caused by ESBL-Producing

*Enterobacteriaceae* - A Case-Control Study in a Low Prevalence Country. PLoS One **8**(7): 1–7. doi:10.1371/journal.pone.0069581.

Spurbeck, R.R., Dinh, P.C., Walk, S.T., Stapleton, A.E., Hooton, T.M., Nolan, L.K., Kim, K.S., Johnson, J.R., and Mobley, H.L.T. 2012. *Escherichia coli* Isolates That Carry *vat*, *fyuA*, *chuA*, and *yfcV* Efficiently Colonize the Urinary Tract. Infect. Immun. **80**(12): 4115–4122. doi:10.1128/IAI.00752-12.

Stamatakis, A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**(9): 1312–1313. doi:10.1093/bioinformatics/btu033.

Steenbergen, S.M., Jirik, J.L., and Vimr, E.R. 2009. *YjhS* (*NanS*) is required for *Escherichia coli* to grow on 9-O-acetylated N-acetylneuraminic acid. J. Bacteriol. **191**(22): 7134–7139. doi:10.1128/JB.01000-09.

Steinsland, H., Lacher, D.W., Sommerfelt, H., and Whittam, T.S. 2010. Ancestral lineages of human enterotoxigenic *Escherichia coli*. J. Clin. Microbiol. **48**(8): 2916–2924. doi:10.1128/JCM.02432-09.

Strange, J.E.S., Leekitcharoenphon, P., Møller, F.D., and Aarestrup, F.M. 2021. Metagenomics analysis of bacteriophages and antimicrobial resistance from global urban sewage. Sci. Rep. **11**(1): 1–11. Nature Publishing Group UK. doi:10.1038/s41598-021-80990-6.

Su, Y., Ma, G., Zheng, Y., Qin, J., Li, X., Ge, Q., Sun, H., & Liu, B. (2023). Neonatal Meningitis-Causing *Escherichia coli* Induces Microglia Activation which Acts as a Double-

Edged Sword in Bacterial Meningitis. Int. J. Mol. Sci. **24**(12): 1–12. doi:10.3390/ijms24129915

Suess, E., Berg, M., Bouchet, S., Cayo, L., Hug, S.J., Kaegi, R., Voegelin, A., Winkel, L.H.E., Tessier, E., Amouroux, D., and Buser, A.M. 2020. Mercury loads and fluxes from wastewater: A nationwide survey in Switzerland. Water Res. **175**: 115708. Elsevier Ltd. doi:10.1016/j.watres.2020.115708.

Takahashi, A., Kanamaru, S., Kurazono, H., Kunishima, Y., Tsukamoto, T., Ogawa, O., and Yamamoto, S. 2006. *Escherichia coli* isolates associated with uncomplicated and complicated cystitis and asymptomatic bacteriuria possess similar phylogenies, virulence genes, and O-serogroup profiles. J. Clin. Microbiol. **44**(12): 4589–4592. doi:10.1128/JCM.02070-06.

Tamura, K., Sakazaki, R., Kosako, Y., and Yoshizaki, E. 1986. *Leclercia adecarboxylata* Gen. Nov., Comb. Nov., formerly known as *Escherichia adecarboxylata*. Curr. Microbiol. **13**(4): 179–184. doi:10.1007/BF01568943.

Tanabe, H., Tanabe, H., Yamasaki, K., Yamasaki, K., Furue, M., Furue, M., Yamamoto, K., Yamamoto, K., Katoh, A., Katoh, A., Yamamoto, M., Yamamoto, M., Yoshioka, S., Yoshioka, S., Tagami, H., Tagami, H., Utsumi, R., and Utsumi, R. 1997. Growth phase-dependent transcription of *emrKY*, a homolog of multidrug efflux *emrAB* genes of *Escherichia coli*, is induced by tetracycline. J. Gen. Appl. Microbiol. **43**: 257–263.

Tanner, W.D., VanDerslice, J.A., Goel, R.K., Leecaster, M.K., Fisher, M.A., Olstadt, J., Gurley, C.M., Morris, A.G., Seely, K.A., Chapman, L., Korando, M., Shabazz, K.A., Stadsholt, A., VanDeVelde, J., Braun-Howland, E., Minihane, C., Higgins, P.J., Deras, M., Jaber, O.,

Jette, D., and Gundlapalli, A. V. 2019. Multi-state study of *Enterobacteriaceae* harboring extended-spectrum beta-lactamase and carbapenemase genes in U.S. drinking water. Sci. Rep. **9**(1): 1–8. Springer US. doi:10.1038/s41598-019-40420-0.

Taylor, L.H., Latham, S.M., and Woolhouse, M.E.J. 2001. Risk factors for human disease emergence. Philos. Trans. R. Soc. Lond. B. Biol. Sci. **356**(1411): 983–989. doi:10.1098/rstb.2001.0888.

Tenaillon, O., Skurnik, D., Picard, B., and Denamur, E. 2010. The population genetics of commensal *Escherichia coli*. Nat. Rev. Microbiol. **8**(3): 207–217. Nature Publishing Group. doi:10.1038/nrmicro2298.

Terlizzi, M.E., Gribaudo, G., and Maffei, M.E. 2017. UroPathogenic *Escherichia coli* (UPEC) infections: Virulence factors, bladder responses, antibiotic, and non-antibiotic antimicrobial strategies. Front. Microbiol. **8**: 1566. doi:10.3389/fmicb.2017.01566.

Thorpe, H.A., Bayliss, S.C., Hurst, L.D., and Feil, E.J. 2017. Comparative analyses of selection operating on nontranslated intergenic regions of diverse bacterial species. Genetics **206**(1): 363–376. doi:10.1534/genetics.116.195784.

Tiwari, S.K., van der Putten, B.C.L., Fuchs, T.M., Vinh, T.N., Bootsma, M., Oldenkamp, R., La Ragione, R., Matamoros, S., Hoa, N.T., Berens, C., Leng, J., Álvarez, J., Ferrandis-Vila, M., Ritchie, J.M., Fruth, A., Schwarz, S., Domínguez, L., Ugarte-Ruiz, M., Bethe, A., Huber, C., Johanns, V., Stamm, I., Wieler, L.H., Ewers, C., Fivian-Hughes, A., Schmidt, H., Menge, C., Semmler, T., and Schultsz, C. 2023. Genome-wide association reveals host-specific

genomic traits in *Escherichia coli*. BMC Biol. **21**(1): 1–14. BioMed Central. doi:10.1186/s12915-023-01562-w.

Tobe, T., and Sasakawa, C. 2002. Species-specific cell adhesion of enteropathogenic *Escherichia coli* is mediated by type IV bundle-forming pili. Cell. Microbiol. **4**(1): 29–42. doi:10.1046/j.1462-5822.2002.00167.x.

Topp, E., Welsh, M., Tien, Y.C., Dang, A., Lazarovits, G., Conn, K., and Zhu, H. 2003. Strain-dependent variability in growth and survival of *Escherichia coli* in agricultural soil. FEMS Microbiol. Ecol. **44**(3): 303–308. doi:10.1016/S0168-6496(03)00055-2.

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., Bouvet, O., Calteau, A., Chiapello, H., Clermont, O., Cruveiller, S., Danchin, A., Diard, M., Dossat, C., El Karoui, M., Frapy, E., Garry, L., Ghigo, J.M., Gilles, A.M., Johnson, J., Le Bouguénec, C., Lescat, M., Mangenot, S., Martinez-Jéhanne, V., Matic, I., Nassif, X., Oztas, S., Petit, M.A., Pichon, C., Rouy, Z., Ruf, C. Saint, Schneider, D., Tourret, J., Vacherie, B., Vallenet, D., Médigue, C., Rocha, E.P.C., and Denamur, E. 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet. **5**(1). doi:10.1371/journal.pgen.1000344.

Touchon, M., Perrin, A., De Sousa, J.A.M., Vangchhia, B., Burn, S., O'Brien, C.L., Denamur, E., Gordon, D., and Rocha, E.P.C. 2020. Phylogenetic background and habitat drive the genetic diversification of *Escherichia coli.* PLoS Genet. **16**(6): 1–43. doi:10.1371/journal.pgen.1008866.

Tramonti, A., De Canio, M., Delany, I., Scarlato, V., and De Biase, D. 2006. Mechanisms of transcription activation exerted by *GadX* and *GadW* at the *gadA* and *gadBC* gene promoters of the glutamate-based acid resistance system in *Escherichia coli*. J. Bacteriol. **188**(23): 8118–8127. doi:10.1128/JB.01044-06.

Trofa, A.F., Ueno-Olsen, H., Oiwa, R., and Yoshikawa, M. 1999. Dr. Kiyoshi Shiga: Discoverer of the dysentery bacillus. Clin. Infect. Dis. **29**(5): 1303–1306. doi:10.1086/313437.

Turner, S.M., Chaudhuri, R.R., Jiang, Z.D., DuPont, H., Gyles, C., Penn, C.W., Pallen, M.J., and Henderson, I.R. 2006. Phylogenetic comparisons reveal multiple acquisitions of the toxin genes by enterotoxigenic *Escherichia coli* strains of different evolutionary lineages. J. Clin. Microbiol. **44**(12): 4528–4536. doi:10.1128/JCM.01474-06.

Tymensen, L.D., Pyrdok, F., Coles, D., Koning, W., Mcallister, T.A., Jokinen, C.C., Dowd, S.E., and Neumann, N.F. 2015. Comparative accessory gene fingerprinting of surface water *Escherichia coli* reveals genetically diverse naturalized population. J. Appl. Microbiol. **119**(1): 263–277. doi:10.1111/jam.12814.

Uluseker, C., Kaster, K.M., Thorsen, K., Basiry, D., Shobana, S., Jain, M., Kumar, G., Kommedal, R., and Pala-Ozkok, I. 2021. A Review on Occurrence and Spread of Antibiotic Resistance in Wastewaters and in Wastewater Treatment Plants: Mechanisms and Perspectives. Front. Microbiol. **12**(October). doi:10.3389/fmicb.2021.717809.

Ungureanu, G., Santos, S., Boaventura, R., and Botelho, C. 2015. Arsenic and antimony in water and wastewater: Overview of removal techniques with special reference to latest

advances in adsorption. J. Environ. Manage. **151**: 326–342. doi:10.1016/j.jenvman.2014.12.051.

Vamosi, J.C., Scott Armbruster, W., Scott Armbruster, W., Scott Armbruster, W., and Renner, S.S. 2014. Evolutionary ecology of specialization: Insights from phylogenetic analysis. Proc. R. Soc. B Biol. Sci. **281**(1795). doi:10.1098/rspb.2014.2004.

Van Den Beld, M.J.C., and Reubsaet, F.A.G. 2012. Differentiation between *Shigella*, enteroinvasive *Escherichia coli* (EIEC) and noninvasive *Escherichia coli*. Eur. J. Clin. Microbiol. Infect. Dis. **31**(6): 899–904. doi:10.1007/s10096-011-1395-7.

Van Hoek, A.H.A.M., Mevius, D., Guerra, B., Mullany, P., Roberts, A.P., and Aarts, H.J.M. 2011. Acquired antibiotic resistance genes: An overview. Front. Microbiol. **2**(September): 1–27. doi:10.3389/fmicb.2011.00203.

Van Ingen, J., Turenne, C.Y., Tortoli, E., Wallace, R.J., and Brown-Elliott, B.A. 2018. A definition of the *Mycobacterium avium* complex for taxonomical and clinical purposes, a review. Int. J. Syst. Evol. Microbiol. **68**(11): 3666–3677. doi:10.1099/ijsem.0.003026.

Vinayamohan, P.G., Pellissery, A.J., and Venkitanarayanan, K. 2022. Role of horizontal gene transfer in the dissemination of antimicrobial resistance in food animal production. Curr. Opin. Food Sci. **47**: 100882. Elsevier. doi:10.1016/j.cofs.2022.100882.

Walk, S.T. 2015. The “Cryptic” *Escherichia*. EcoSal Plus, **6**(2). doi:10.1128/ecosalplus.esp-0002-2015.

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

Wall, E., Majdalani, N., and Gottesman, S. 2018. The Complex *Rcs* Regulatory Cascade. Annu. Rev. Microbiol. **72**: 111–139. doi:10.1146/annurev-micro-090817-062640.

Wang, H., Zhong, Z., Luo, Y., Cox, E., and Devriendt, B. 2019. Heat-stable enterotoxins of enterotoxigenic *Escherichia coli* and their impact on host immunity. Toxins, **11**(1): 1–12. doi:10.3390/toxins11010024.

Wang, L., Wakushima, M., Aota, T., Yoshida, Y., Kita, T., Maehara, T., Ogasawara, J., Choi, C., Kamata, Y., Hara-Kudo, Y., and Nishikawa, Y. 2013. Specific properties of enteropathogenic *Escherichia coli* isolates from diarrheal patients and comparison to strains from foods and fecal specimens from cattle, swine, and healthy carriers in Osaka City, Japan. Appl. Environ. Microbiol. **79**(4): 1232–1240. doi:10.1128/AEM.03380-12.

Wang, Z., Fang, Y., Zhi, S., Simpson, D.J., Gill, A., Mcmullen, L.M., Neumann, N.F., and Gänzle, M.G. 2020. The Locus of Heat Resistance Confers Resistance to Chlorine and Other Oxidizing Chemicals in *Escherichia coli*. Appl. Environ. Microbiol. **86**(4): 1–16.

Warner, D.M., Yang, Q., Duval, V., Chen, M., Xu, Y., and Levy, S.B. 2013. Involvement of *MarR* and *YedS* in carbapenem resistance in a clinical isolate of *Escherichia coli* from China. Antimicrob. Agents Chemother. **57**(4): 1935–1937. doi:10.1128/AAC.02445-12.

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. 2009. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. Bioinformatics **25**(9): 1189–1191. doi:10.1093/bioinformatics/btp033.

Welch, R.A., Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D., Buckles, E.L., Liou, S.R., Boutin, A., Hackett, J., Stroud, D., Mayhew, G.F., Rose, D.J., Zhou, S., Schwartz, D.C., Perna, N.T., Mobley, H.L.T., Donnenberg, M.S., and Blattner, F.R. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. **99**(26): 17020–17024. doi:10.1073/pnas.252529799.

Wenzel, H., Kaminski, R.W., Clarkson, K.A., Maciel, M., Smith, M.A., Zhang, W., and Oaks, E. V. 2017. Improving chances for successful clinical outcomes with better preclinical models. Vaccine, **35**(49): 6798–6802. Elsevier Ltd. doi:10.1016/j.vaccine.2017.08.030.

Whelan, F.J., Rusilowicz, M., and McInerney, J.O. 2020. Coinfinder: Detecting significant associations and dissociations in pangenomes. Microb. Genomics **6**(3). doi:10.1099/mgen.0.000338.

White, A.P., Sibley, K.A., Sibley, C.D., Wasmuth, J.D., Schaefer, R., Surette, M.G., Edge, T.A., and Neumann, N.F. 2011. Intergenic sequence comparison of *Escherichia coli* isolates reveals lifestyle adaptations but not host specificity. Appl. Environ. Microbiol. **77**(21): 7620–7632. doi:10.1128/AEM.05909-11.

Whitlock, M.C. 1996. The Red Queen Beats the Jack-Of-All-Trades: The Limitations on the Evolution of Phenotypic Plasticity and Niche Breadth. Am. Nat. **148**: S65–S77. [University of

Chicago Press, American Society of Naturalists]. Available from http://www.jstor.org/stable/2463048.

Wickham, H. 2009. Elegant Graphics for Data Analysis: ggplot2. *In* Applied Spatial Data Analysis with R.

Wijetunge, D.S.S., Gongati, S., Debroy, C., Kim, K.S., Couraud, P.O., Romero, I.A., Weksler, B., and Kariyawasam, S. 2015. Characterizing the pathotype of neonatal meningitis causing *Escherichia coli* (NMEC). BMC Microbiol. **15**(1): 211. doi:10.1186/s12866-015-0547-9.

Winter, J., Ilbert, M., Graf, P.C.F., Özcelik, D., and Jakob, U. 2008. Bleach Activates a Redox-Regulated Chaperone by Oxidative Protein Unfolding. Cell **135**(4): 691–701. doi:10.1016/j.cell.2008.09.024.

Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., Wieler, L.H., Karch, H., Reeves, P.R., Maiden, M.C.J., Ochman, H., and Achtman, M. 2006. Sex and virulence in *Escherichia coli*: An evolutionary perspective. Mol. Microbiol. **60**(5): 1136–1151. doi:10.1111/j.1365-2958.2006.05172.x.

Wold, A.E., Caugant, D.A., Lidin-Janson, G., de Man, P., and Svanborg, C. 1992. Resident Colonic *Escherichia coli* Strains Frequently Display Uropathogenic Characteristics. J. Infect. Dis. **165**(1): 46–52. doi:10.1093/infdis/165.1.46.

Wood, J.M., Becraft, E.D., Krizanc, D., Cohan, F.M., and Ward, D.M. 2020. Ecotype Simulation 2: An improved algorithm for efficiently demarcating microbial species from large

sequence datasets. bioRxiv: 2020.02.10.940734. Available from https://doi.org/10.1101/2020.02.10.940734%0A

Woolhouse, M.E.J., Taylor, L.H., and Haydon, D.T. 2001. Population biology of multihost pathogens. Science **292**(5519): 1109–1112. doi:10.1126/science.1059026.

Wu, Y.H., Cheng, M.F., Lai, C.H., Lin, H.H., Hung, C.H., and Wang, J.L. 2014. The role of Sequence Type (ST) 131 in adult community-onset non-ESBL-producing *Escherichia coli* bacteraemia. BMC Infect. Dis. **14**(1): 1–7. doi:10.1186/s12879-014-0579-z.

Yang, X., Wang, H., He, A., and Tran, F. 2018. Biofilm formation and susceptibility to biocides of recurring and transient *Escherichia coli* isolated from meat fabrication equipment. Food Control **90**: 205–211. Elsevier Ltd. doi:10.1016/j.foodcont.2018.02.050.

Yang, X., Tran, F., Zhang, P., and Wang, H. 2021. Genomic and phenotypic analysis of heat and sanitizer resistance in *Escherichia coli* from beef in relation to the locus of heat resistance. Appl. Environ. Microbiol. **87**(23): 1–17. American Society for Microbiology. doi:10.1128/AEM.01574-21.

Yang, Z., Kovar, J., Kim, J., Nietfeldt, J., Smith, D.R., Moxley, R.A., Olson, M.E., Fey, P.D., and Benson, A.K. 2004. Identification of common subpopulations of non-sorbitol-fermenting, β-glucuronidase-negative *Escherichia coli* O157:H7 from bovine production environments and human clinical samples. Appl. Environ. Microbiol. **70**(11): 6846–6854. doi:10.1128/AEM.70.11.6846-6854.2004.

Yin, W., Wang, Y., Liu, L., and He, J. 2019. Biofilms: The microbial “protective clothing” in extreme environments. Int. J. Mol. Sci. **20**(14). doi:10.3390/ijms20143423

Yu, D., Banting, G., and Neumann, N.F. 2021. A review of the taxonomy, genetics, and biology of the genus *Escherichia* and the type species *Escherichia coli*. Can. J. Microbiol. **67**(8): 553–571. doi:10.1139/cjm-2020-0508

Yu, D., Ryu, K., Zhi, S., Otto, S.J.G., and Neumann, N.F. 2022. Naturalized *Escherichia coli* in Wastewater and the Co-evolution of Bacterial Resistance to Water Treatment and Antibiotics. Front. Microbiol. **13**(May). doi:10.3389/fmicb.2022.810312.

Yu, G., Smith, D.K., Zhu, H., Guan, Y., and Lam, T.T.Y. 2017. Ggtree: an R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data. Methods Ecol. Evol. **8**(1): 28–36. doi:10.1111/2041-210X.12628.

Yue, M., Han, X., Masi, L. De, Zhu, C., Ma, X., Zhang, J., Wu, R., Schmieder, R., Kaushik, R.S., Fraser, G.P., Zhao, S., McDermott, P.F., Weill, F.X., Mainil, J.G., Arze, C., Fricke, W.F., Edwards, R.A., Brisson, D., Zhang, N.R., Rankin, S.C., and Schifferli, D.M. 2015. Allelic variation contributes to bacterial host specificity. Nat. Commun. **6**. doi:10.1038/ncomms9754.

Zaslaver, A., Bren, A., Ronen, M., Itzkovitz, S., Kikoin, I., Shavit, S., Liebermeister, W., Surette, M.G., and Alon, U. 2006. A comprehensive library of fluorescent transcriptional reporters for *Escherichia coli*. Nat. Methods **3**(8): 623–628. doi:10.1038/nmeth895.

Zhang, H., Chen, X., Nolan, L.K., Zhang, W., and Li, G. 2019a. Identification of host adaptation genes in extraintestinal pathogenic *Escherichia coli* during infection in different hosts. Infect. Immun. **87**(12): 1–12. doi:10.1128/IAI.00666-19.

Zhang, S., Li, S., Gu, W., Den Bakker, H., Boxrud, D., Taylor, A., Roe, C., Driebe, E., Engelthaler, D.M., Allard, M., Brown, E., McDermott, P., Zhao, S., Bruce, B.B., Trees, E., Fields, P.I., and Deng, X. 2019b. Zoonotic source attribution of *Salmonella enterica* serotype Typhimurium using genomic surveillance data, United States. Emerg. Infect. Dis. **25**(1): 82–91. doi:10.3201/eid2501.180835.

Zhang, S., Wang, Y., Lu, J., Yu, Z., Song, H., Bond, P. L., and Guo, J. 2021. Chlorine disinfection facilitates natural transformation through ROS-mediated oxidative stress. ISME J. **15**(10), 2969–2985. doi:0.1038/s41396-021-00980-4

Zhang, T., Shi, X.C., Xia, Y., Mai, L., and Tremblay, P.L. 2019c. *Escherichia coli* adaptation and response to exposure to heavy atmospheric pollution. Sci. Rep. **9**(1): 1–13. doi:10.1038/s41598-019-47427-7.

Zhang, Y., and Lin, K. 2012. A phylogenomic analysis of *Escherichia coli*/*Shigella* group: Implications of genomic features associated with pathogenicity and ecological adaptation. BMC Evol. Biol. **12**(1). doi:10.1186/1471-2148-12-174.

Zhang, Z., Zhou, K., Tran, D., and Saier, M. 2022. Insertion Sequence (IS) Element-Mediated Activating Mutations of the Cryptic Aromatic β-Glucoside Utilization (*BglGFB*) Operon Are Promoted by the Anti-Terminator Protein (*BglG*) in *Escherichia coli*. Int. J. Mol. Sci. **23**(3). doi:10.3390/ijms23031505.

Zhi, S., Li, Q., Yasui, Y., Edge, T., Topp, E., and Neumann, N.F. 2015. Assessing host-specificity of *Escherichia coli* using a supervised learning logic-regression-based analysis of single nucleotide polymorphisms in intergenic regions. Mol. Phylogenet. Evol. **92**: 72–81. Academic Press. doi:10.1016/J.YMPEV.2015.06.007.

Zhi, S., Banting, G., Li, Q., Edge, T.A., Topp, E., Sokurenko, M., Scott, C., Braithwaite, S., Ruecker, N.J., Yasui, Y., McAllister, T., Chui, L., and Neumann, N.F. 2016a. Evidence of Naturalized Stress-Tolerant Strains of *Escherichia coli* in Municipal Wastewater Treatment Plants. Appl. Environ. Microbiol. **82**(18): 5505–5518. doi:10.1128/AEM.00143-16.Editor.

Zhi, S., Li, Q., Yasui, Y., Banting, G., Edge, T.A., Topp, E., McAllister, T.A., and Neumann, N.F. 2016b. An evaluation of logic regression-based biomarker discovery across multiple intergenic regions for predicting host specificity in *Escherichia coli*. Mol. Phylogenet. Evol. **103**: 133–142. Elsevier Inc. doi:10.1016/j.ympev.2016.07.016.

Zhi, S., Banting, G.S., Ruecker, N.J., and Neumann, N.F. 2017. Stress resistance in naturalised waste water *E. coli* strains. J. Environ. Eng. Sci. **12**(2): 42–50. doi:10.1680/jenes.16.00021.

Zhi, S., Banting, G., Stothard, P., Ashbolt, N.J., Checkley, S., Meyer, K., Otto, S., and Neumann, N.F. 2019. Evidence for the evolution, clonal expansion and global dissemination of water treatment-resistant naturalized strains of *Escherichia coli* in wastewater. Water Res. **156**: 208–222. Elsevier Ltd. doi:10.1016/j.watres.2019.03.024.

Zhi, S., Stothard, P., Banting, G., Scott, C., Huntley, K., Ryu, K., Otto, S., Ashbolt, N., Checkley, S., Dong, T., Ruecker, N.J., and Neumann, N.F. 2020. Characterization of water

treatment-resistant and multidrug-resistant urinary pathogenic *Escherichia coli* in treated wastewater. Water Res. **182**: 115827. Elsevier Ltd. doi:10.1016/j.watres.2020.115827.

Zhi, S., Banting, G., and Neumann, N.F. 2022. Development of a qPCR assay for the detection of naturalized wastewater *E. coli* strains. J. Water Health **20**(4): 727–736. doi:10.2166/wh.2022.014.

Zhou, J., Ren, H., Hu, M., Zhou, J., Li, B., Kong, N., Zhang, Q., Jin, Y., Liang, L., and Yue, J. 2020. Characterization of *Burkholderia cepacia* Complex Core Genome and the Underlying Recombination and Positive Selection. Front. Genet. **11**(May): 1–15. doi:10.3389/fgene.2020.00506.

Zhou, Z., Li, X., Liu, B., Beutin, L., Xu, J., Ren, Y., Feng, L., Lan, R., Reeves, P.R., and Wang, L. 2010. Derivation of *Escherichia coli* O157:H7 from its O55:H7 precursor. PLoS One, **5**(1). doi:10.1371/journal.pone.0008700.

Zurfluh, K., Stephan, R., Klumpp, J., Nüesch-Inderbinen, M., Hummerjohann, J., Bagutti, C., and Marti, R. 2017. Complete Genome Sequence of *Escherichia coli* ABWA45, an rmtB-Encoding Wastewater Isolate. Genome Announc. **5**(34): 1–2.

Zwirzitz, B., Wetzels, S.U., Dixon, E.D., Stessl, B., Zaiser, A., Rabanser, I., Thalguter, S., Pinior, B., Roch, F.F., Strachan, C., Zanghellini, J., Dzieciol, M., Wagner, M., and Selberherr, E. 2020. The sources and transmission routes of microbial populations throughout a meat processing facility. npj Biofilms Microbiomes **6**(1): 1–12. Springer US. doi:10.1038/s41522-020-0136-z.