Molecular epidemiology of the multidrug-resistant *Escherichia coli* sequence type 131-*H30* lineage among U.S. children

Arianna Danielle Miles-Jay

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

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Janet Baseman, Chair

Danielle Zerr

Brad Cookson

Program Authorized to Offer Degree:
Epidemiology

University of Washington

Abstract

Molecular epidemiology of the multidrug-resistant *Escherichia coli* sequence type 131-*H30* lineage among U.S. children

Arianna Danielle Miles-Jay

Chair of the Supervisory Committee:
Janet Baseman
Epidemiology

Escherichia coli sequence type 131-*H30* is a globally important pathogen implicated in rising rates of antimicrobial resistance among extraintestinal *E. coli* infections. *H30* causes both community- and healthcare-associated infections, and is associated with resistance to several commonly used antimicrobial agents. This dissertation addresses several knowledge gaps about the epidemiology and transmission dynamics of *H30* among U.S. children through the integration of high-resolution molecular data from clinical extraintestinal *E. coli* isolates with patient epidemiologic data. I observed that although *H30* is less common among extraintestinal *E. coli* collected from children compared to reported estimates among adults, it is similarly dominant among very antimicrobial-resistant isolates. Additionally, *H30* is especially dominant among young children when compared to other types of antimicrobial-resistant extraintestinal *E. coli*. Whole genome sequencing analyses provided proof of principle that putative transmission clusters of *H30* can be identified from passively collected clinical isolates. Integration of data describing patient healthcare contact into a temporal phylogenomic analysis revealed that ancestral *H30* isolates were more likely to be community-associated than healthcare-associated. Finally, when evaluating the evolutionary dynamics of resistance to trimethoprim-sulfamethoxazole, a commonly used

antimicrobial agent in pediatrics, I found that the acquisition of resistance to this agent likely occurred prior to the differentiation of specific *H30* subtypes. Together, these findings highlight that high-resolution molecular analyses of isolates collected during routine clinical care, when combined with patient data, can offer valuable insights into resistance and transmission dynamics of concerning antimicrobial-resistant pathogens like *H30*.

Table of Contents

List of Figures

List of Tables

Acknowledgments

This dissertation would not have been possible without the many people who supported me throughout this process. First, to my committee members – Drs. Janet Baseman, Danielle Zerr, Scott Weissman, Vladimir Minin, and Brad Cookson – thank you for valuing my ideas and independence and for trusting my instincts enough to agree to guide me down this path. I would especially like to recognize my good fortune in acquiring not one, but two, remarkable women scientist mentors to look up to in Drs. Danielle Zerr and Janet Basemen. In particular, to Danielle: thank you for staying with me through all of the iterations and ups and downs of this project, for reading all of my first drafts, and for being a relentless cheerleader; and to Janet: thank you for imparting lessons on the importance of plan B, perspective, and persistence.

To my Master's degree mentors, Drs. Steven Pergam and Ali Rowhani-Rahbar: thank you for encouraging me to achieve my potential, for being the first to tell me that if I wanted to be a scientist, I could, and for making sure I knew that your confidence in my abilities never wavered. To those that helped me get the laboratory work completed – Huxley Smart, Jeff Myers, and the Northwest Genomics Center – thank you for your companionship and expertise. Thanks to Amanda Adler for being a fountain of knowledge and for your meticulous management of the original research study and databases. To all of those around the world that helped me navigate the choppy waters of learning high-throughput computing and bioinformatics; to Dr. Torsten Seemann for developing top-notch open-source tools; to Dr. Kat Holt for your correspondence and encouragement; and to all of the others I met near and far that welcomed me into the genomic epidemiology community: thank you.

To those patients that contributed data and samples to this study: thank you. To my funders and sponsors throughout this process – The National Institutes of Health, The Uni-

versity of Washington School of Public Health, The Institute for Translational Health Sciences, and Seattle Children's Hospital and Research Institute – thank you for your support. In particular, thanks to the Seattle Children's Bioinformatics and High Throughput Analytics team, for providing top-notch computing resources and readily available consultation. Your availability and willingness to assist allowed me to develop my own data analytics workflow, which was invaluable. The availability of these resources will undoubtedly expedite innovation among current and future Seattle Children's scientists.

Thank you to my PhD student colleagues, particularly Drs. Jessica Williams-Nguyen and Lauren Schwartz, for your steady camaraderie and encouragement. To Dr. Sarah Sullivan-Singh, thank you for sharing your wisdom and your heart. To other friends, family, and colleagues who always believed in me and who celebrated my victories: thank you. To my parents, thank you for raising me to believe that the sky is the limit. To my dog, Loki, thank you for serving as a daily and tangible embodiment of joy and the playfulness in life. And finally to my husband, Eric, for being my number one fan, my steadfast partner, and for confidently standing with me through thick and thin. Thank you. I am happy you chose me.

Dedication

For Abuela, as promised.

Introduction

Escherichia coli is one of the most widely known and extensively characterized species of bacteria. While most strains of *E. coli* are harmless members of the human intestinal microflora, some strains are pathogenic and cause a wide variety of intestinal and non-intestinal infections. Pathogenic *E. coli* are typically divided into two groups: intestinal pathogenic *E. coli* and extraintestinal pathogenic *E. coli* (ExPEC). The more well-known intestinal pathogenic *E. coli* cause severe diarrhea and are commonly associated with food-borne outbreaks.[1] ExPEC are less well-known, but very common; the largest burden of ExPEC infections stem from urinary tract infections (UTIs), of which they account for 80-90%.[2] In children, UTIs are one of the most prevalent bacterial infections, affecting 2-4% of children and comprising a significant portion of emergency room visits and hospitalizations each year.[3]

The natural history and epidemiology of ExPEC infections is complex. Transmission is thought to be fecal-oral and intestinal colonization is considered a prerequisite to infection.[4] However, ExPEC can asymptomatically colonize the intestine and may never cause illness,[5] resulting in silent transmission events and significant challenges in identifying linked cases. This complexity has become especially vexing as rising rates of antimicrobial resistance make ExPEC increasingly difficult to treat.[6]

Whole genome sequencing (WGS) has been a valuable tool in shedding light on some of the complex epidemiologic dynamics of ExPEC. For instance, WGS-based studies have shown that while ExPEC are immensely genetically diverse, there are a handful of dominant clones — or groups of strains with similar genetic backbones — that have been especially globally successful.[7] Additionally, only a small number of these clones cause the vast majority of antimicrobial resistant ExPEC infections.[8]

E. coli sequence type (ST) 131 was the first ExPEC clone to be characterized as a “high-

risk clone" due to its rapid and seemingly simultaneous global emergence, resistance to multiple classes of antibiotics, and robust virulence.[9,10] This clone burst onto the scene in the early 2000's and was recognized for its association with extended-spectrum beta-lactamase (ESBL) production — which confers resistance to extended-spectrum cephalosporins — among community-associated infections, a previously uncommon combination. Soon after its identification, high-resolution molecular techniques revealed that it was not ST131 generally, but rather a particular subclone of ST131 called *H30* — named for the signature presence of a specific allele of the gene coding the mannose-binding adhesin FimH (*fimH30*) — that was actually harboring the unusual antimicrobial resistance factors.[11]

Among adults in the U.S., *H30* is estimated to comprise about 50% of ESBL-producing *E. coli* infections and between 10% and 20% and of all clinical *E. coli* infections, and has been linked to certain host factors including older age, healthcare contact, compromised hosts, and recent antibiotic use.[11–14] It has also been associated with adverse outcomes such as persistent infections, hospitalization, and sepsis.[14,15] Despite the large burden of UTIs in children, and the recorded dominance of ST131-*H30* among antimicrobial-resistant ExPEC, there is very limited available data about the epidemiology of ST131-*H30* in children.

This dissertation integrates high-resolution molecular data with patient-level epidemiologic data from an existing multicenter case-control study to clarify the molecular epidemiology of the *E. coli* ST131-*H30* clone among U.S. children. The bulk of work about the epidemiology of *H30* has focused on elderly populations and long-term care facilities; this body of work is the first to focus specifically on the epidemiology of this pathogen in a pediatric population.[16–18] In Chapter 1, the overall burden of *H30* as well as patient correlates of *H30* infection are characterized. In Chapter 2, the ability for WGS to identify putative transmission clusters of *H30* among clinical isolates collected from U.S. children is assessed. In Chapter 3, a temporal phylogenetic analysis of *H30* is performed, combining WGS data, patient data, and antimicrobial resistance data, to identify links between *H30's* population structure and patient and antimicrobial resistance characteristics.

Chapter 1

Epidemiology and antimicrobial resistance characteristics of the Escherichia coli ST131-*H30* lineage among U.S. children

This chaper is adapted from a manuscript, co-authored with Dr. Scott Weissman, Amanda Adler, Dr. Veronika Tchesnokova, Dr. Evgeni Sokurenko, Dr. Janet Baseman, and Dr. Danielle Zerr, that has been published in Clinical Infectious Diseases.[19]

1.1 Introduction

Extraintestinal *Escherichia coli*, a common cause of urinary tract and bloodstream infections across all ages, have displayed increasing rates of antimicrobial resistance over the past two decades.[6] This increase has been attributed to the emergence and rapid clonal expansion of *E. coli* sequence type (ST) 131, which has transformed the population structure of extraintestinal *E. coli* infections worldwide.[10,20–22] Molecular epidemiologic studies have shown that a subclone of ST131, termed *H30*, has driven the global dissemination of ST131.[15,23–25] The clonal structure of ST131-*H30* is tightly linked to antimicrobial resistance; the vast majority of *H30* isolates are fluoroquinolone-resistant due to mutations in the *gyrA* and *parC* chromosomal genes (isolates known as *H30*-R or clade C), while nested subclones are additionally associated with the production of CTX-M-type extended-spectrum beta-lactamases (ESBLs) that confer resistance to extended-spectrum cephalosporins (Figure B.1).[15,25–28]

Although *E. coli* ST131-*H30* (hereafter, *H30*) has been recognized as a clone of signifi-

cant public health importance,[10,29] there is a lack of data about its epidemiology in children. Most studies that have included *H30* isolates from children have occurred over short time periods at single centers and have accumulated few *H30* isolates.[13,14,30] Among adults in the US, *H30* is estimated to comprise about 50% of ESBL-producing *E. coli* infections and 10%-20% of all extraintestinal *E. coli* infections, and has been linked to host factors including older age, healthcare contact, local or systemic compromise, and recent antibiotic use.[13,14,23,30,31] Associations with adverse outcomes such as persistent infections, new infections, sepsis, and hospitalization have also been reported in adult populations.[14,15,32] Understanding the epidemiology of *H30* in pediatric populations is important, as its dominance among multidrug-resistant (MDR) extraintestinal *E. coli* makes it a likely culprit of many difficult-to-treat infections in children. Proper treatment of urinary tract infections – the most common type of infection caused by extraintestinal *E. coli* – is especially critical in pediatric populations, as young children are more prone to upper urinary tract infection with potential short- and long-term complications such as renal scarring and decreased renal function.[33,34]

We sought to address this knowledge gap using data from a multiyear, multicenter prospective case-control study of extraintestinal *E. coli* infections to quantify the burden and identify clinical and demographic correlates of infection with *H30* in a U.S. pediatric population. In addition, we describe and compare the antimicrobial resistance characteristics of *H30* and non- *H30 E. coli* isolates.

1.2 Methods

1.2.1 Patients and isolates

All isolates and clinical data came from a multicenter case-control study that prospectively collected isolates and is described in detail elsewhere.[35] In brief, between September 1,

2009 and September 30, 2013, four freestanding U.S. children's hospitals (referred to here as West, Midwest 1, Midwest 2, and East) used standard clinical microbiology techniques to identify and collect all extended-spectrum cephalosporin-resistant (ESC-R) *E. coli* collected from urine or other normally sterile sites during routine clinical care of both inpatient and outpatient children < 22 years of age. ESC-R isolates were defined as those non-susceptible to ceftriaxone, cefotaxime, ceftazidime, cefepime, or aztreonam. Patients could contribute multiple ESC-R isolates if the subsequent isolate was collected ≥ 15 days after the previous ESC-R isolate. For each resistant isolate, three consecutive *E. coli* isolates that were susceptible to the aforementioned agents, referred to here as extended-spectrum cephalosporin-susceptible (ESC-S) isolates, were collected without respect to any patient or microbiological characteristics beyond temporal proximity to the ESC-R isolates and prior enrollment in the study (patients could only contribute one ESC-S isolate). Demographic and clinical data were collected from the medical records; methods for categorizing underlying medical conditions, capturing antibiotic exposure, and characterizing the clinical significance of urine isolates (likely UTI vs. not) were described previously.[35,36] The Institutional Review Board at each hospital approved the study protocol.

1.2.2 Laboratory methods

Methods for antibiotic susceptibility testing and typing of resistance phenotypes and determinants were described previously. Briefly, ESC-R phenotypes (ESBL vs. AmpC) were characterized using a combination of disk diffusion and E-tests. Genetic determinants of extended-spectrum cephalosporin resistance were identified by PCR using primers for genes encoding common extended-spectrum cephalosporinases.[35] *H30* isolates were identified using the *fumC/fimH* genotyping scheme.[37] Isolates belonging to the *H30Rx* sublineage were identified by PCR detection of sublineage-specific single nucleotide polymorphisms.[15]

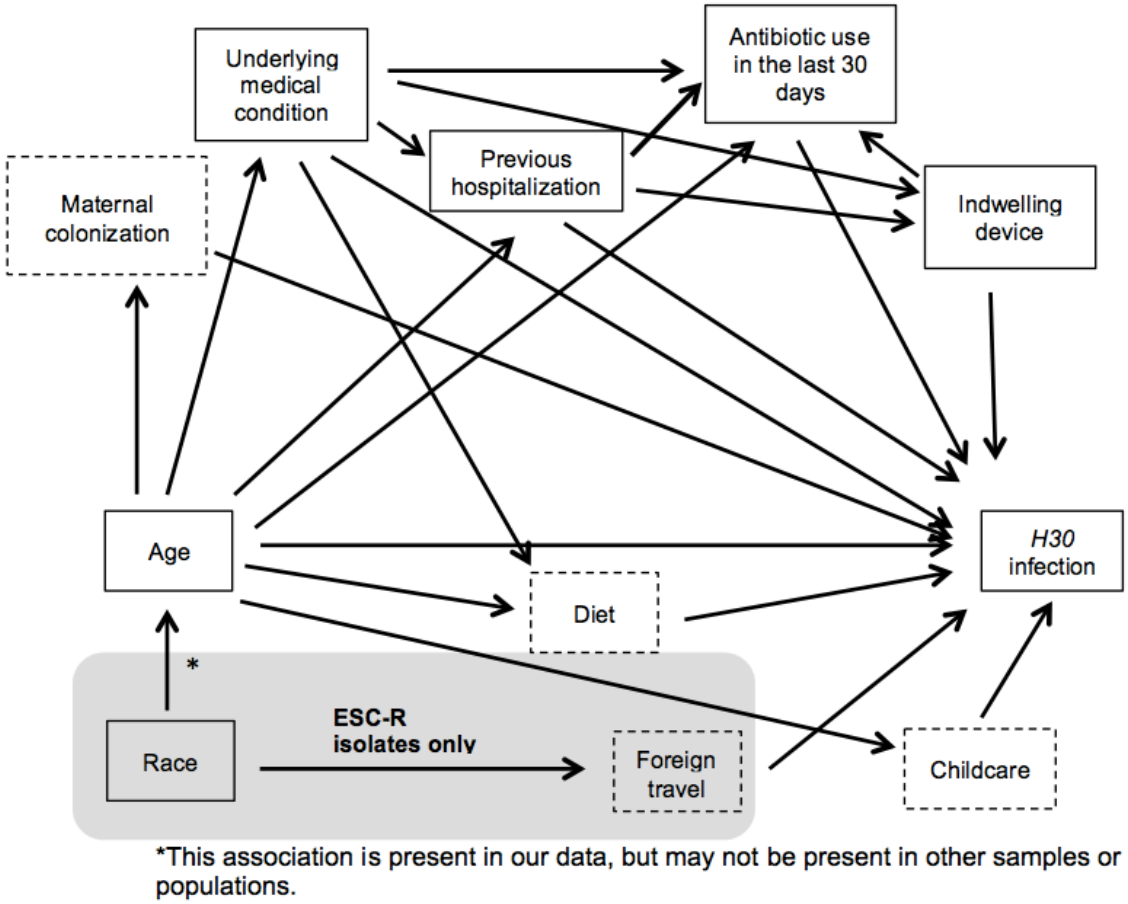

Figure 1.1: Conceptual framework used for building multivariable models of the relationship between host clinical and demographic factors and ST131-H30 infection among children with extraintestinal E. coli infections. Unmeasured variables are surrounded by dashed boxes.

1.2.3 Statistical analyses

1.2.3.1 Prevalence estimates

The period prevalence of *H30* was estimated using the prevalence of *H30* among ESC-R isolates (which were captured completely), the prevalence of *H30* among the collected ESC-S isolates (which were captured partially), and the total number of *E. coli* isolates collected from each clinical microbiology laboratory during the study period, excluding repeat isolates from a given patient if they were collected within 15 days of the first isolate. The 15-day cut point was chosen to reflect the end of a typical course of antibiotic treatment, which is < 14 days. This cut point was prescribed in the original study protocol, and was used by each study site to provide denominator data for the total number of *E. coli* isolates that came through the laboratory during the study period. The total number of *E. coli* isolates from each study hospital was only available between October 1, 2009 and September 30, 2013, so any isolates collected during September 2009 were excluded (Figure B.2). After calculating the prevalence of *H30* among ESC-R and ESC-S isolates separately, the overall prevalence among all clinical *E. coli* isolates was estimated using a weighted average of the stratified prevalence estimates, with the weights being the relative proportions of the two mutually exclusive groups (ESC-R and ESC-S) among the total number of *E. coli* isolates reported. All data used for these calculations can be found in Table A.2. This approach makes two assumptions: 1) all ESC-R isolates were captured; and 2) the collected ESC-S isolates are representative of all ESC-S isolates.

To estimate uncertainty around these prevalence estimates, we used a resampling-based approach to calculate 95% interval estimates. Since our data were sampled to over-represent ESC-R isolates, in our bootstrap analyses, we weighted each ESC-R observation with the prevalence of ESC-R isolates among all isolates in the population, and each ESC-S observation with the prevalence of ESC-S isolates among all isolates in the population (Table A.2). We applied the non-parametric bootstrap procedure using simpleboot::one.boot

in R with 10,000 bootstrap replicates. We reported the bias-corrected and accelerated bootstrap (BCa) interval for these estimates.[38] The same approach was applied to calculate the prevalence and 95% CI for *H30Rx*.

1.2.3.2 Host correlates of infection

Only the first isolate from each unique individual was considered in the host factor analyses. Host factors were compared between patients with *H30* vs. non- *H30* isolates, stratified by ESC-R status and adjusting for study hospital where sample size allowed. Variables were first compared using chi-square tests with continuity corrections. When the strata-specific sample sizes were too small for the chi-square approximation to be valid, Fisher's Exact tests were used. Factors with a p-value of <0.05 were further examined as candidate predictors of interest.

The magnitude of the association between each predictor of interest and *H30* infection was then assessed using univariable and multivariable log-binomial regression models. The log-binomial regression models were implemented using the logbin package in R with the adaptive barrier method selected for maximum likelihood estimation. Log-binomial regression was chosen over logistic regression because the outcome (*H30*) was common among ESC-R isolates; if logistic regression were used, the odds ratio would not approximate the relative risk. For each predictor of interest, the relative risk (RR) and 95% confidence intervals (CIs) from three models are presented: 1) a univariable model that estimates the crude (unadjusted) total effect of the predictor of interest on the outcome; 2) a multivariable model that estimates the total effect of the predictor of interest on the outcome, adjusted for potential confounders; and 3) a multivariable model that estimates the direct effect of the predictor of interest on the outcome, adjusted for potential confounders as well as for potential mediators. All multivariable models adjusted for study hospital; additional potential confounders and mediators were selected according to the conceptual framework illustrated in Figure 1.1 using causal diagram principles.

All covariates were chosen a priori, without regard to associations in the data, except for Asian race. In our data, young age was strongly associated with Asian race only among ESC-R isolates (data not shown), and Asian race also displayed an association with *H30* infection among ESC-R isolates. Therefore, we included Asian race (yes/no) as a potential confounder of the association between patient age and *H30* infection among ESC-R isolates. The conceptual framework in Figure 1.1 highlights the hypothesized mechanism for this observed association.

Finally, we conducted post-hoc analyses of the interaction between age and underlying medical condition. Analyses of the mechanistic interaction between categorized patient age and presence of an underlying medical condition were conducted by examining the stratified and joint effects of these two variables. Multivariable log-binomial models adjusting for study hospital were constructed; the adaptive barrier method was selected for maximum likelihood estimation. To examine the joint effects, a 4 level categorical variable representing the combinations of age and underlying medical condition was included as the predictor of interest. We chose to measure the interaction on the additive scale, as this is typically most useful for assessing the public health importance of interactions. Interaction contrast ratios (ICRs) were calculated to measure the extent of interaction on the additive scale: an ICR of 0 represents no interaction, an ICR <0 indicates a negative interaction, and an ICR > 0 indicates a positive interaction.[39] Preventative factors were recoded as risk factors, and the stratum with the lowest risk was used as the referent category.[40] Uncertainty in the ICR was estimated via a non-parametric bootstrap approach using the boot package in R with antithetic resampling; 1000 replicates were performed and 95% BCa intervals were presented.[38]

1.2.3.3 Antimicrobial resistance characteristics

We examined co-resistance to commonly used antimicrobial agents in the first *E. coli* isolate collected per individual, stratifying by ESC-R and ESC-S status to maintain consistency with

the sampling scheme of the parent study. *H30* isolates were additionally stratified into *H30Rx* and *H30* -non-Rx (Figure B.1) and compared to non- *H30* isolates. Comparisons between *H30* and non- *H30* isolates and *H30Rx* and non- *H30* isolates were quantified using Chi-square tests using the chisq.test function in R; Fisher's Exact tests were used when sample size did not allow for a chi-square approximation. Among ESC-R isolates, ESC-R-associated resistance mechanisms and determinants were also identifed and compared. All analyses were conducted using R version 3.3.1 (R Core Team, 2016).

1.3 Results

1.3.1 Isolates and prevalence estimates

A total of 339 ESC-R isolates from 278 patients and 1008 ESC-S isolates from 1008 patients were available for analyses.(Figure B.2) The estimated prevalence of *H30* among all clinical *E. coli* isolates at all study hospitals was 5.3% (95% CI 4.6%-7.1%), while the hospital-specific prevalence ranged from 2.7% to 6.2%.(Figure 1.2) The estimated overall prevalence of *H30Rx* was 0.87% (95% CI 0.70%-1.7%).

1.3.2 Host correlates of infection by extended-spectrum cephalosporin resistance status

The first ESC-R isolate from each of the 278 patients with an ESC-R isolate collected during the study period was included in the host correlates analyses (Figure B.2). Among these patients, patient age was associated with *H30* infection and further examined as a predictor of interest (Table 1.1). Our sample size precluded multilevel predictors, so age was categorized into ages 0-5 versus 6-21 years in regression models. After adjusting for potential confounders, age 0-5 was associated with an 83% increased risk of the infecting organ-

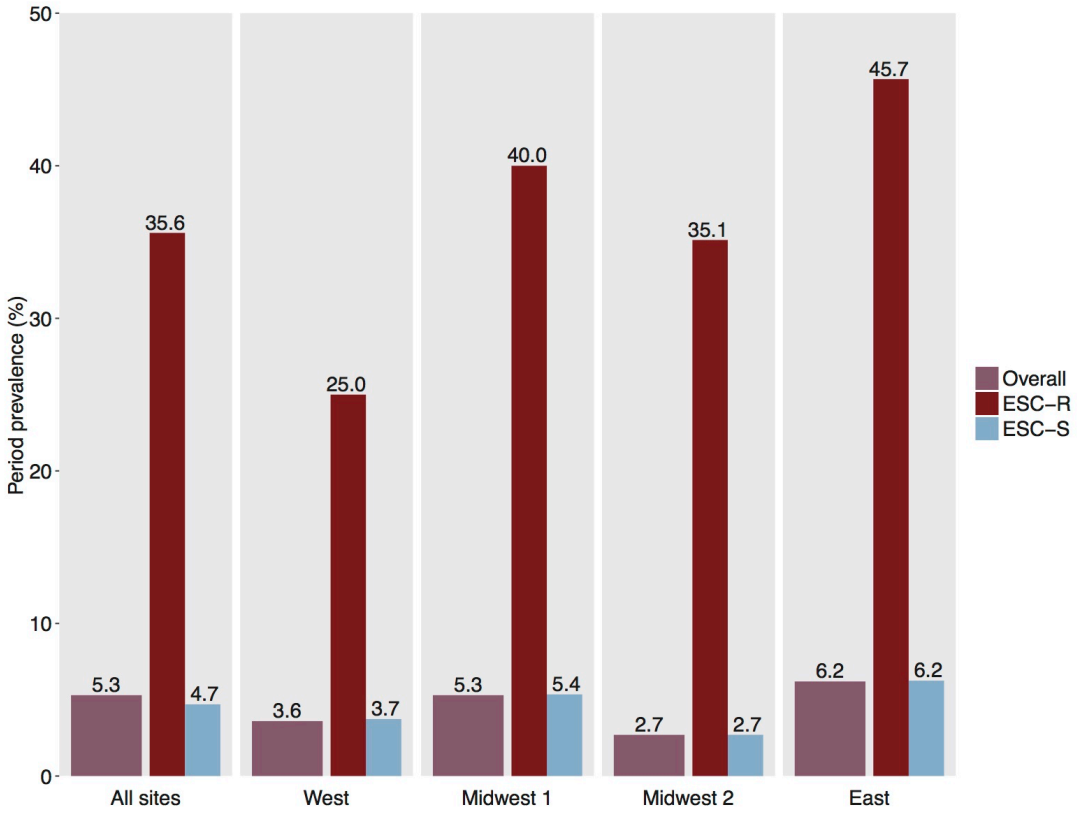

Figure 1.2: Estimated prevalence of ST131-H30 among extraintestinal E. coli infections overall and by study hospital. ESC-R = extended-spectrum cephalosporin-resistant. ESC-S = extended-spectrum cephalosporin-susceptible.

ism being *H30* (RR 1.83, 95%CI 1.19-2.83). There was no evidence that this association was mediated through factors related to underlying illness (Table 1.2), or that underlying illness interacted with age (Table A.3). When restricting the outcome to *H30Rx* infection only (vs. non- *H30* infection) and adjusting for potential confounders, the effect size was stronger (RR 2.25, 95%CI 1.33-3.80).

A total of 1008 patients had one ESC-S isolate collected during the study period. Among these patients, patient age and several factors associated with underlying illness were associated with *H30* infection (Table 1.1). Each of these variables was examined as a predictor of interest except for: (i) history of transplantation, due to small numbers, and (ii) type of infection acquisition, since previous hospitalization and underlying medical conditions were examined independently. Underlying medical condition and indwelling device categories were collapsed into any vs. none. Patient age $\leq$ 5 years was negatively associated with *H30* infection (RR 0.48, 95% CI 0.27-0.87, Table 1.2). Of the variables related to underlying illness, after adjusting for potential confounders, only presence of an underlying medical condition (RR 4.49, 95%CI 2.43-8.31) remained as an independent predictor of *H30* infection. When including potential mediators in the models, the magnitude of the associations between age $\leq$ 5 years and presence of an underlying medical condition with *H30* infection decreased, but the associations remained statistically significant (Table 1.2). Evidence of interaction between age and underlying medical condition was observed; when examining joint effects, underlying medical condition was only significantly associated with *H30* infection in combination with older age, and older age was only significantly associated with *H30* infection in combination with presence of an underlying medical condition (Table 1.3). Since patient age was important in the analyses of both ESC-R and ESC-S isolates, we also visually inspected the distributional differences of age measured continuously. While the non-*H30* age distributions are very similar, the *H30* age distributions display marked differences between ESC-R and ESC-S isolates (Figure 1.3).

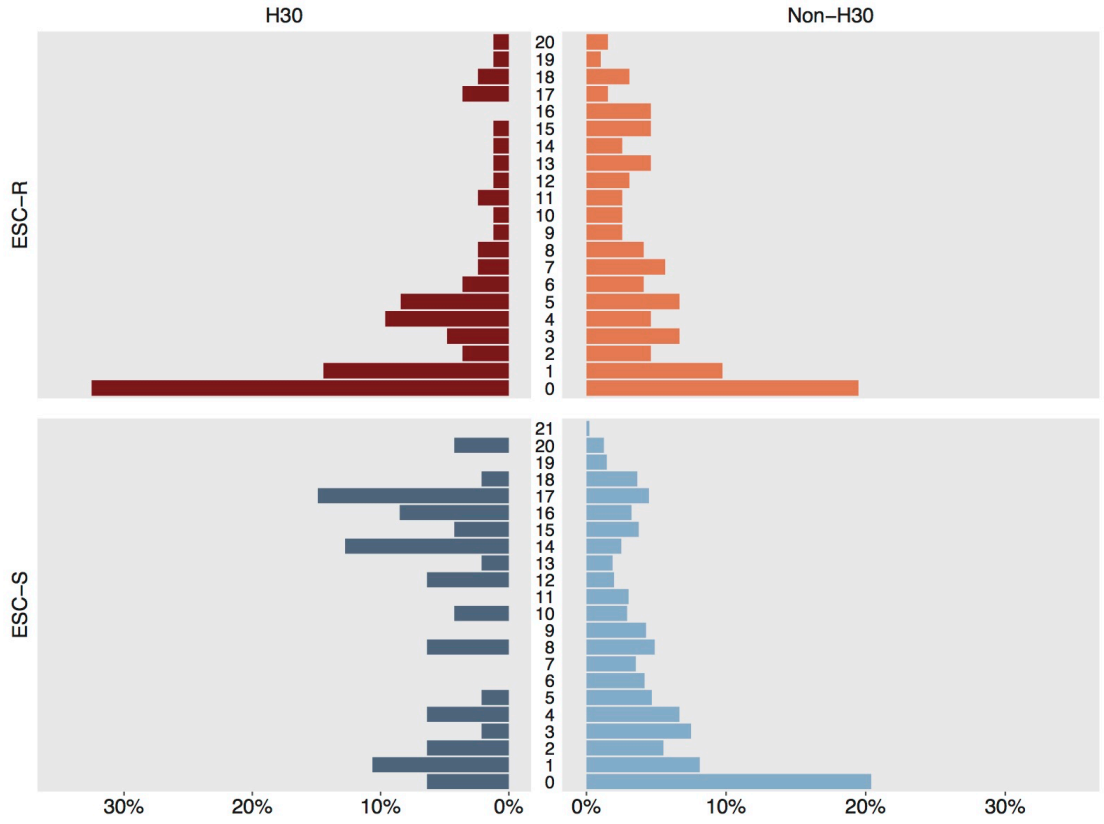

Figure 1.3: Distributions of age (in years) by ST131-H30 and non-ST131-H30 status and extended-spectrum cephalosporin resistance status. ESC-R = extended-spectrum cephalosporin-resistant. ESC-S = extended-spectrum cephalosporin-susceptible.

1.3.3 Antimicrobial resistance characteristics by extended-spectrum cephalosporin resistance and H30-Rx status

A total of 278 ESC-R isolates were examined (the first isolate collected per individual). Among these isolates, nearly all *H30Rx* and *H30*-non-Rx isolates were non-susceptible to fluoroquinolones, compared to less than half of non- *H30* isolates (Table 1.4). Similarly, all ESC-R *H30Rx* and the vast majority of *H30*-non-Rx isolates were ESBL-producing, while non- *H30* isolates were more evenly split between ESBL producers and AmpC producers. *H30* was the most common subclone identified among the ESC-R isolates in the study (Table A.1); it made up 29.9% (83/278) of ESC-R isolates, and when restricting to ESBL-producing isolates only, it made up 43.3% (81/187) of the total. The vast majority of ESBL-producing *H30Rx* isolates had a CTX-M-15 beta-lactamase, while ESBL-producing *H30*-non-Rx isolates were dominated by the CTX-M-27 beta-lactamase; ESBL-producing non- *H30* isolates were more evenly split between CTX-M-15 and CTX-M-14 beta-lactamases (Table 1.4).

Among the 1008 ESC-S isolates examined, fluoroquinolone non-susceptibility was dominant among *H30* isolates, while only a small fraction of non- *H30* ESC-S isolates were non-susceptible to fluoroquinolones (Table 1.4).

1.4 Discussion

We utilized a multiyear, multicenter case-control study of extraintestinal *E. coli* infections in children's hospitals to address a critical knowledge gap about the epidemiology of the globally important ST131- *H30* subclone among US children. Our results can be summarized into three main findings. First, the estimated prevalence of *H30* among pediatric extraintestinal *E. coli* isolates of 5.3% was lower than the 10-20% that has been observed in US adults.[14,23,30] However, *H30* was nearly as dominant among ESBL-producing isolates in children (43.3%) as has been reported in adults (about 50%).[13,31] Second, patient age was

associated with infection due to *H30*, and the nature of this association contrasted sharply between ESC-R and ESC-S infections. Among ESC-R infections, *H30* was associated with young age (≤5 years), while among ESC-S infections, *H30* was associated with older age (6-21 years), as well as with the presence of an underlying medical condition. Third, the antimicrobial resistance characteristics of *H30* and *H30* Rx collected from children were consistent with what has been previously reported.[13,27,31,32,41] ESC-R *H30* isolates were almost always fluoroquinolone-resistant and ESBL-producing, and ESBL-producing *H30* Rx isolates were associated with the CTX-M-15 beta-lactamase, while ESBL-producing *H30*-non-Rx isolates were associated with the CTX-M-27 beta-lactamase.

Other studies have suggested that *H30* is less prevalent among children than adults; however, very few pediatric isolates were included in these studies.[13,30] Interestingly, we observed that *H30* was nearly as dominant among ESBL-producing *E. coli* infections in children as has been reported in adults.[13,31] These findings are consistent with a recent study from a pediatric setting conducted in the Midwestern US.[42] However, in the context of all clinical extraintestinal *E. coli* infections, ESBL-producing organisms are still relatively rare in both adults and children. The bulk of the *H30* isolates circulating in the population are non-ESBL-producing but fluoroquinolone-resistant, and these isolates were much less common in our study than has been observed in adult populations.[13,30] This observation may be explained by differential antibiotic use in these populations. Fluoroquinolones are infrequently prescribed to children due to concerns about toxicity;[43] in our study, about 5% of patients received fluoroquinolones in the year before collection of their first isolate, while 46% of patients received any antibiotic in that same time period (Figure 1.4). Lower rates of fluoroquinolone use likely translate to less selective pressure on fluoroquinolone-resistant organisms such as *H30*. Interestingly, a recent study conducted in adults in Australia and New Zealand, a population that also has low rates of fluoroquinolone use, reported an overall prevalence of *H30* of 3.5%, but a prevalence of *H30* among ESC-R *E. coli* of 39%, which is similar to our findings.[44]

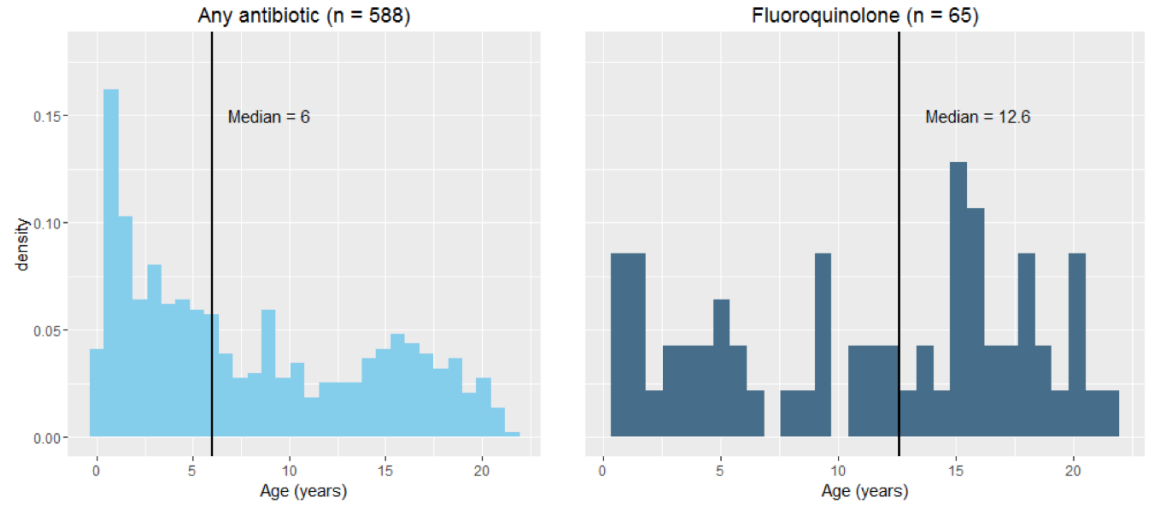

Figure 1.4: Distributions of age (in years) among individuals that received any antibiotic in the year prior to isolate collection compared to individuals that received a fluroquinolone in the year prior to isolate collection.

The association we identified between *H30* and young age among ESC-R isolates is consistent with the findings of a recent longitudinal study showing that among children, the prevalence of ESBL-producing Enterobacteriaceae was highest and increasing most rapidly in children aged 1-5.[45] Why *H30* and *H30Rx* is more frequently found among young children with ESC-R infections compared to older children with ESC-R infections, as well as where young children are acquiring this pathogen, deserves further investigation. Previous studies have portrayed *H30* as an opportunistic pathogen that favors compromised hosts including the elderly,[14] and young children's developing immune systems could be associated with *H30* infection. Maternal infection or colonization may also play a role; a recent study found *H30* colonization during the first several years of life of healthy twins was associated with the mother also being colonized, however, none of these *H30* isolates were ESBL-producing.[46] Finally, while transmission of *H30* between children within healthcare facilities has not been documented, there are reports of transmission of, and persistent colonization with, *H30/ H30Rx* among healthy children within daycares and households.[46–49] Future studies might focus on systematic sampling in the community setting in order to

better elucidate the reservoirs and transmission dynamics of *H30/ H30Rx* among young children.

The association we observed between ESC-S *H30* infections and older children is not consistent with the limited existing data.[30,50] Our post-hoc interaction analyses suggest that age and underlying illness interact, with the strongest risk of an infection being *H30* observed in older children with underlying medical conditions. We hypothesize that these observed associations could be driven by different selective pressures in older, less healthy children: specifically, fluoroquinolones are likely prescribed more frequently to older children than younger children due to less concern about toxicity. This prescribing pattern was borne out in our data; the median age was 12.6 years among patients that received fluoroquinolones in the year prior to their infection, whereas the median age among those that received any antibiotic was 6 years (Figure 1.4). A more refined examination of the role of antibiotic exposure, specifically focusing on fluoroquinolones, is warranted.

Notably, previous studies conducted in adult populations have described *H30* as being associated with healthcare contact and compromised hosts,[14,30] however, we found those associations only among ESC-S *H30* infections. The fact that we observed these patterns among ESC-S isolates is not surprising; compromised hosts and healthcare contact are consistently associated with antimicrobial-resistant infections,[51] and as is shown in Table 1.4, *H30* isolates are more antimicrobial-resistant than other ESC-S isolates. However, we observed that when compared to other ESC-R organisms, there is no evidence of an association between *H30* and underlying illness. This observation raises the question of whether some host correlates observed in previous studies are specific to the *H30* subclone, or just reflect risk factors for MDR extraintestinal *E. coli* in general. Future studies should consider comparing *H30* to other MDR *E. coli* where possible.

A number of limitations need to be considered in the interpretation of these data. First, because of the case-control design of the parent study, the prevalence of *H30* and *H30Rx* among clinical *E. coli* isolates could not be calculated directly. However, we believe

the assumptions employed in our prevalence estimates are reasonable, and that these data provide the best estimate of the prevalence of *H30* in children to date. The design of the parent study was also a strength, as it allowed us to enrich the collection with the less common MDR isolates and examine risk factors for infection with *H30* among those with ESC-R *E. coli* isolates specifically. Second, because this study was an exploratory investigation of an existing dataset, all findings should be interpreted cautiously; there could be residual confounding due to unmeasured or incompletely measured variables, spurious associations identified due to multiple testing, or missed associations due to lack of power. To mitigate this, we attempted to make thoughtful model building decisions and interpretations by using conceptual models rather than taking a purely data-driven approach. Third, the isolates did not undergo multilocus sequence-typing (MLST) or other molecular characterization relevant to *H30* such as typing of the *gyrA* and *parC* alleles. However, the *H30* isolates in this study have since undergone whole genome sequencing, and in silico MLST analyses have confirmed that isolates classified as *H30* are ST131 (data not shown). Finally, although this was a multicenter study, our data were collected from freestanding children's hospitals between 2009 and 2013, so the results may not be generalizable to other settings, and epidemiologic patterns may have shifted during the subsequent several years. Despite these limitations, this study significantly improves our understanding of the impact of *H30* in children, and is one of the most robust examinations of the clinical burden of, and risk factors for, *H30* infections to date.

1.4.1 Conclusion

Although *E. coli* ST131- *H30* is not as prevalent among children as has been reported in adults, perhaps as a result of low rates of fluoroquinolone use in pediatrics, this clone is dominant among ESC-R extraintestinal *E. coli* infections in children. In particular, ESBL-producing *H30* appear to disproportionately affect young children relative to other ESC-R *E. coli*, even when accounting for other underlying host factors. More densely sampled studies

are needed to elucidate the reservoirs and transmission dynamics of this difficult-to-treat pathogen in a pediatric population.

Table 1.1: Selected demographic and clinical characteristics of patients with H30 and Non-H30 isolates, stratified by extended-spectrum cephalosporin resistance status

	ESC-R			ESC-S		
	H30 (n = 83)	Non-H30 (n = 195)	p-value[a]	H30 (n = 47)	Non-H30 (n = 961)	p-value[a]
Age, years			0.008			<0.001
0-5	60 (72.3)	98 (50.3)		16 (34.0)	504 (52.4)	
6-10	10 (12.0)	40 (20.5)		4 (8.5)	190 (19.8)	
11-15	6 (7.2)	31 (15.9)		12 (25.5)	126 (13.1)	
16-21	7 (8.4)	26 (13.3)		15 (31.9)	141 (14.7)	
Sex = Male	18 (21.7)	53 (27.2)	0.417	8 (17.0)	130 (13.5)	0.643
Ethnicity = Hispanic[b]	8 (10.0)	36 (19.3)	0.092	4 (8.9)	135 (14.6)	0.397
Race[b,c]			0.087			0.314
White	39 (49.4)	116 (62.0)		29 (63.0)	629 (68.1)	
Black	12 (15.2)	29 (15.5)		12 (26.1)	219 (23.7)	
Asian	22 (27.8)	32 (17.1)		2 (4.3)	51 (5.5)	
Native American	4 (5.1)	2 (1.1)		1 (2.2)	6 (0.7)	
Pacific Islander	2 (2.5)	7 (3.7)		1 (2.2)	10 (1.1)	
>1 race	0 (0.0)	1 (0.5)		1 (2.2)	8 (0.9)	
Site of culture			0.233			0.753
Urine[c,d]	78 (94.0)	173 (88.7)		45 (95.7)	923 (96.0)	
Blood	2 (2.4)	15 (7.7)		2 (4.3)	32 (3.3)	
Other sterile site	3 (3.6)	7 (3.6)		0 (0.0)	6 (0.6)	
Type of acquisition[b,e]			0.921			<0.001
Community-associated	28 (33.7)	65 (33.3)		14 (29.8)	599 (62.7)	
Healthcare-associated	45 (54.2)	103 (52.8)		30 (63.8)	297 (31.1)	
Hispital-associated	10 (12.0)	27 (13.8)		3 (6.4)	60 (6.3)	
Hospitalized in the past 6 months = Yes[b]	25 (30.1)	69 (35.4)	0.477	13 (27.7)	143 (15.0)	0.032
Underlying medical condition[b]			0.888			<0.001
Urologic[f]	30 (36.1)	75 (38.7)		26 (55.3)	185 (19.3)	
Malignancy	4 (4.8)	13 (6.7)		1 (2.1)	26 (2.7)	
Other condition	16 (19.3)	35 (18.0)		6 (12.8)	104 (10.8)	
No condition	33 (39.8)	71 (36.6)		14 (29.8)	644 (67.2)	
Antibiotic use in the past 30 days = Yes[b]	34 (41.0)	85 (43.6)	0.785	16 (34.0)	176 (18.4)	0.013
History of transplantation = Yes[b]	3 (3.6)	19 (9.8)	0.134	5 (10.6)	22 (2.3)	0.003
Received immunosupressants in the last year = Yes[b,g]	9 (10.8)	37 (19.1)	0.131	6 (12.8)	62 (6.5)	0.167
Device type			0.148			<0.001
Central venous catheter	7 (8.4)	28 (14.4)		3 (6.4)	53 (5.5)	
Foley catheter	6 (7.2)	5 (2.6)		3 (6.4)	11 (1.1)	
Other device	14 (16.9)	26 (13.4)		10 (21.3)	55 (5.7)	
No device	56 (67.5)	135 (69.6)		31 (66.0)	840 (87.6)	
Hospital			0.156			0.349
West	22 (26.5)	78 (40.0)		13 (27.7)	341 (35.5)	
Midwest 2	24 (28.9)	51 (26.2)		16 (34.0)	284 (29.6)	
Midwest 1	11 (13.3)	23 (11.8)		3 (6.4)	108 (11.2)	
East	26 (31.3)	43 (22.1)		15 (31.9)	228 (23.7)	

Note:

Abbreviations: ESC-R, extended-spectrum cephalosporin-resistant; ESC-S, extended-spectrum cephalosporin-susceptible.

[a] P-values generated via chi-square tests unless otherwise indicated.

[b] Number does not add to n because of missing data.

[c] P-values generated via Fisher exact test.

[d] All isolates collected from urine and without missing data were characterized as likely urinary tract infection (UTI); 7 isolates with missing data could not be classified (3 extended-spectrum cephalosporin-resistant and 4 extended-spectrum cephalosporin-susceptible

[e] Type of acquisition was defined as follows: community associated, culture obtained in an outpatient setting or <48 hours after hospital admission from an otherwise healthy patient without hospitalization in the previous 6 months; healthcare associated, culture obtained in an outpatient setting or <48 hours after hospital admission from a patient who had been hospitalized in the previous 6 months and/or had a chronic medical condition requiring frequent healthcare or prolonged/recurrent antibiotic courses; and hospital associated, culture obtained >48 hours after hospital admission or <48 hours after hospital discharge from a patient without signs or symptoms of infection on admission

[f] Diagnoses included in the urologic category are congenital urological abnormality, neurogenic bladder, and vesicoureteral reflux.

[g] Immunosuppressants included antineoplastic agents, high-dose glucocorticoids (less than or equal to 2 mg/kg of body weight), tumor necrosis factor inhibitors, calcineurin inhibitors, and mycophenolate mofetil

Table 1.2: Total and direct effect of selected factors on risk of H30 infection vs. infection with other E. coli types using log-binomial regression models stratified by extended-spectrum cephalosporin resistance status.

	ESC-R			ESC-S		
	Total effect RR (95% CI)		Direct effect RR (95% CI)	Total effect RR (95% CI)		Direct effect RR (95% CI)
	Crude	Adjusted[a]	Adjusted[a]	Crude	Adjusted[a]	Adjusted[a]
Age 0-5 years	1.98 (1.30-3.01)	1.83 (1.19-2.83)[b]	1.91 (1.24-2.96)[c]	0.48 (0.27-0.87)	–	0.52 (0.29-0.94)[d]
Antibiotics in last 30 days	–	–	–	2.18 (1.22-3.91)	1.18 (0.64-2.20)[e]	–
Underlying medical condition	–	–	–	4.46 (2.42-8.21)	4.49 (2.43-8.31)[f]	3.53 (1.73-7.17)[g]
Hospitalization in past 6 months	–	–	–	2.08 (1.12-3.84)	1.22 (0.65-2.30)[h]	1.02 (0.51-2.01)[i]
Presence of indwelling device	–	–	–	3.33 (1.87-5.92)	1.54 (0.78-3.04)[j]	1.53 (0.77-3.01)[d]

Note:
Abbreviations: CI, confidence interval; ESC-R, extended-spectrum cephalosporin-resistant; ESC-S, extended-spectrum cephalosporin-susceptible; RR, relative risk.
[a] All models adjusted for study hospital.
[b] Additional covariates: Asian race (yes/no).
[c] Additional covariates: Asian race (yes/no), underlying medical condition (yes/no), antibiotics in the last 30 days (yes/no), hospitalization in the past 6 months (yes/no).
[d] Additional covariates: underlying medical condition (yes/no), antibiotics in the last 30 days (yes/no), hospitalization in the past 6 months (yes/no).
[e] Additional covariates: age (0–5 or 6–21), hospitalization in the past 6 months (yes/no), underlying medical condition (yes/no), indwelling device (yes/no).
[f] Additional covariates: age (0–5 or 6–21).
[g] Additional covariates: age (0–5 or 6–21), hospitalization in the past 6 months (yes/no), antibiotics in the last 30 days (yes/no), indwelling device (yes/no).
[h] Additional covariates: age (0–5 or 6–21), underlying medical condition (yes/no).
[i] Additional covariates: age (0–5 or 6–21), underlying medical condition (yes/no), antibiotics in the last 30 days, indwelling device (yes/no).
[j] Additional covariates: underlying medical condition (yes/no), hospitalization in the past 6 months (yes/no).

Table 1.3: Analysis of interaction between age and underlying medical condition risk of H30 infection vs. infection with other E. coli types using log-binomial regression models among ESC-S isolates

	Age		
	0-5 years	6-21 years	
	RR (95% CI)[a]	RR (95% CI)[a]	RRs (95% CI)[a] for Age 0-5 vs. Age 6-21 within strata of underlying medical condition
Presence of an underlying medical condition	2.80 (0.90-8.70)	8.66 (3.38-22.2)	0.32 (0.14-0.72)
No underlying medical condition	1.52 (0.51-4.50)	1.0 (ref)	1.51 (0.50-4.53)
RRs (95% CI) for underlying medical condition within age strata	1.99 (0.74-5.33)	8.81 (3.44-22.6)	

Note:
Interaction contrast ratio (ICR), -6.38 (95% confidence interval, -23.5 to -1.15). When interpreting the ICR, deviation from 0 indicates evidence of interaction on the additive scale (see Supplementary Methods).
Abbreviations: CI, confidence interval; RR, relative risk.
[a] RRs adjusted for study hospital.

Table 1.4: Selected antimicrobial resistance characteristics of H30Rx, H30-non-Rx, and non-H30 isolates stratified by extended-spectrum cephalosporin resistance status.

	ESC-R n = 278					ESC-S n = 1008				
	H30 (n = 83)			p-value vs. non-H30[a]		H30 (n = 47)			p-value vs. non-H30[a]	
	Rx (n = 64)	Non-Rx (n = 19)	Non-H30 (n = 195)	Rx	Non-Rx	Rx (n = 5)	Non-Rx (n = 42)	Non-H30 (n = 961)	Rx	Non-Rx
Co-resistance										
Ciprofloxacin	62 (96.9)	18 (94.7)	76 (39.0)	<0.001	<0.001	5 (100)	36 (85.7)	25 (2.6)	<0.001	<0.001
Gentamicin	28 (43.8)	6 (31.6)	73 (37.4)	0.453	0.798	0 (–)	13 (31.0)	34 (3.5)	1.000	<0.001
TMP/SMX	43 (67.2)	15 (78.9)	121 (62.1)	0.555	0.226	1 (20.0)	26 (61.9)	240 (25.0)	1.000	<0.001
TMP/SMX & ciprofloxacin	41 (64.1)	15 (78.9)	64 (32.8)	<0.001	<0.001	1 (20.0)	23 (54.8)	15 (1.6)	0.080	<0.001
All three	19 (29.7)	5 (26.3)	36 (18.5)	0.084	0.374	0 (–)	8 (19.0)	2 (0.2)	1.000	<0.001
ESC-R type				<0.001	0.007					
ESBL only	64 (100)	17 (89.5)	102 (52.6)			–	–	–	–	–
AmpC only	0 (–)	2 (5.4)	88 (45.4)			–	–	–	–	–
ESBL & AmpC	0 (–)	0 (–)	4 (2.06)			–	–	–	–	–
Undetermined	0 (–)	0 (–)	1 (0.5)							
ESBL determinants[c]	n=64	n=17	n = 106							
CTX-M-15	60 (93.8)[b]	3 (17.6)	48 (45.3)	<0.001	0.06	–	–	–	–	–
CTX-M-14	0 (–)	2 (11.8)	44 (41.5)	<0.001	0.037	–	–	–	–	–
CTX-M-27	1 (1.6)	10 (58.8)	1 (0.9)	1.000	<0.001	–	–	–	–	–
CTX-M others	0 (–)	1 (5.3)	7 (6.6)	0.046	1.000	–	–	–	–	–
ESBL SHV	0 (–)	0 (–)	3 (2.8)	0.292	1.000	–	–	–	–	–
ESBL TEM	0 (–)	0 (–)	0 (–)	–	–	–	–	–	–	–
None identified	3 (4.7)	1 (5.3)	4 (3.8)	1.000	0.531	–	–	–	–	–
AmpC determinants[c]	n=0	n=2	n=92							
CMY-2	–	1 (50.0)	79 (96.3)	–	0.277	–	–	–	–	–
DHA	–	0 (–)	2 (2.2)	–	1.000	–	–	–	–	–
FOX	–	0 (–)	2 (2.2)	–	1.000	–	–	–	–	–
None identified	–	1 (50.0)	10 (10.9)	–	1.000	–	–	–	–	–

Note:
Abbreviations: AmpC, AmpC-type-beta-lactamase; ESBL, extended-spectrum beta-lactamase; ESC-R, extended-spectrum cephalosporin-resistant; ESC-S, extended-spectrum cephalosporin-susceptible; TMP\/SMX, trimethoprim\/sufamethoxazole.

[a] P values generated via chi-squared test; Fisher exact test was used when expected frequencies were <5.

[b] One of these isolates had both a CTX-M-15 gene identified as well as a KPC-3 carbapenemase gene, and was resistant to meropenem.

[c] Total exceeds 100\\% as isolates could have >1 determinant identified.)

Chapter 2

Whole genome sequencing of clinical isolates of the *Escherichia coli* Sequence Type 131 *H30* lineage reveals putative transmission clusters among U.S. children

2.1 Introduction

Extraintestinal pathogenic *Escherichia coli* (ExPEC) cause a wide range of non-intestinal illnesses, ranging from uncomplicated urinary tract infection to potentially fatal bacteremia.[52] Unlike intestinal pathogenic *E. coli*, which are commonly associated with outbreaks, ExPEC infections are generally considered sporadic, and tracking the transmission of ExPEC has not historically been a clinical or public health priority. However, the emergence and widespread dissemination of antimicrobial resistant lineages such as sequence type (ST) 131- *H30* , the most common extended-spectrum beta-lactamase (ESBL) producing ExPEC lineage, has brought new interest to understanding the transmission dynamics of these common pathogens.[11] While most research on the transmission of ESBL-producing ExPEC has focused on the older populations,[17] some evidence suggests that children may play an important role in ESBL-producing ExPEC transmission.[53,54]

The transmission dynamics of ST131- *H30* (hereafter, *H30*) are challenging to study. Like other ExPEC lineages, *H30* is capable of asymptomatically colonizing the gut without transitioning to extraintestinal infection.[55] This potential for long-term intestinal colonization likely results in many “silent” transmission events.[42] Whole genome sequencing and

phylogenetic methods have the power to shed light on pathogen transmission dynamics and potentially uncover some of these silent transmission events. Here, we investigated the ability of whole genome sequencing to reveal putative transmission clusters among passively collected clinical isolates of *H30* from children across the U.S.

2.2 Methods

2.2.1 Strain collection and whole-genome sequencing

All isolates and clinical data came from the same multicenter case-control study described in section 1.2.1. Briefly, between September 1, 2009 and September 30, 2013, four free-standing children's hospitals—referred to here as "West," "Midwest 1," "Midwest 2," and "East"— collected *E. coli* isolates during the course of standard clinical care from individuals <22 years old. All extended-spectrum cephalosporin-resistant and a subset of extended-spectrum cephalosporin-sensitive isolates were collected.[35] The Institutional Review Board at each hospital approved the study protocol. *H30* isolates were identified using the *fumC/fimH* genotyping scheme;[37] only the first *H30* isolate per individual was included.

Genomic DNA was extracted from the isolates using the Qiagen QIAamp DNA Mini Kit following instructions from the manufacturer. Starting with 300-500ng of DNA, samples were sheared in a 96-well format using a Covaris LE220 focused ultrasonicator targeting 380 base pair (bp) inserts. The resulting sheared DNA was cleaned with Agencourt AMPure XP beads to remove sample impurities prior to library construction. End-repair, A-tailing, and ligation were performed as directed by KAPA Hyper Prep Kit (Kapa Biosystems, Wilmington, MA). Following ligation, the samples were subjected to an AMPure XP cleanup to remove excess adapters. Samples were then amplified by 5 cycles of PCR using the KAPA HiFi HotStart DNA Polymerase, followed by a final AMPure XP cleanup. Final libraries were quantified by fluorometric assay (Quant-it, Invitrogen) and the molecular weight distribu-

tion of library fragments was checked using the Agilent Bioanalyzer. Barcoded genome libraries were pooled and loaded on the Illumina NextSeq instrument, where cluster generation and paired-end 150bp read sequencing occurred sequentially.

2.2.2 Genomic data quality filtering and pre-processing

The quality of the sequencing reads was assessed using FastQC;[56] Trimmomatic v0.36 was used to trim adapters and low quality bases (quality score of less than 3) at the ends of reads.[57] Combined output from FastQC and Trimmomatic from all samples was visualized using the program MultiQC v1.6.dev0.[58]

2.2.3 Variant detection and alignment construction

Quality filtered and trimmed reads were mapped to the EC958 *H30* reference genome and single nucleotide variants (SNV) were called and filtered using the Snippy v4.2.3 pipeline.[59,60] Default settings were used with the exception of applying a minimum read depth of 5X and a minimum proportion of reads that differ from the reference of 0.75. Sites identified as likely phage regions in the EC958 reference genome by PHASTER were filtered.[61] A pseudo genome alignment, with the identified and filtered variants instantiated into the EC958 reference genome, was then provided to Gubbins v2.3.4 to identify variants in likely recombinant sites.[62] SNVs in likely recombinant sites as identified via Gubbins were masked. Mapping quality and variant quality of the remaining sites was evaluated using Qualimap v2.2.2-dev and bcftools v1.9 and visualized using MultiQC v1.6.dev0.[58,63,64] Remaining sites were concatenated into two SNV-based alignments; one made up of pseudo genomes including monomorphic sites and one with only the concatenated polymorphic sites.

2.2.4 Accessory gene analyses

The quality filtered and trimmed short reads were additionally assembled into contigs using the Shovill pipeline and annotated using Prokka.[65,66] ABRicate was then used to identify acquired antimicrobial resistance genes; as found in the ResFinder database, genes known to be associated with virulence in *E. coli* ; as found in the ecoli_vf repository, and known plasmid replicons; as found in the PlasmidFinder database.[67–70]

2.2.5 Phylogenetic analyses

A maximum-likelihood (ML) phylogenetic tree was constructed from the full pseudo genome alignment using IQ-tree v1.6.7.1, with a general time reversible model of nucleotide substitution, an among-site rate heterogeneity of gamma with 4 categories, and 1,000 bootstrap replicates.[71] The ML-phylogenetic tree was rooted using isolate MW20107-A, which was deemed to be ancestral to the main clade. The output ML-phylogenetic tree from IQ-Tree was visualized using the R package ggtree, and selected patient and bacterial characteristics were mapped onto the tree.[72] Pagel's lambda was used to measure the tendency for more closely related isolates to be isolated from the same collection site.[73] Available patient data about previous hospitalizations, presence of an underlying medical condition, and previous antibiotic use were collected from medical records and mapped onto the putative clusters. Antibiotic resistance phenotypes, measured as previously described, and presence or absence of known resistance and virulence genes and plasmids, captured using the ABRicate, were also mapped onto the putative clusters and summarized descriptively.[35,69]

2.2.6 SNV distance matrix and transcluster

A pairwise SNV distance matrix was constructed from the full pseudogene alignment using the program snp-dists.[74] Pairwise SNV distances were plotted within and between collection sites; selected comparisons between groups of pairwise SNV distances were assessed using one-sided t-tests. The minimum SNV distance between two isolates from discordant collection sites was used to define a threshold for identification of putative transmission clusters. The transcluster package in R was used to estimate the number of uncaptured transmission events separating isolates in each putative cluster.[75] The input into transcluster included a range of mutation rates estimated using BEAST, and a range of transmission rates from existing literature.[53,55,76,77]

2.3 Results

One hundred thirty *H30* isolates were identified out of a total of 1,347 *E. coli* that were screened. Three of the 130 *H30* isolates were determined to be non- *H30* after *in-silico* analysis and one isolate was identified to have been mislabeld, leaving 126 *H30* isolates in the remainder of the analyses. After quality filtering, 3,433 variable sites were identified and included in the whole-genome-based SNV alignment.

There were 7,875 different pairwise comparisons made, with the pairwise SNV distance ranging from 0 to 165 SNVs after quality filtering. We hypothesized that that the mean SNV distance within pairs across discordant sites would be greater than the overall mean SNV distance, but this hypothesis was not supported by the data (one-sided p-value = 0.424, Figure 2.1A). Similarly, we hypothesized that the mean SNV distance within pairs across the most geographically proximate discordant sites (Midwest 1 and Midwest 2) would be less than the mean SNV distance within pairs across geographically distant discordant sites, but this hypothesis was also not supported by our data (one-sided p-value

= 0.9682, Figure 2.1B). The minimum SNV distance between isolates from discordant collection sites was 14 SNVs. Based on the improbability of epidemiologicaly relevant transmission clusters spanning different geographic sites, we chose <14 SNVs as a cutoff for defining a putative transmission cluster. (Figure 2.1A) Using this threshold, eight putative clusters were identified involving seventeen isolates, seven clusters containing two isolates and one cluster containing three isolates. The putative cluster with three isolates (Cluster 1) consisted of one pair separated by 15 SNVs, which was just beyond the selected cutoff, but since the other two pairs within the cluster were separated by <14 SNVs, all three isolates were included in further analyses. The transcluster package in R estimated that there were between 1 and 19 transmission events separating individuals within clusters (Figure 2.2A).

Out of the eight identified putative clusters, documented epidemiologic data associated with four clusters (clusters 2,6,7,8) was consistent with possible direct transmission. Clusters 2 and 6 involved individuals with documented overlapping dates of hospitalization (Figure 2.2B and C), indicating plausible nosocomial transmission events. While isolates from these clusters differed by 12 and 10 SNVs, respectively, the within-cluster difference in isolation dates was 179 and 199 days, so the potential for long-term colonization and within-host evolution to inflate the estimated number of transmission events was high. Cluster 7 and 8 consisted of isolates that differed by 0-1 SNVs after quality filtering and isolates that were collected between 1 and 7 days of one another.(Figure 2.2A) While there was no documentation of overlapping hospitalizations within Cluster 7 or 8, both individuals within Cluster 7 had surgical site infections associated with neurological procedures, while both individuals within Cluster 8 were paraplegic. These connections are consistent with a plausible epidemiological link in inpatient or outpatient care, although conclusively establishing such a link is outside the scope of this data.

Although visual inspection of the maximum-likelihood phylogenetic tree did not reflect obvious associations between the phylogeny and collection site, a test of phylogenetic

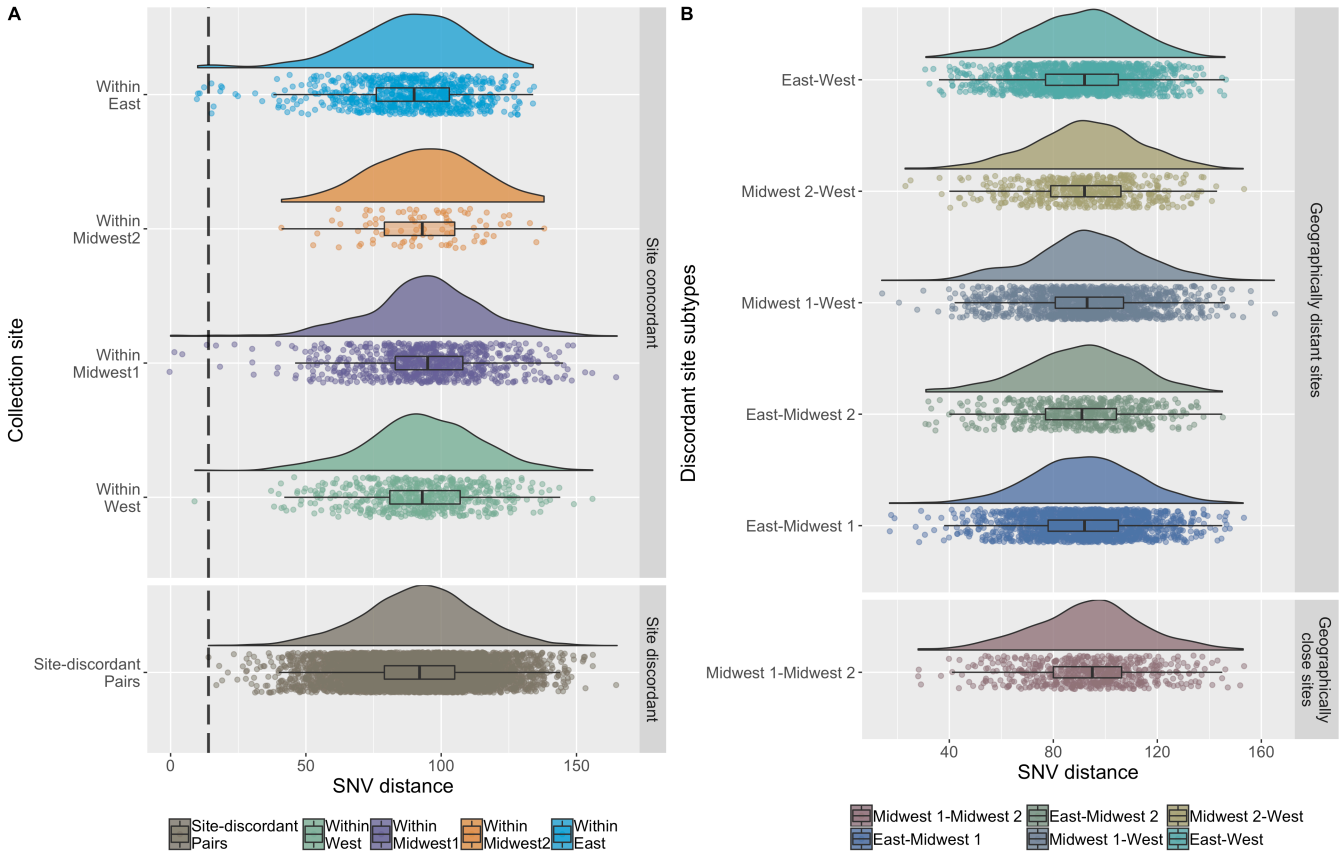

Figure 2.1: Distributions of pairwise single nucleotide variant (SNV) differences between H30 clinical isolates from 4 children's hospitals in the U.S. for (A) pairs containing two isolates from the same (concordant) collection site and pairs containing two isolates from two different (discordant) collection sites and for (B) pairs from different combinations of discordant collection sites. The black dashed line in A indicates the selected SNV threshold for defining putative transmission clusters (<14 SNVs), based on the observation that this was the smallest number of SNVs recorded between isolates from discordant collection sites.

signal of collection site rejected the no-signal model, suggesting that isolates from the same collection site are more likely to be more closely related than isolates from different collection sites (Figure 2.4, p = 8.36e-06). However, when excluding the 17 isolates identified to be included in putative transmission clusters, there was no longer evidence to reject the no-signal model (p = 1.00). Finally, visual inspection of accessory gene data revealed that while most clusters had consistent accessory gene profiles, there were some differences. (Figure 2.3)

2.4 Discussion

Using clinical isolates collected from 4 free-standing children's hospitals over 4 years, we identified putative transmission clusters of *H30* isolates among U.S. children, including two clusters with documented overlapping hospitalization dates and two clusters with other potential epidemiologic links. We also observed that those isolates identified to be part of putative clusters appeared to account for the observed association between collection site and phylogenetic relatedness in these data, and that beyond the most closely related isolates, more closely related isolates were not more likely to be isolated from the same collection site than from a different collection site.

The available data about the transmission dynamics of *H30* is limited. Most evidence is based off of point prevalence studies of colonization in single sites, and to our knowledge, no studies used phylogenetic methods to assess putative transmission clusters. However, available evidence does suggest that household transmission is more common than nosocomial transmission, with the potential exception of rehabilitation and long-term care facilities populated by elderly individuals.[17,55] Interestingly, all documented cases of household transmission involve children, including several documented instances of mother-to-baby transmission.[49,78] While our study was not set up to identify community/household transmission, the limited evidence of nosocomial transmission across four sites over four years

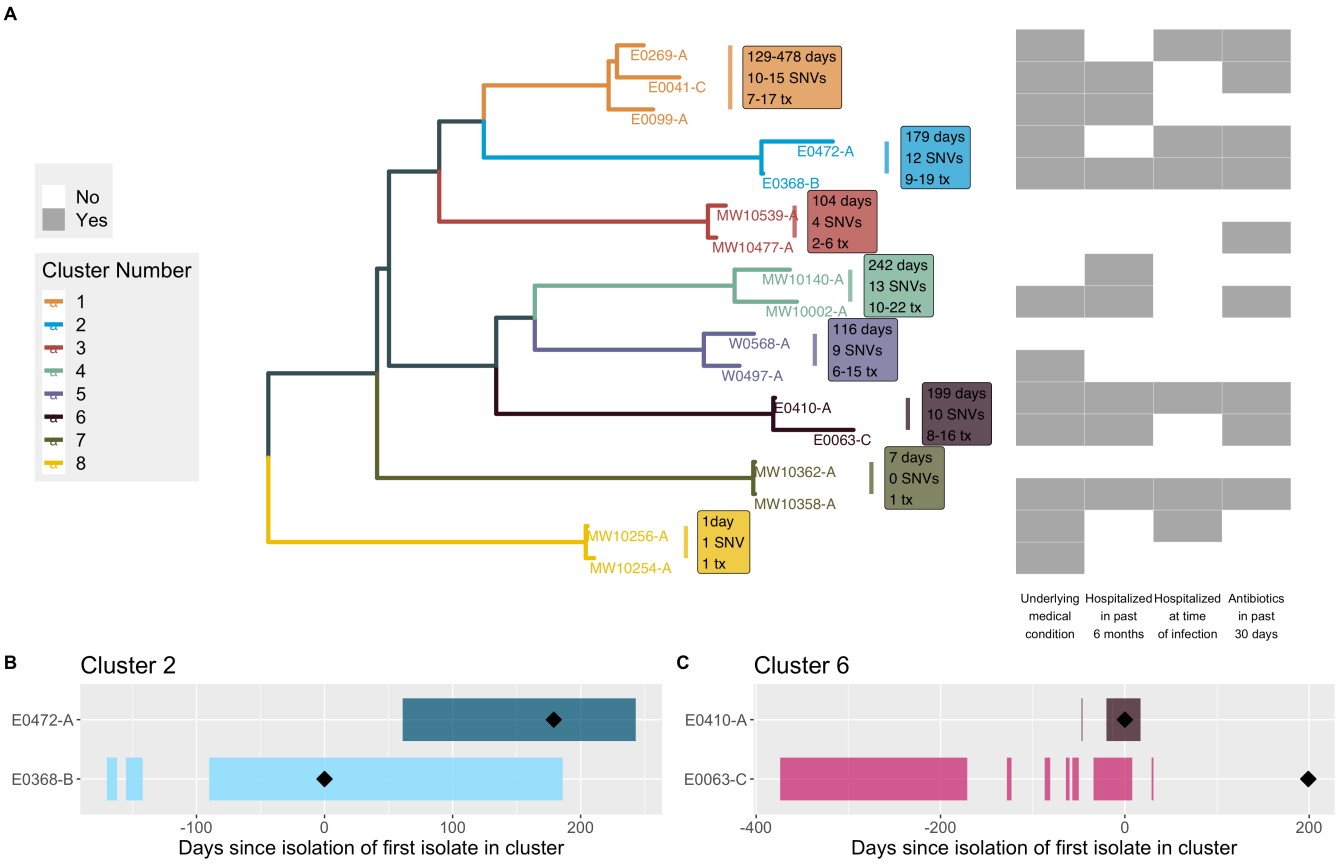

Figure 2.2: A) Unrooted maximum-likelihood phylogeny of 8 identified putative transmission clusters annotated with i) the number of days separating collection of the isolates within a cluster ii) the number of single nucleotide variants separating isolates and ii) the estimated number of transmission events separting isolates, calcualted using the R package transcluster. The presence or absence of selected patient characteristics are mapped to the right of the phylogeny. tx = transmission, SNV = single nucleotide variant. B and C) Schematic of documented hospitalization-days in the one year prior to isolate collection up until the discharge of the admission associated with the isolate collection, if any, for clusters 2 and 6, respectively. Black diamonds represent the day of H30 isolate collection.

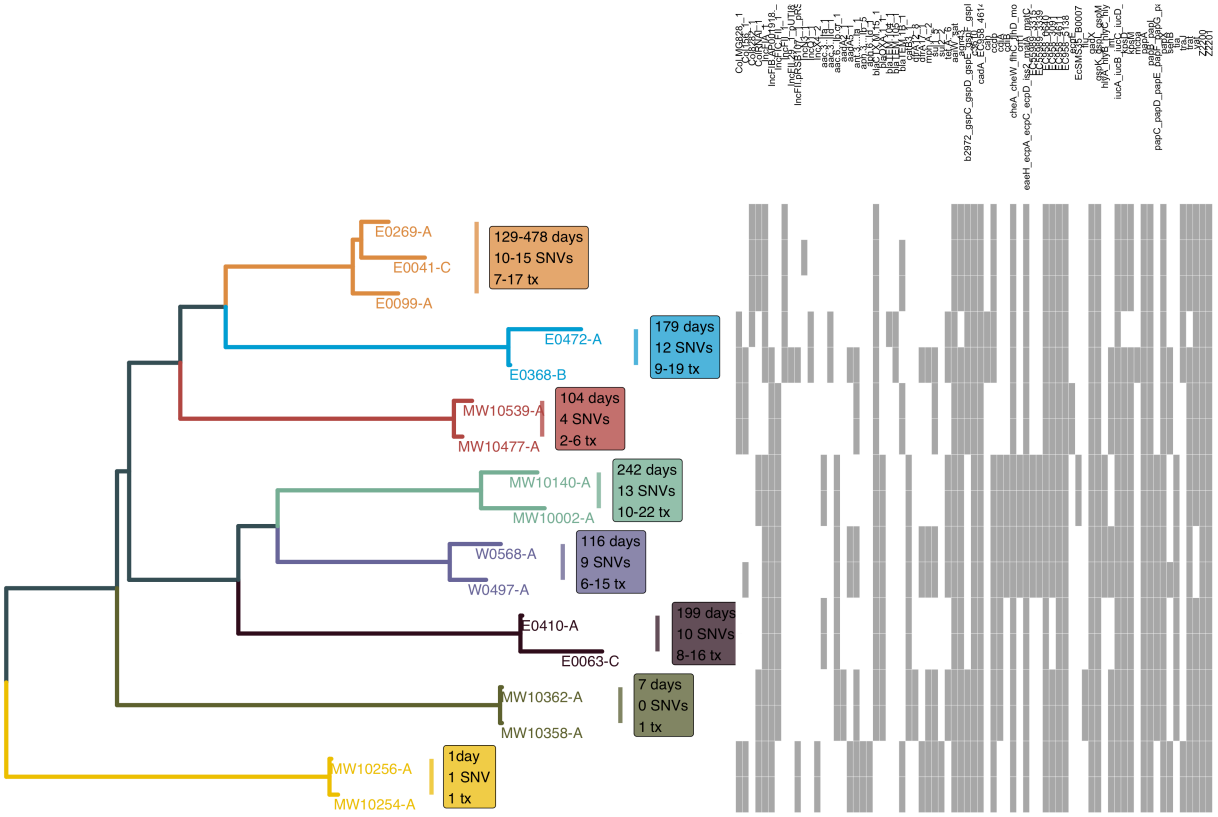

Figure 2.3: Unrooted maximum-likelihood phylogeny of 8 identified putative transmission clusters annotated with i) the number of dates separating collection of the isolates within a cluster ii) the number of single nucleotide variants separating isolates and ii) the estimated number of transmission events separting isolates, calcualted using the R package transcluster. The presence of plasmids identified from the PlasmidFinder database, acquired antimicrobial resistance genes identified from the ResFinder database, and virulence factors identified from the ecolivf database are mapped to the right of the phylogeny. Virulence factors that occurred together one hundred percent of the time in the larger dataset are lumped together into one column. tx = transmission, SNV = single nucleotide variant.

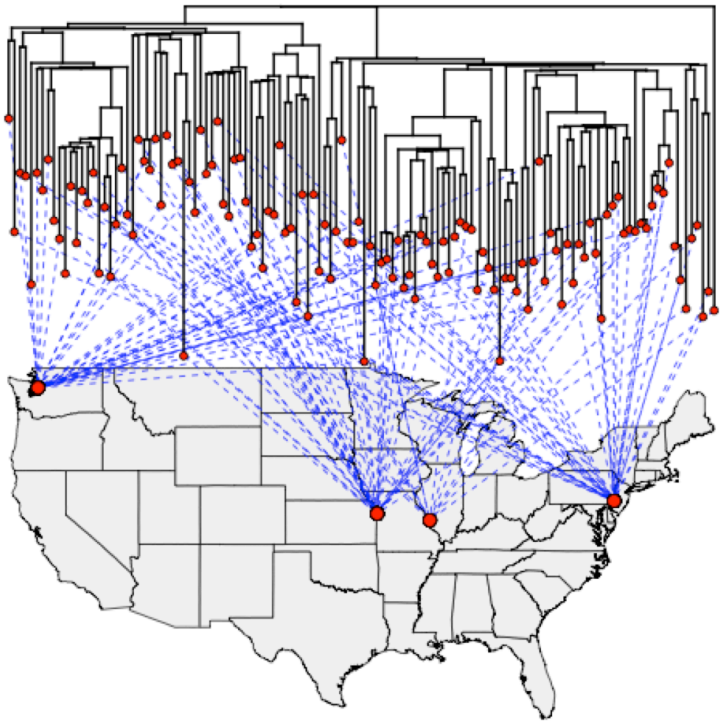

Figure 2.4: Maximum-likelihood phylogeny of all 126 ST131-H30 isolates mapped to their respective collection sites. Visual inspection does not reveal obvious clustering, but there is a statistically significant phylogenetic signal between collection site and the phylogeny as measured by Pagel's lambda (p = 8.36e-06)

is consistent with the idea that nosocomial transmission of *H30* is not the predominant source of spread of *H30* among U.S. children. The relative importance of person-to-person transmission of *H30* generally, compared to other possible mechanisms of spread such as via food or other community-based reservoirs is still unclear.[79] Future work with more detailed collection of epidemiologic context, including data from commuinity-based contacts, are needed to address this question.

The multicenter design of this isolate collection allowed us to explore the local vs. national transmission dynamics of *H30*. Given the fine-scale phylogenetic analysis we conducted, we hypothesized that there would be evidence of clustering by collection site reflected in the phylogeny as a result of local circulation of strains. However, the lack of evidence of clustering by collection site beyond the identified putative clusters, which represent only the lower tail of the distribution of SNV distances, is consistent with the the conceptualization of *H30* as a nationally disseminated strain. This observation supports the idea that this and future instances of identification of closely related *H30* isolates, perhaps within the 14 SNV threshold identified here, are likely to be connected by direct or indirect transmission rather than solely reflecting differential local circulation of strains between sites.

Overall, these results provide proof of principle that—at least among the included population of U.S. children—targeted whole genome sequencing of clinical ExPEC isolates can reveal potential transmission clusters. If used in real-time, phylogenetic analyses of closely related isolates could guide contact sampling and collection of epidemiologic data. This would provide much-needed detail regarding the transmission dynamics of epidemic lineages of ExPEC such as *H30*, in particular, the role of person-to-person transmission. In addition, modern methods like the probabilistic transmission estimation approach used here, which explicitly incorporate sampling date, mutation rate, and transmission rate in estimating transmission events from whole genome sequencing data, have the potential to be especially useful for clarifying the transmission dynamics of pathogens that spread

silently like ExPEC. Improving our understanding of the transmission dynamics of *H30* is an essential prerequisite to evaluating the utility of interventions for infection prevention via transmission interruption.

The results of this study should be interpreted in the context of multiple limitations. First, the available epidemiologic data was limited— including a lack of detail about the location of specific wards during overlapping hospitalizations— and, as such, all observations of plausible transmission should be interpreted cautiously. Future investigations that occur closer to the time of infection could use this "plausible" classification as an impetus for further data collection. Second, the selection of a SNV cutoff is dependent on multiple analytical decisions that are hard to standardize. Despite this, the multicenter design of the available data provided a unique opportunity to define a conservative cutoff where even indirect transmission was epidemiologically unlikely, which we believe to be a reasonable approach given the nature of this investigation and the limited transmission data available about *H30*. Our use of probabilistic transmission event estimation methods also provided context about the epidemiologic relevance of the clusters by translating SNV differences to estimated transmission events. This study also had several strengths. Our multicenter design included a large number of *H30* isolates and used whole genome sequencing, the highest resolution technology available for strain typing, as well as cutting-edge statistical approaches to define putative transmission clusters. We were thus able to demonstrate the utility of modern molecular epidemiologic approaches to address an important public health issue in an understudied pediatric population.

As antimicrobial resistance rates among ExPEC rise, there is new urgency to improving our understanding of the transmission dynamics of these common pathogens. While there are many ExPEC lineages in circulation, multidrug resistance is concentrated among only a handful of dominant lineages, including *E. coli* ST131-*H30*. Using whole genome sequencing of passively collected *E. coli* ST131- *H30* clinical isolates from US children, we demonstrated the ability to reveal putative transmission clusters and plausible nosocomial

transmission events. Targeted genomic surveillance of ExPEC isolates collected during the course of standard clinical care could help address the vast knowledge gaps about the transmission of worrisome ExPEC lineages like *H30*.

Chapter 3

Associations between the population structure of the *Escherichia coli* sequence type 131 *H30* lineage, patient characteristics, and antimicrobial resistance among U.S. children: a phylogenomic analysis

3.1 Introduction

Extraintestinal pathogenic *E. coli* (ExPEC) cause urinary tract, bloodstream, and other non-intestinal infections and are responsible for substantial morbidity and mortality across all ages.[52] In the past two decades, increasing rates of antimicrobial resistance have complicated the treatment of ExPEC infections.[6] Multiple molecular epidemiologic surveillance studies have implicated the proliferation of a particular lineage of ExPEC — known as sequence type (ST) 131- *H30* or Clade C — in the rising rates of resistance to extended spectrum cephalosporins observed among ExPEC.[11,20,22,41] In addition to being responsible for between 40 and 50% of extended-spectrum cephalosporin resistant infections in both adults and children, ST131- *H30* is nearly ubiquitously resistant to fluoroquinolones and commonly resistant to trimethoprim-sulfamethoxazole (TMP-SMX) and gentamicin.[31,80] It is also known to possess more virulence-associated factors than other multidrug resistant ExPEC.[13,16]

These factors, along with ST131- *H30's* (hereafter, *H30*) rapid global dissemination, have earned it the label of an epidemic lineage warranting thorough examination. High-resolution study of *H30* can both enhance our ability to identify the emergence of future

similar lineages and to develop interventions to prevent contemporary *H30* infections. Progress has been made on the former: several previous studies have used whole genome sequencing (WGS) and phylogenomic methods to elucidate the details of the evolutionary history of ST131 and its *H30* sub lineage, including estimating the timing, location, and gene acquisition events associated with its emergence.[15,25,26,28,81] However, many foundational epidemiologic questions relevant to the latter goal remain unanswered. For example, *H30* is known to cause both community and healthcare-associated infections, but it is still unknown whether *H30* emerged in healthcare or in the community, and where it currently continues to predominantly spread.[82] Additionally, while genetic diversity within *H30* has been observed, it is unknown if certain subtypes are associated with heterogeneity in patient characteristics such as patient age.

When integrated with epidemiologic data, WGS of pathogen genomes offers a powerful approach for exploring these epidemiologic questions. Such data integration has been shown to shed light on the the spread of other pathogens between sexual and social networks and different types of hosts.[83–85] However, this approach has not yet been applied to *H30*. In this study, we combine WGS and patient data to explore some of these outstanding questions about the epidemiology of *H30* in U.S children. In particular, we seek to address the community vs. healthcare dynamics of *H30* and to identify whether particular *H30* subtypes are associated with different age groups. Additionally, we delve into the evolutionary dynamics of resistance to TMP-SMX, an antimicrobial agent that is especially important in children.

3.2 Methods

3.2.1 Isolate collection, whole genome sequencing, and bioinformatic methods

All isolates and clinical data came from the same multicenter case-control study described in section 1.2.1. Briefly, between September 1, 2009 and September 30, 2013, four free-standing children's hospitals—referred to here as "West," "Midwest 1," "Midwest 2," and "East"— collected *E. coli* isolates during the course of standard clinical care from individuals <22 years old. All extended-spectrum cephalosporin-resistant (the original cases) and a subset of extended-spectrum cephalosporin-sensitive isolates (the original controls) were collected.[35] The Institutional Review Board at each hospital approved the study protocol. *E. coli* ST131-*H30* isolates were identified using the *fumC/fimH* genotyping scheme;[37] only the first ST131-*H30* isolate per individual was included. Those isolates identified as ST131-*H30* underwent whole genome sequencing as described in section 2.2.1. Short reads were quality filtered, mapped to a *H30* reference genome, and single nucleotide variants (SNVs) were called as described in section 2.2.2 and section 2.2.3 to create a SNV-based core-genome alignment. Short reads were also assembled into contigs, and known acquired antimicrobial resistance genes, plasmid replicons, and virulence factors were identified as described in section 2.2.4.

3.2.2 Collection of patient data

Selected patient data was collected from medical records as previously described.[35] Factors that were central to these analyses included patient age, as well as patient factors associated with classifying infections as healthcare-associated vs. community-associated: presence of an underlying medical condition, record of hospitalization in the 6 months preceding isolate collection, hospitalization at the time of isolate collection, and record of antibiotic use

in the 30 days preceding isolate collection. Phenotypic antimicrobial resistance was also evaluated as previously described.[35]

3.2.3 Comparing population structure to patient and antimicrobial resistance characteristics

In order to generate meaningful proportions of selected patient and antimicrobial resistance characteristics by *H30* clades, the sampling scheme of the original case-control study had to be accounted for. We were able to estimate the expected proportion of ESC-R isolates in the overall population of *H30* clinical isolates to be 12%; this estimate was generated using data provided by each study site about the total number of extraintestinal *E. coli* collected during the study period (Table A.2). Using this estimate, we created 9 "downsampled" datasets of 49 isolates each where ESC-R isolates were randomly sampled to the estimated unbiased prevalence of 12%. These data were used for all formal comparisons of characteristics by *H30* clade. Categorical variables were compared using chi square tests with simulated p-values, and continuous characteristics were compared using two-sample t-tests.

3.2.4 Temporal phylogenomic analyses

In order to assess whether the sequence data contained sufficient temporal signal to conduct molecular-clock based phylogenomic analyses, a linear regression of root-to-tip distance and sampling date was performed using R scripts available in Murrary et al.[86] The statistical significance of the temporal signal was assessed by permuting sampling dates 1000 times and plotting the distribution of correlation coefficients; this was also undertaken using scripts from Murray et al.[86] Additionally, a pairwise linear distance regression was run to assess the correlation between difference in sampling date and pairwise SNV distance.

Temporal phylogenomic analyses were carried out using BEAST v1.10.4.[87] A general

time reversible nucleotide substitution model was used with a gamma model of rate heterogeneity. A strict molecular clock was selected and a nonparametric Bayesian Gaussian Markov random field skyline coalescent model was used to allow for variation in the population dynamics over evolutionary time.[88] Markov chain Monte Carlo chain lengths for each model were 100 million with sampling every 10,000 steps. The Stamatakis correction for ascertainment bias was applied to take into account the inclusion of variable sites only as input.[89] The program Tracer v1.7 was used to evaluate MCMC chain convergence and to examine marginal posterior distributions of parameters; 10% of the chain was removed as burn-in.[90]

Discrete traits, including selected patient characteristics and antimicrobial resistance factors, were included in the BEAST analyses for ancestral state reconstruction using an asymmetric trait evolution model.[91,92] Robust stochastic counting methods were applied to count the number of transitions between selected traits over evolutionary time along with 95% highest posterior density (HPD) intervals.[93,94] A maximum clade credibility tree was created using Treeannotator v 1.10.4 with 10% of the trees discarded as burn-in.[95] This tree was used as the summary tree in all tree visualizations, which were constructed using the R package ggtree.[72]

Since discrete trait analyses are known to be sensitive to sampling effects, we set out to assess the robustness of the trait mapping results. First, discrete trait mapping analyses were repeated on the previously described datasets that were downsampled for ESC-R status and changes in the transition counts and 95% HPD intervals were visualized. In addition, since the isolates were clinical *E. coli* isolates and the collection sites were tertiatry care hospitals, we anticipated our collection was enriched for isolates collected from individuals with underlying medical conditions. To address the impact of this potential sampling bias, we assessed the sensitivity of healthcare-associated results to overrepresentation of individuals with underlying medical conditions. We first took the same approach as described above with the overrepresentation of ESC-R isolates: 9 subsampled datasets

were created where the prevalence of underlying medical conditions was reduced by half to a prevalence of 29%. We then repeated the discrete trait mapping analyses as previous described and summarized mean and 95% HPD intervals for transition counts. Given the results of these analyses, we additionally employed a novel alternate ancestral state reconstruction and transition counting approach in BEAST2 v2.5.0 called BASTA. BASTA is based on an approximation of the structured coalescent and has been demonstrated to be less sensitive to sampling bias.[96,97]

3.3 Results

3.3.1 Summary of patient characteristics and phenotypic antimicrobial resistance

In total, 130 *E. coli* isolates identified as ST131- *H30* from unique individuals underwent WGS. Three of the 130 isolates that were identified as *H30* were determined to be misclassified due to a recombination event at the *fimH* locus and were excluded from further analyses. One isolate did not meet inclusion criteria of coming from a unique individual and was also excluded from further analyses, resulting in 126 remaining isolates. Overall, the majority of individuals were female (78.3% in ESC-R and 81.4% is ESC-S), the vast majority of isolates originated from urine (94% in ESC-R and 95.3% in ESC-S), and antimicrobial resistance to other classes of antibiotics was common. (Table 3.1)

3.3.2 Phylogenomic analysis of ST131-H30 in children

After initial quality filtering, 13,770 variant sites were identified out of the roughly 5.1 Mb EC958 reference genome. Recombination analyses estimated that 10,005 (72.7%) of these SNVs were in putative recombinant regions, while an additional 332 SNVs were identified

Table 3.1: Selected patient characteritics and phenotypic antimicrobial resistance by extended-spectrum cephalosporin resistance status (only includes one isolate collected per individual).

	ESC-R (n = 83)	ESC-S (n = 43)
Patient characteristic		
Age (years)	4.9	10.7
Male	18 (21.7)	8(18.6)
Site of infection		
Blood	2 (2.4)	2 (4.7)
Urine	78 (94)	41 (95.3)
Other sterile site	3 (3.6)	0
Underlying medical condition	50 (60.2)	31 (72.1)
Hospitalization in the previous 6 months	25 (30.1)	13 (30.2)
Antibiotic use in the previous 30 days	34 (41)	15 (34.9)
Hospitalized at the time of infection	10 (12)	2(4.7)
Non-susceptibility		
Ciprofloxicin	80 (96.3)	40 (93)
Trimethoprim-sulfamethoxazole	58 (69.9)	25(58.1)
Gentamicin	34 (40.1)	12 (27.9)

Note:
Abbreviations: ESC-R, extended-spectrum cephalosporin-resistant; ESC-S, extended-spectrum cephalosporin-susceptible.

to be in phage-associated regions not captured by recombination analyses, resulting in non-recombinant core SNV count of 3,433 sites. A phylogeny constructed from unfiltered sites had several topological changes when compared to the recombination filtered phylogeny (Figure B.5).

Temporal signal analyses demonstrated that there was sufficient clock-like signal in the data to proceed with analyses in BEAST. BEAST analyses produced an estimated mutation rate of of 3.33 x 10^{-7} substitutions per site per year (95% highest posterior density [HPD] interval 2.23 x 10^{-7} to 4.57 x 10^{-7}). The previously defined clades C1 and C2 were readily observed in the maximum clade credibility phylogeny. One isolate was consistent with membership in clade C0, a previously identified intermediate clade.[26] That isolate is highlighted in the phylogeny but was grouped with clade C1 for all epidemiologic comparisons.(Figure 3.1) BEAST estimated that the divergence of clade C0 from C1 and clade C1 from C2 were temporally very close in the early 1980s; the large overalp in the HPD in-

tervals precluded determination of which divergence event occurred first using these data. (Figure 3.1) The predicted population growth trajectory was estimated to be fairly constant before the late 1970's, with rapid growth from the late 1970's to the late 2000's, followed by a slight decline by 2013. (Figure B.6)

3.3.3 Association between population structure and patient factors

Visual inspection of patient characteristics mapped onto the *H30* phylogeny revealed potential associations between clade, patient age, and presence of an underlying medical condition (Figure 3.1). When formally comparing the distribution of patient characteristics in the datasets downsampled for ESC-R status, older patient age and presence of an underlying medical condition remained associated with clade C1 (Figures 3.3 and 3.2). No significant associations were observed between previous hospitalization and clade (Figure 3.4). The low prevalence of patients that were hospitalized at the time of infection or received antibiotics in the previous 30 days precluded formal comparisons.

To explore the healthcare-associated vs. community-associated origins of *H30*, we carried out discrete trait stochastic mapping analyses in BEAST. Initial results based on the complete dataset suggested that the ancestral *H30* isolates were more likely to be isolated from individuals with underlying illness than healthy individuals, (Figure 3.5A) while they were less likely to be collected from individuals that were recently or currently hospitalized (Figure B.7A). These results persisted when repeating the analyses on the datasets downsamped for ESCR status. (Figure 3.5C(ii) and Figure B.7A) However, when assessing the sensitivity of these results to overrepresentation of individuals with underlying medical conditions by repeating the analysis on datasets downsampled for individuals with underlying medical conditions, the observed association with underlying medical condition flipped directions: the ancestral isolates in these analyses were predicted to be from healthy individuals, and there were more predicted transitions from healthy individuals to those with underlying medical conditions.(Figure 3.5C(iii)) An alternative analysis based on the

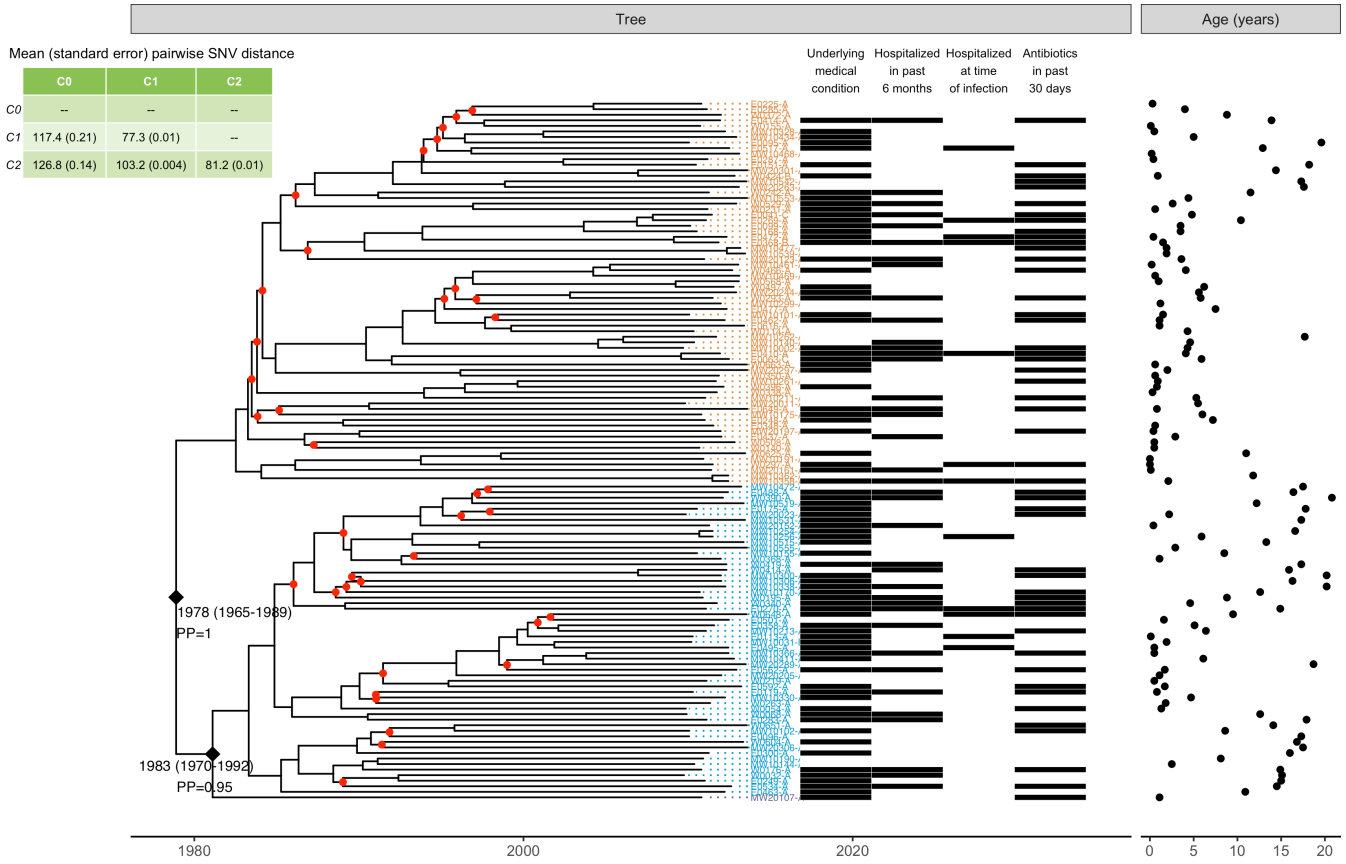

Figure 3.1: Maximum clade credibility tree of H30 based on BEAST analyses using a strict molecular clock, GMRF skyline tree prior, and general time reversibile nucleotide substitution model. The tips of the tree are constrained by isolate dates and the time scale is shown at the bottom of the tree. The tip labels of the tree are colored according to H30 clade, with orange representing clade C2, blue representing clade C1, and purple representing the intermediate clade C0 isolate. Red circles at selected nodes indicate nodes where the posterior probabilty was less than 75 percent. A table in the upper left hand corner provides the mean pairwise SNV distance separating isolates in each clade. Black diamonds mark divergence dates of interest with estimated node ages and 95 percent highest posterior density intervals of node ages, along with posterior probability support for that node. The presence or absence of selected patient characteristics are mapped to the right of the phylogeny, with black indicating presence. Patient age in years is also mapped to the right of the phylogeny. SNV = single nucleotide polymorphism, PP = posterior probability.

structured coalescent using the full dataset was consisent with the analyses downsampled for underlying medical condition.(Figure 3.5B) Together, these findings suggest that the results from the traditional discrete trait stochastic mapping analysis were likely strongly influenced by oversampling of individuals with underlying medical conditions, and that the findings of the results based on the novel analytic approach based on the structured coalescent, which suggested a community-based origin of *H30* ancestors, were likely more reliable.

3.3.4 Association between population structure and antimicrobial resistance to TMP-SMX

Visual inspection of selected antimicrobial resistance characteristics mapped onto the *H30* phylogeny highlighted the previously identified associations between the CTX-M-15 beta-lactamase and clade C2, as well as the association between a subclade of clade C1 and the CTX-M-27 beta-lactamase. It also revealed that phenotypic resistance to TMP-SMX was spread throughout the phylogeny, although the *sul2* resistance determinant associated with sulfanamide resistance appeared associated with clade C1. Finally, there were no significant associations between clade and count of acquired AMR genes (Figures 3.6 and 3.8).

We delved deeper into the evolutionary dynamics of TMP-SMX resistance using BEAST discrete trait mapping analyses to identify whether TMP-SMX resistance pre-existed *H30*'s divergence. Initial results suggested that the ancestral isolates were largely non-susceptible to TMP-SMX, and that non-susceptibility was lost in several different instances across the phylogeny. These loss events were frequently linked with predicted loss of the *sul1* gene and *dfrA17* genes (Figure 3.7). Conversely, the ancestral isolates were predicted to not have the *sul2* gene, so there were more predicted transitions from *sul2* absence to *sul2* presence than visa versa. These observations about phenotypic TMP-SMX resistance and the *sul1* gene held in the analyses downsampled for ESC-R isolates, as well as the analyses downsampled

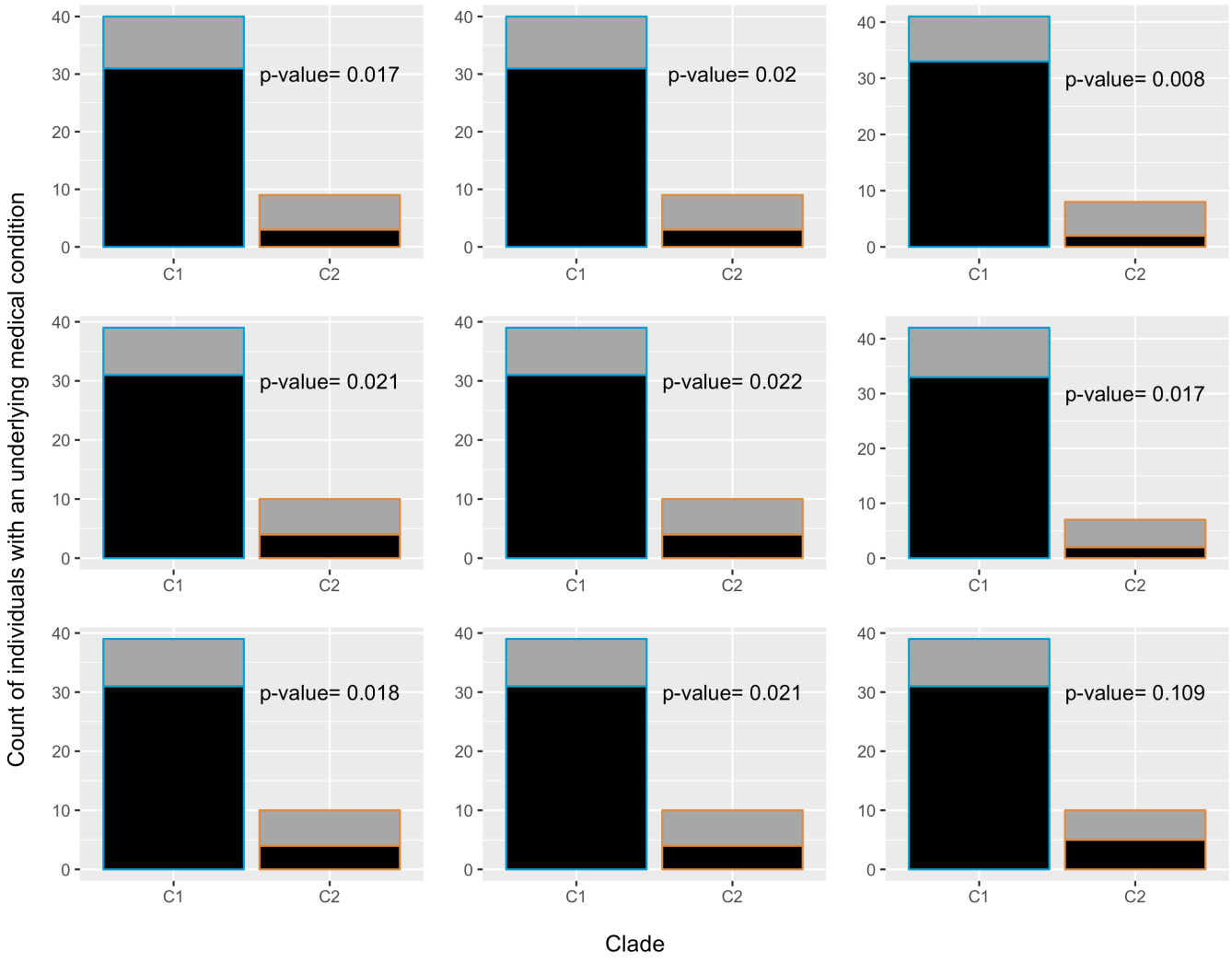

Figure 3.2: Proportion of individuals with a documented underlying medical condition by H30 clades in 9 subsampled datasets of 49 isolates each. In each subsampled dataset, the prevalence of extended-spectrum cephalosporin-resistant isolates was downsampled to 12 percent in order to reflect the estimated prevalence of extended-spectrum cephalosporin resistance in the general population of clinical H30 isolates in children. Black sections of each bar represent proportion with an underlying medical condition, while grey sections of each bar represent the proportion of individuals without an underlying medical condition. P-values calculated using chi square tests with simulated p-values.

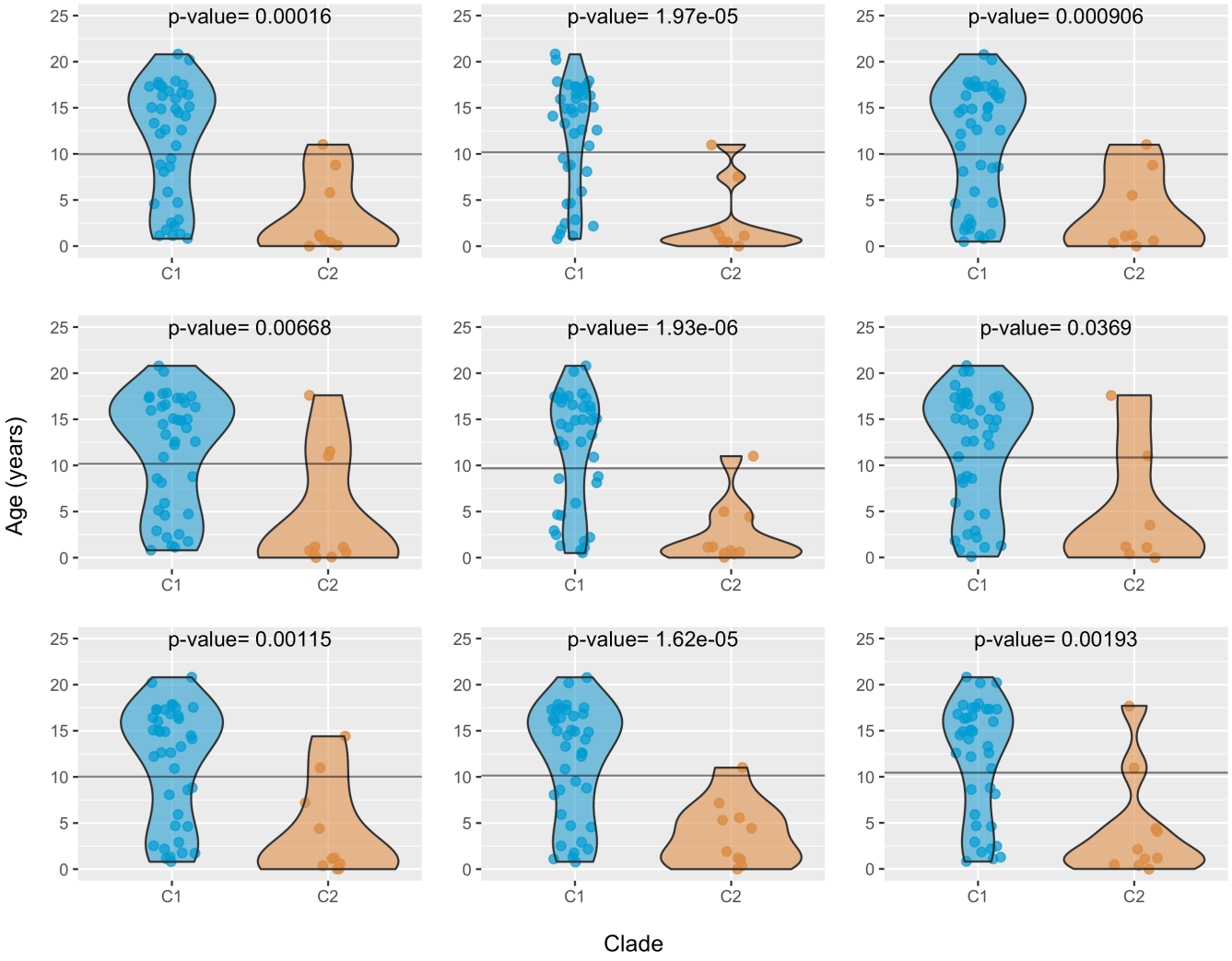

Figure 3.3: Distribution of patient age in years by H30 clades in 9 subsampled datasets of 49 isolates each. In each subsampled dataset, the prevalence of extended-spectrum cephalosporin-resistant isolates was downsampled to 12 percent in order to reflect the estimated prevalence of extended-spectrum cephalosporin resistance in the general population of clinical H30 isolates in children. P-values calculated using two sided student's t-tests comparing the mean in clade C1 to the mean in clade C2. Horizontal line represents the overall mean age in each subsampled dataset.

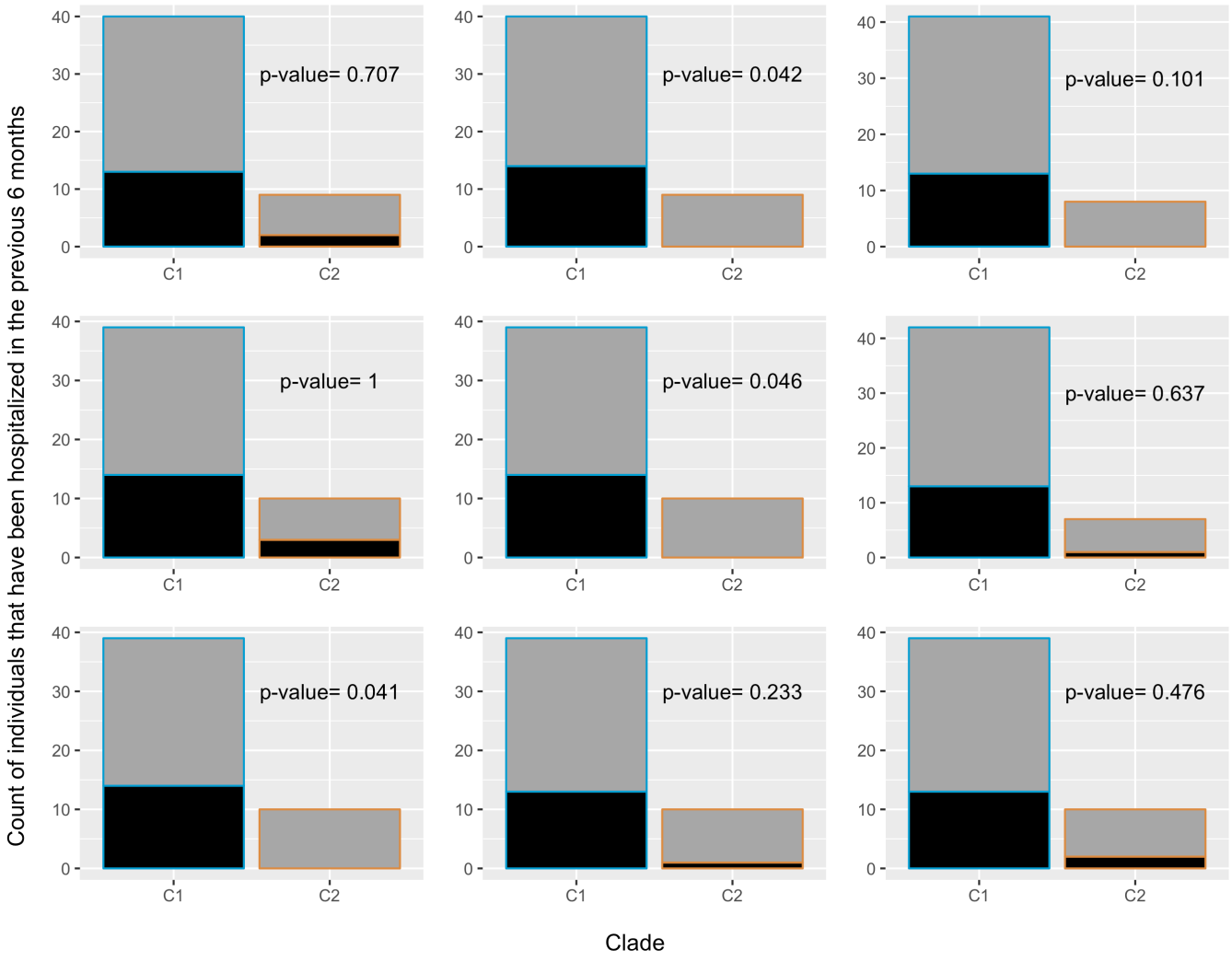

Figure 3.4: Proportion of individuals that had been hospitalized in the 6 months prior to isolate collection by H30 clades in 9 subsampled datasets of 49 isolates each. In each subsampled dataset, the prevalence of extended-spectrum cephalosporin-resistant isolates was downsampled to 12 percent in order to reflect the estimated prevalence of extended-spectrum cephalosporin resistance in the general population of clinical H30 isolates in children. Black sections of each bar represent proportion of individuals that had been hospitalized in the 6 months prior to isolate collection while grey sections of each bar represent the proportion of individuals that were not hospitilzed in the 6 months prior to isolate collection. P-values calculated using chi square tests with simulated p-values.

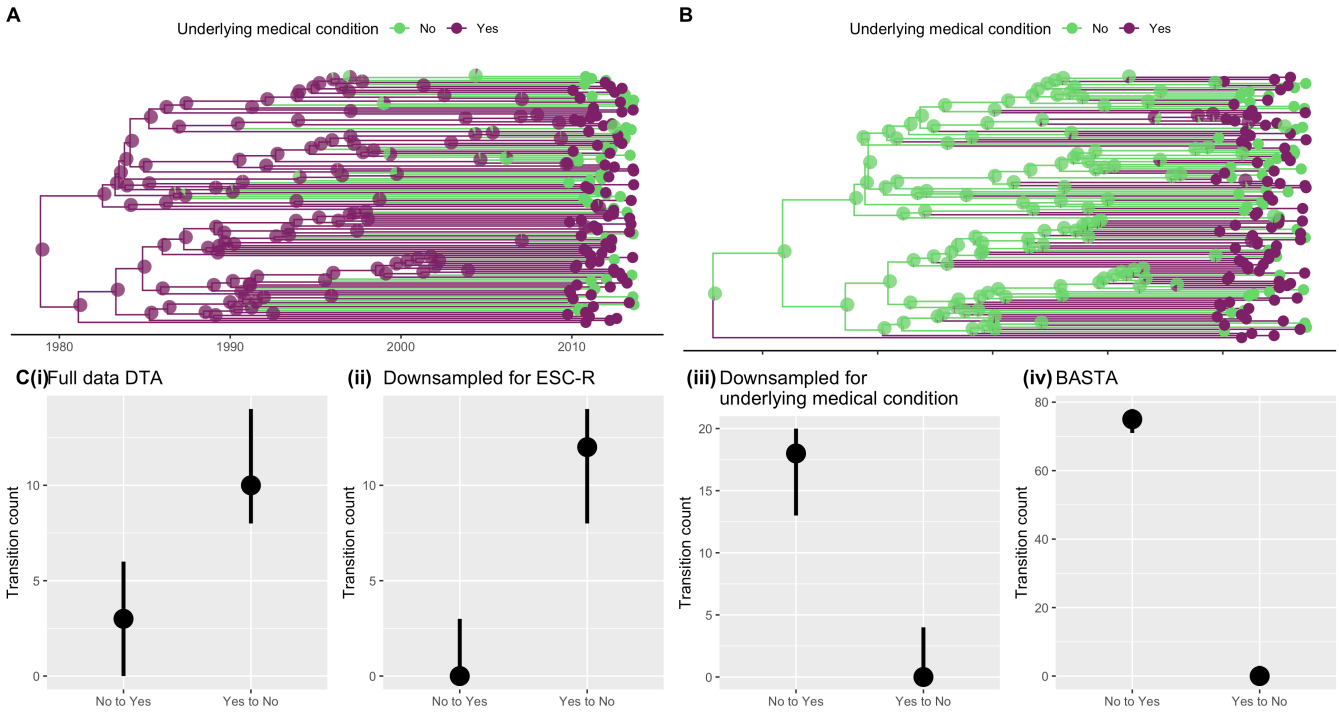

Figure 3.5: Results of discrete trait mapping analyses with presence of an underlying medical condition as the discrete trait. A) Results from a traditional discrete trait mapping analysis with the complete dataset. Maximum clade credibility tree summarizing output from BEAST with time on the x-axis. Pie chart at nodes represent the posterior probability of each ancestor being collected from an individual with an underlying medical condition. B) Results from BASTA, an alternative method for trait mapping that is based on an approximation of the structured coalescent and is thought to be less sensitive to sampling biases. Maximum clade credibility tree summarizing output from BEAST; x-axis not directly interpretable due to not adjusting for the exclusion of constant sites in BASTA analysis. Pie charts at nodes represent the posterior probability of each ancestor being collected from an individual with an underlying medical condition. C) Mean and 95 percent highest posterior density intervals for counts of estimated transitions between individuals with an underlying medical condition vs. not over evolutionary time. The figure farthest to the left represents the results of the traditional discrete trait analysis with the full dataset. The figure second from the left represents the averaged results from 9 separate BEAST runs on the 9 datasets of 49 isolates each where the prevalence of extended-spectrum cephalosporin isolates was fixed at 12 percent. The figure second from the right represents the averaged results from 9 separate BEAST runs on the 9 datasets of 55 isolates each where the prevalence of isolates collected from individuls with an underlying medical condition was fixed at 29 percent. The figure on the right shows the same results from the BASTA run on the complete dataset.

for patients with an underlying medical condition, while the *sul2* observations and *dfrA17* observations became more uncertain. (Figure 3.7 panels C and D)

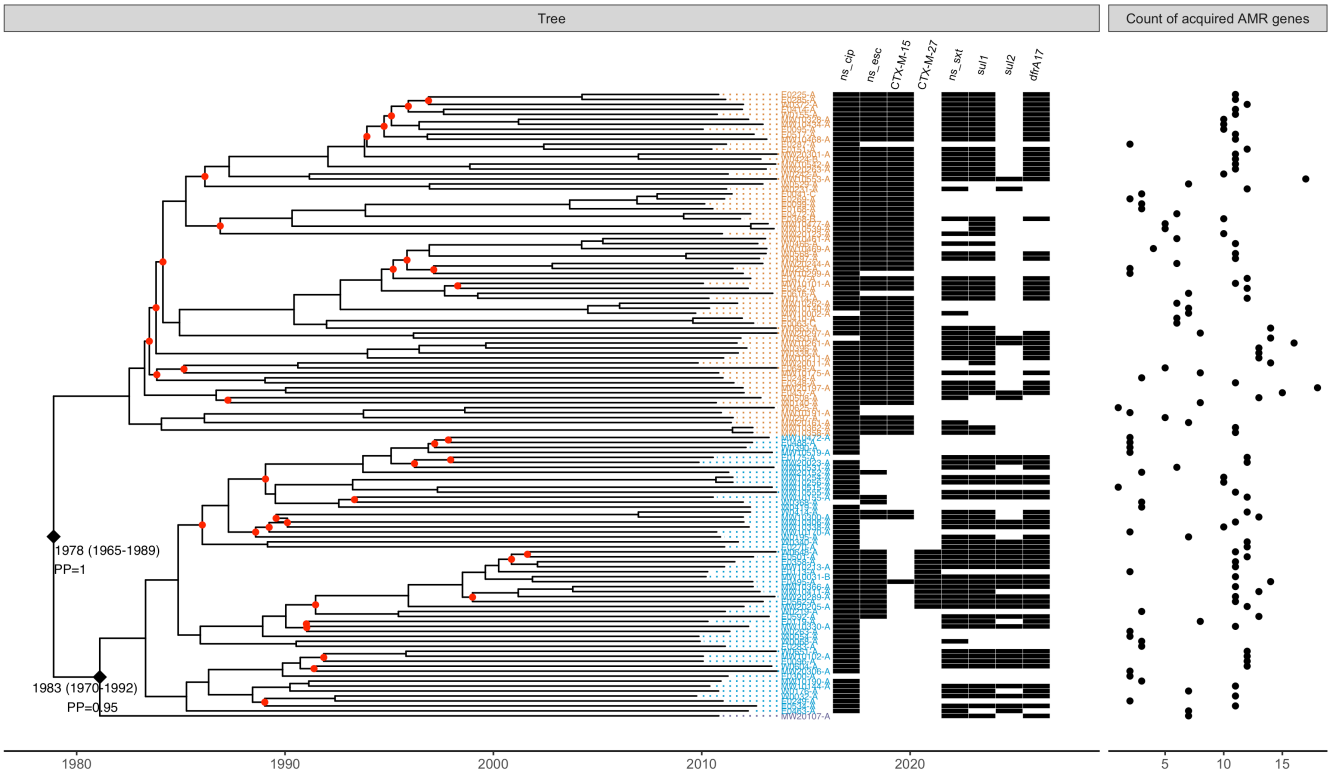

Figure 3.6: Maximum clade credibility tree of H30 based on BEAST analyses using a strict molecular clock, GMRF skyline tree prior, and general time reversibile nucleotide substitution model. The tips of the tree are constrained by isolate dates and the time scale is shown at the bottom of the tree. The tip labels of the tree are colored according to H30 clade, with orange representing clade C2, blue representing clade C1, and purple representing the intermediate clade C0 isolate. Red circles at selected nodes indicate nodes where the posterior probabilty was less than 75 percent. Black diamonds mark divergence dates of interest with estimated node ages and 95 percent highest posterior density intervals of node ages, along with posterior probability support for that node. The presence or absence of selected antimicrobial resistance characteristics are mapped to the right of the phylogeny, with black indicating presence. A count of identified acquired antimicrobial resistance genes according to the ResFinder database is located to the far right. SNV = single nucleotide polymorphism, PP = posterior probability, ns sxt = non-susceptible to trimethoprim-sulfamethoxazole, ns esc = non-susceptible to extended-spectrum cephalopsorins, ns cip = non-susceptible to ciprofloxacin

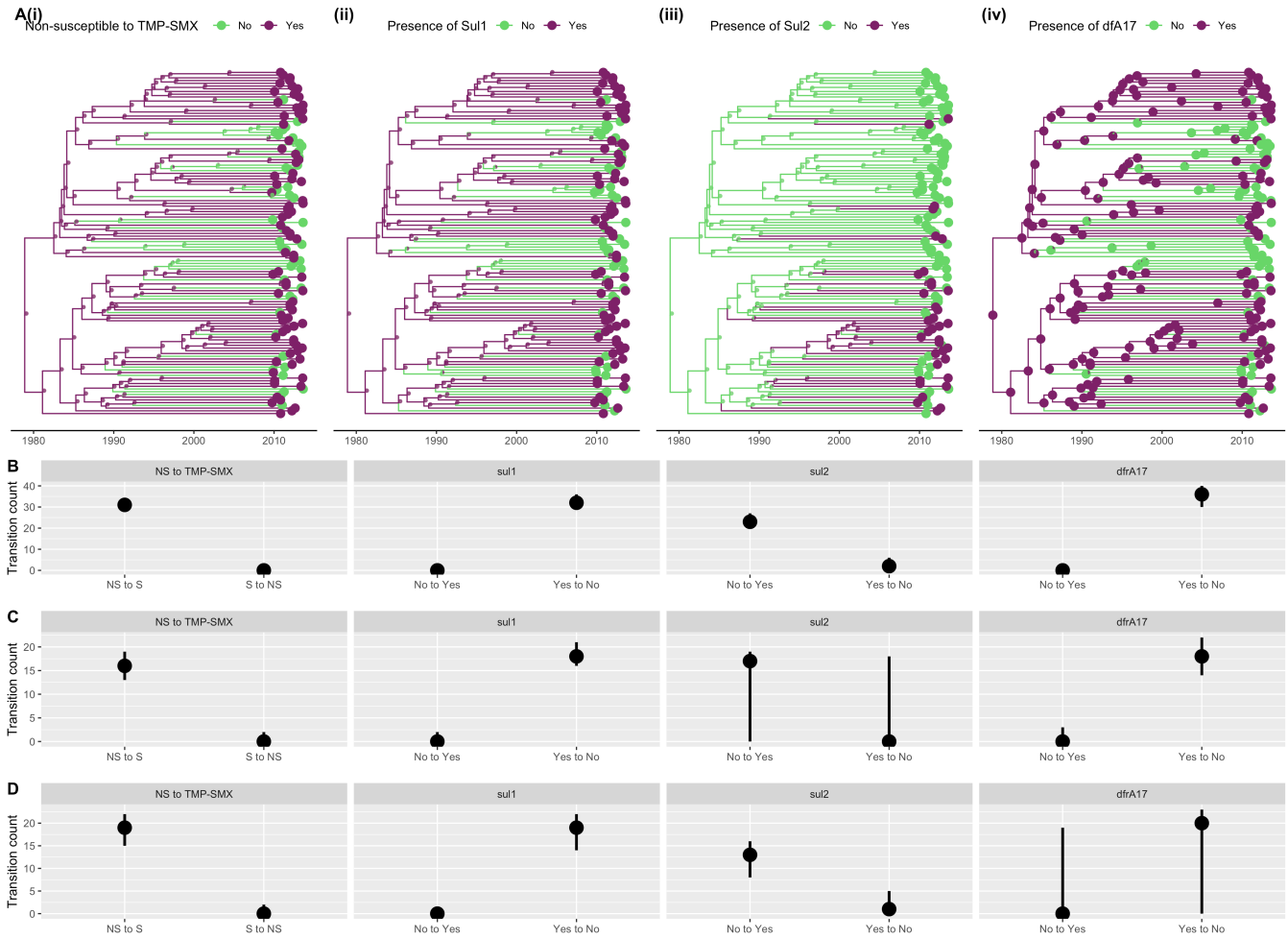

Figure 3.7: Results of discrete trait mapping analyses with various factors associated with resistance to trimethoprim-sulfamethoxazole (TMP-SMX) mapped as discrete traits. A) Results from a traditional discrete trait mapping analysis with the complete dataset. Maximum clade credibility tree summarizing output from BEAST with time on the x-axis. Pie chart at nodes represent the posterior probability of each ancestor i) being phenotypically non-susceptible to TMP-SMX, ii) posessing the sul1 gene, which confers resistance to sulfonamides, iii) posessing the sul2 gene, which confers resistance to sulfonamides, and iv) posessing the dfrA17 gene, which confers resistance to trimethoprim. B) Mean and 95 percent highest posterior density intervals for counts of estimated transitions between states associated with resistance to TMP-SMX over evolutionary time from the full dataset discrete trait analysis. C) Same as B except averaged over 9 BEAST runs on the 9 datasets of 49 isolates each where the prevalence of extended-spectrum cephalosporin-resistant isolates was fixed at 12 percent. D) Same as B except averaged over the 9 datasets of 55 isolates each where the prevalence of isolates collected from individuls with an underlying medical condition was fixed at 29 percent.

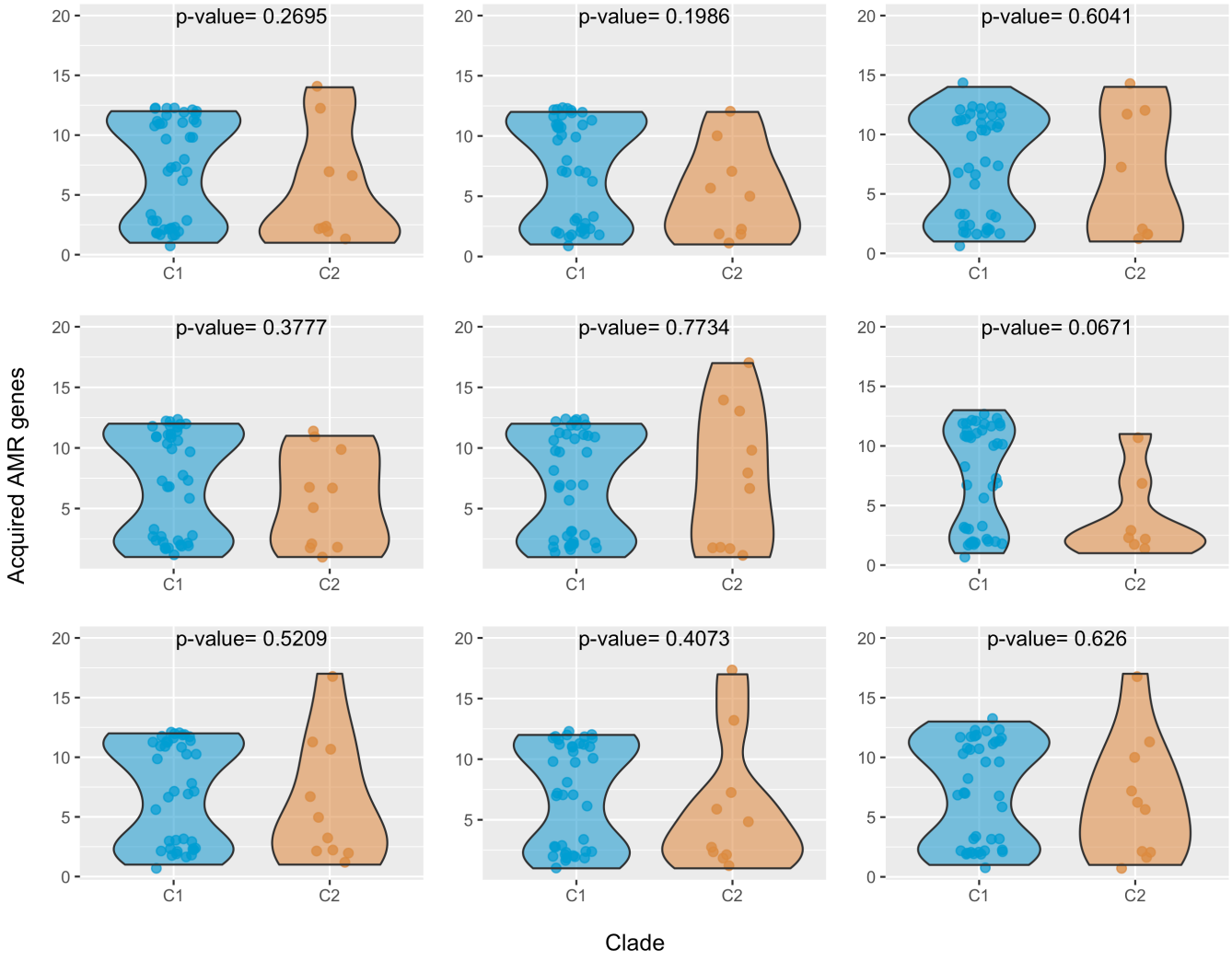

Figure 3.8: Distribution of count of identified acquired antimicrobial resistance genes as identified in the ResFinder database by H30 clades in 9 subsampled datasets of 49 isolates each. In each subsampled dataset, the prevalence of extended-spectrum cephalosporin-resistant isolates was downsampled to 12 percent in order to reflect the estimated prevalence of extended-spectrum cephalosporin resistance in the general population of clinical H30 isolates in children. P-values calculated via two sided student's t-tests.

3.4 Discussion

We integrated whole genome sequencing and patient data from a collection of *E. coli* ST131-*H30* isolates to investigate the dynamics between *H30*'s population structure and patient characteristics, to explore the healthcare vs. community origins of the lineage, and to characterize the evolutionary dynamics of antimicrobial resistance to TMP-SMX. We observed that the previously identified clades C1 and C2 impacted different portions of the pediatric population: clade C2 is associated with younger age and the absence of a documented underlying medical condition when compared to clade C1. We also observed that discrete trait stochastic mapping methods applied to presence of an underlying medical condition were senstive to oversampling of individuals with an underlying medical condition, but when taking this overrepresentation into account, ancestral *H30* isolates were more likely to be from individuals without healthcare-associated factors with evidence of numerous transitions to less healthy patients over time. Finally, we did not observe associations between clade and phenotypic resistance to TMP-SMX, or identify clonal expansion of any particular clades of *H30* as the culprit of high rates of TMP-SMX resistance among *H30* isolates.

The importance of patient age and underlying medical condition is consistent with our findings in Chapter 1, which compared these characteristics between *H30* isolates and other *E. coli* clones. There, we observed that among ESC-R infections, *H30* was associated with younger children, while among ESC-S infections, *H30* was associated with older children and presence of an underlying medical condition. Interestingly, in both the analysis in Chapter 1 and this analysis, comparing to other highly antimicrobial resistant organisms eliminates the ability to attribute the observed associations to antibiotic resistance writ large. Instead, it encourages us to consider other factors that might mediate the association between certain *H30* subtypes and young age, such as immunological factors, transmission dynamics, or specific virulence factors that are either associated with the transition from colonization to infection or colonization persistence. While the uneven sampling and small numbers in this study precluded such analyses here, a *post hoc* summary of the overall

virulence factor count in our downsampled datasets did reveal that clade C2 had a higher overall count of known virulence factors than clade C1.(Figure B.8) Overall, this observed link between clade C2 and young children is cause for concern and warrants future study.

Whether or not ST131-*H30* is more dominant in healthcare settings or in the community has been an ongoing open question since it was first identified.[82] This question is especially important for healthcare infection prevention professionals, because if healthcare is driving *H30's* spread, more resources should be put into developing interventions to interrupt transmission in the healthcare setting. We sought to capitalize on our combination of WGS and patient data about healthcare contact to explore the evolutionary history of these healthcare vs. community dynamics. While our initial results suggested that the ancestral isolates came from individuals with underlying medical conditions, the sensitivity analyses demonstrated that this association was very sensitive to changes in sampling frequency of individuals with an underlying medical condition, causing us to reverse our conclusions to healthy individuals being the most likely hosts of ancestral *H30* isolates. Despite the caution required in interpreting these analyses, it is notable that our results are not consistent with healthcare contact in particular being central to the dissemination of *H30*. This line of thought is also supported by the small number of putative transmission clusters identified in Chapter 2. Future studies from more diverse patient populations with carefully designed sampling are needed.

We also sought to shed light on the evolution of resistance to TMP-SMX, an agent that is especially vital in the pediatric population due to infrequent use of fluroquinolones in children,[98] but has historically received less attention despite high rates of resistance in *H30*. Specifically, we looked for evidence of defining gene acquisition events or clonal expansion that might be responsible for these high rates of resistance — comparable to the previously defined events that conferred fluoroquinolone resistance and extended-spectrum cephalosporin resistance within this same clone.[15,25,26] We observed that the ancestral isolates were consistently predicted to be non-susceptible to TMP-SMX, suggesting that if there

were any such defining events, they occurred outside the scope of our data. In addition, the frequent presence of multiple separate genetic determinants that are known to confer resistance to TMP-SMX is consistent with the hypothesis that this particular phenotype is likely more associated with indirect rather than direct selective forces.[99]

Finally, there are methodological findings from these analyses that are worth noting. Since bacteria evolve more slowly than other measurably evolving pathogens such as viruses, there is often uncertainty about the utility of applying temporal phylogenomic methods to collections of isolates that do not span decades.[100] Although not the primary aim of these analyses, our BEAST-based estimates for evolutionary parameters such as substitution rate, demographic history, and dates of clade divergence are consistent with other recent studies of ST131 that utilized independent and temporally wider-ranging data sets.[26,28,81] This supports the viability of these methods for future studies of ExPEC, so long as the sensitivity to changes in model specification are assessed and unnecessarily complex models are avoided. This study is also one of the first, to our knowledge, to compare the more well-established discrete trait mapping methods to newer methods based on approximations of the structured coalescent in bacterial data.[97] Interestingly, the results of the more traditional discrete trait mapping analyses that downsampled for the overrepresented trait were consistent with the structured coalescent-based results. Overall, this work highlights both the power and challenges associated with using phylogenomic methods to investigate epidemiologic questions in bacteria.

In summary, by integrating patient data and WGS data from a collection of ST131-*H30* isolates from four pediatric hospitals over four years, we observed that previously defined genetic heterogeneity within the *H30* clone is associated with patient age and presence of an underlying medical condition within U.S. children. We also observed that the community is the most likely source of ancestral *H30* isolates, although these results should be interpreted cautiously based on demonstrated sensitively to sampling. Finally, we found that the evolutionary dynamics of TMP-SMX resistance in *H30* are complex, and that if

there were particular influential gene acquisition events that led to high rates of TMP-SMX resistance in *H30*, they occurred prior to the emergence of the *H30* lineage. Future efforts geared toward elucidating the transmission dynamics of *H30* may benefit from focusing on young children and community settings.

Conclusion

Molecular epidemiology of infectious diseases originated as a typing method—a way to categorize infections into smaller, more homogenous categories to better define outbreaks and subsequently expedite the implementation of preventative measures. In recent years, rapid increased accessibility of whole genome sequencing— which provides access to **all** of the pathogen genomic data rather than just a small portion— has expanded the types of epidemiologic questions that can be addressed using molecular data.[101] The development of this new genomic toolbox for infectious disease epidemiologists is timely, as between emerging infections, re-emerging vaccine-preventable disease, and dramatically rising rates of antimicrobial resistance, we are facing some of the biggest infectious disease threats since the advent of modern medicine.

In this dissertation, I applied molecular and genomic epidemiologic methods to a concerning antimicrobial-resistant pathogen, *E. coli* ST131- *H30*, to study its epidemiology among U.S. children. In Chapter 1, I defined the burden of ST131- *H30* in children, showing that although less common overall than in adult populations, it is similarly dominant among very antimicrobial-resistant isolates. Additionally, young children were disproportionately affected by *H30* compared to other antimicrobial-resistant clones. In Chapter 2, I presented a proof of principle of the utility of isolates collected from clinical microbiology laboratories for identifying putative transmission clusters of *H30* among children. I also applied a novel probabilistic approach for quantifying uncaptured transmission events, which shows great promise for silently spreading pathogens like *H30*. In Chapter 3, I identified associations between *H30* subtypes and patient characteristics, again identifying patient age as an important factor. I also carried out a high-resolution temporal phylogenomic analyses on *H30*, and by integrating data associated with patient healthcare contact, found that ancestral *H30* isolates were more likely to be community-associated than healthcare-associated. Finally, I used this same framework to explore the evolutionary dynamics of

resistance to trimethoprim-sulfamethoxazole, a commonly used antimicrobial agent in pediatrics, and observed that the acquisition of resistance to this agent likely occurred prior to the differentiation of specific *H30* clades.

While recent work suggests that the population growth of *E. coli* ST131- *H30* has subsided, it is still currently one of a handful of dominant ExPEC clones, and deserves future study with the goal of protecting the most vulnerable patients, such as young children, from infection.[81] Additionally, as humans continue to apply evolutionary pressure through our use of antibiotics in clinical settings and in the environment, and as interclone competition continues, it is inevitable that new dominant ExPEC clones will emerge. The threat of a widespread, virulent, pan-resistant ExPEC clone is disturbingly close, and we need to apply our most sophisticated methods to study these pathogens so that we can be proactive instead of reactive when the next high-risk clone emerges.

Finally, in addition to addressing significant knowledge gaps about the epidemiology of *H30* in children, this work highlights the methodological challenges associated with genomic epidemiologic work. The data used in these analyses, which were collected during a case-control study, highlight the perils of uneven sampling in addressing epidemiologic questions. Rigorous epidemiologic principles, such as including carefully chosen comparison groups and addressing sampling biases, need to have a bigger presence in the world of molecular and genomic epidemiology, so that data and summary measures are not misinterpreted. Additionally, while all data requires some level of processing and filtering, the sheer amount of data used in genomic epidemiologic studies raises concerns about standardizing methods and how each data processing decision impacts downstream analyses. This is particularly true with bacteria, as the genomes are more complex and numerous data processing steps are required to facilitate phylogenetic analyses. More effort and resources need to be applied to systematically studying the effects of bioinformatic decisions on epidemiologic analyses, in order to continue to support the transition of these methods out of academia and into applied public health.

Overall, this work represents a leap forward in our understanding of the epidemiology of the problematic *E. coli* ST131-*H30* clone among children in the U.S. It also demonstrates that when combined with epidemiologic data, whole genome sequencing of pathogens offers great promise for the advancement of infectious disease epidemiology. In order to maximize the translational impact of genomic and computational methodological advancements, there is a need for individuals who can bridge the understanding between infectious disease epidemiology and genomics, specifically for those with both a solid foundation in epidemiologic methods and the ability to manage, analyze, and synthesize genomic data. This dissertation has served as a vehicle for my development into such an individual. I feel fortunate to have had the opportunity to conduct this work, and I look forward to continuing to develop and apply these newly acquired skills for the betterment of public health.

Appendix A

Supplementary tables

Table A.1: Most common CH types by ESC-R status (only including the first isolate collected per individual).

CH type	ESC-R (n = 278)	ESC-S (n = 1008)
H30 (40-30)	83 (29.9)	47 (4.7)
Other ST-131-associated types		
40-41	7 (2.5)	26 (2.6)
40-22	3 (1.1)	2 (0.2)
40-27	3 (1.1)	8 (0.8)
Other common types		
37-27	10 (3.6)	1 (0.1)
26-5	10 (3.6)	9 (0.9)
11-54	7 (2.5)	11 (1.1)
35-27	7 (2.5)	105 (10.4)
38-41	3 (1.1)	102 (10.1)
14-27	0 (–)	47 (4.7)
24-10	0 (–)	46 (4.6)
All others	145 (52.2)	604 (59.9)

Table A.2: Raw numbers used for the ST131-H30 and H30Rx prevalence estimates. These numbers can include repeat isolates from a given patient if they were collected within 15 days of the first isolate.

	West	Midwest 1	Midwest 2	East	All sites
ESC-R	116	100	37	81	334
H30 (all)	29	40	13	37	119
H30Rx	25	22	10	25	82
ESC-S	348	299	111	240	998
H30 (all)	13	16	3	15	47
H30Rx	1	2	0	2	5
Total # of E. coli)[a]	3021	6668	2163	9947	21799

Note:
Abbreviations: ESC-R = extended-spectrum cephalosporin-resistant. ESC-S = extended spectrum cephalosporin-susceptible
[a] This count was provided by each study center and was used to calculate weights for the overall estimate of prevalence.

Table A.3: Analysis of interaction between age and underlying medical condition risk of H30 infection vs. infection with other E. coli types using log-binomial regression models among ESC-R isolates

	Age		
	0-5 years	**6-21 years**	
	RR (95% CI)[a]	RR (95% CI)[a]	RRs (95% CI)[a] for Age 0-5 vs. Age 6-21 within strata of underlying medical condition
Presence of an underlying medical condition	2.07 (1.22-3.52)	1.0 (ref)	2.04 (1.20-3.45)
No underlying medical condition	2.07 (1.18-3.64)	1.11 (0.52-2.36)	1.92 (0.98-3.75)
RRs (95% CI) for underlying medical condition within age strata	1.00 (0.67-1.49)	0.87 (0.41-1.87)	

Note:
Interaction contrast ratio (ICR), -0.10 (95% confidence interval, -1.67 to 1.03). When interpreting the ICR, deviation from 0 indicates evidence of interaction on the additive scale (see Supplementary Methods).
Abbreviations: CI, confidence interval; RR, relative risk.
[a] RRs adjusted for study hospital.

Appendix B

Supplementary figures

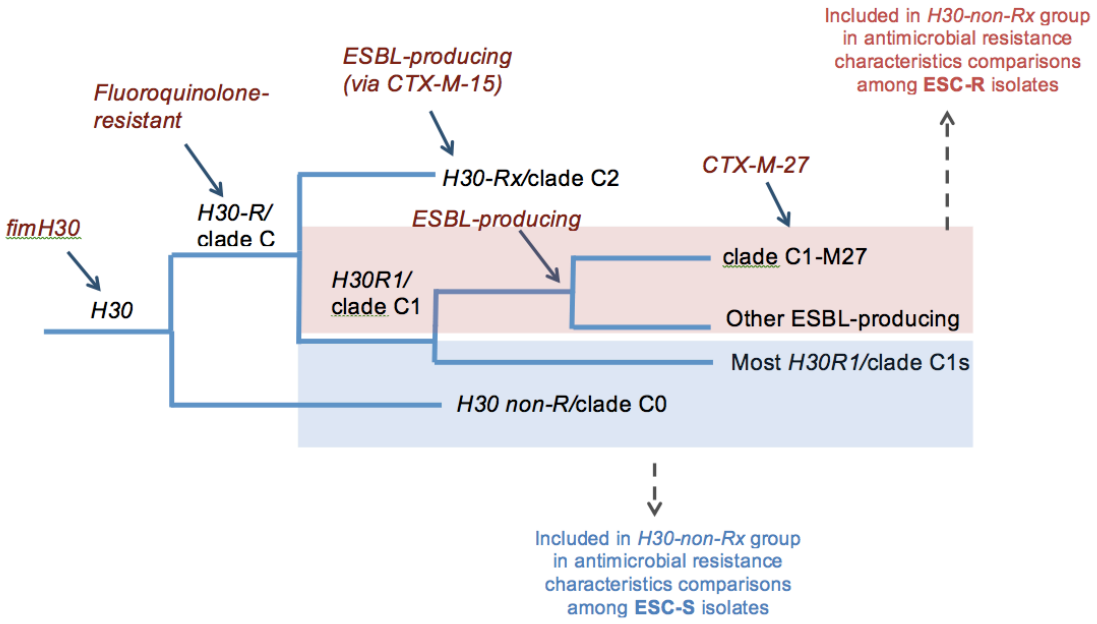

Figure B.1: Schematic dendrogram of ST131-H30 population structure as interpreted in Chapter 1, with associated phenotypic and genotypic characteristics. Solid arrows indicate strong associations identified in existing literature, but are not universal; organisms that have undergone recombination or lost mobile genetic elements may not possess these characteristics. ESBL = extended-spectrum beta-lactamase.

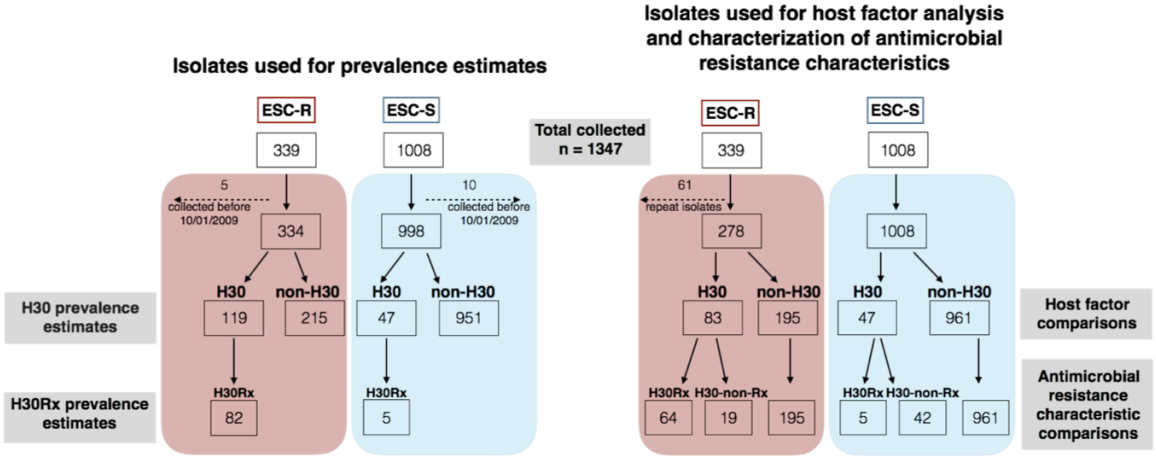

Figure B.2: Schematic of subsets of isolates and data used in each presented analyses in Chapter 1

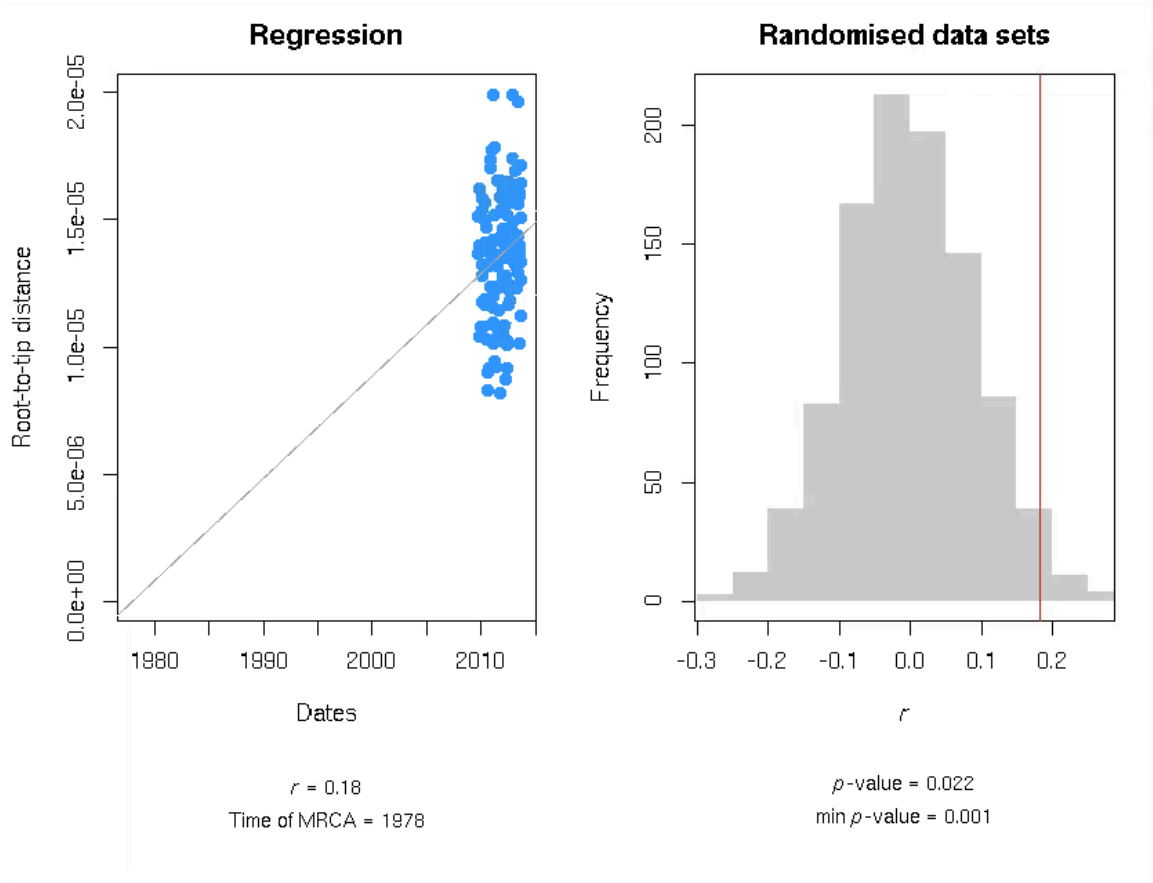

Figure B.3: Results of a root-to-tip regression on the H30 isolates included in this study.The tip dates, assigned using the collection date of the isolate, is regressed upon the distance from the estimated root in the maximum likelihood phylogeny to determine temporal signal.

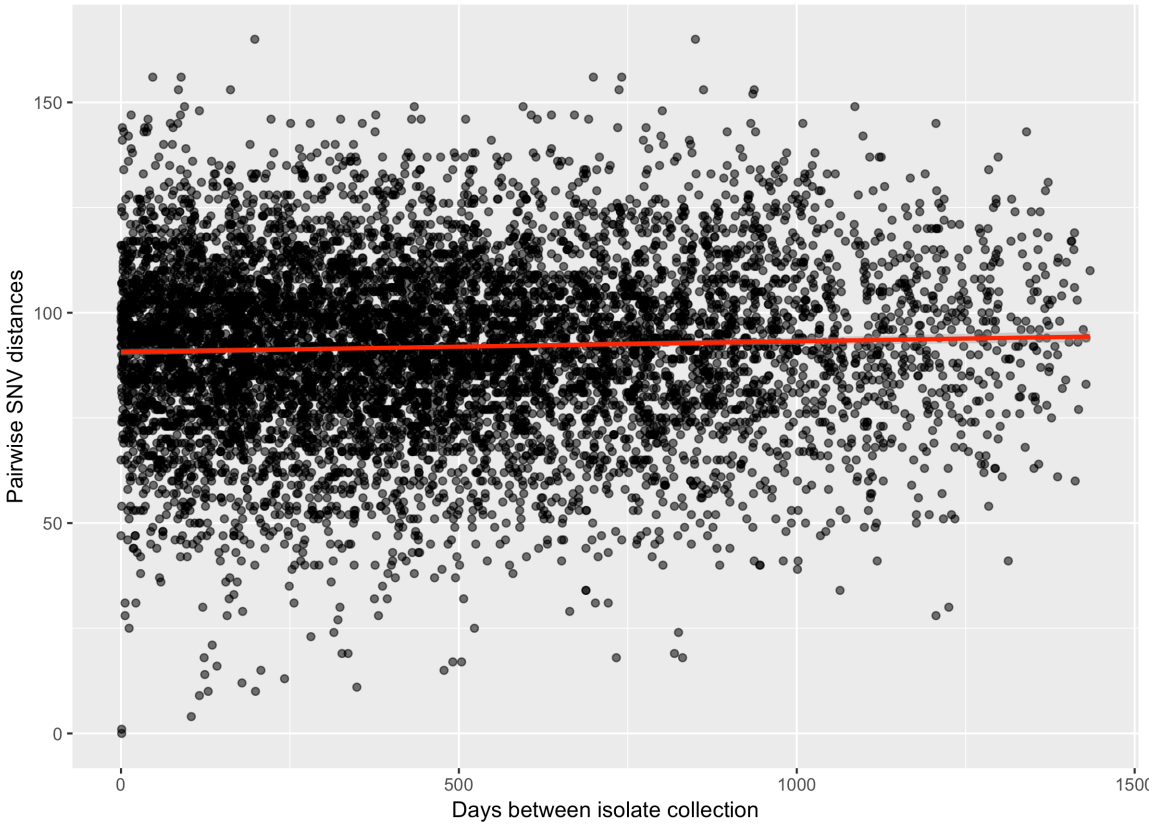

Figure B.4: Scatter plot of pairwise single nucleotide variant (SNV) distances against the days between isolate collection. A linear regression model produced a slope of 0.00257 and a p-value of 0.0002, indicating that there is a slight but statistically significant positive linear trend. These results led to the selection of a strict molecular clock model for all dated phylogenetic analyses

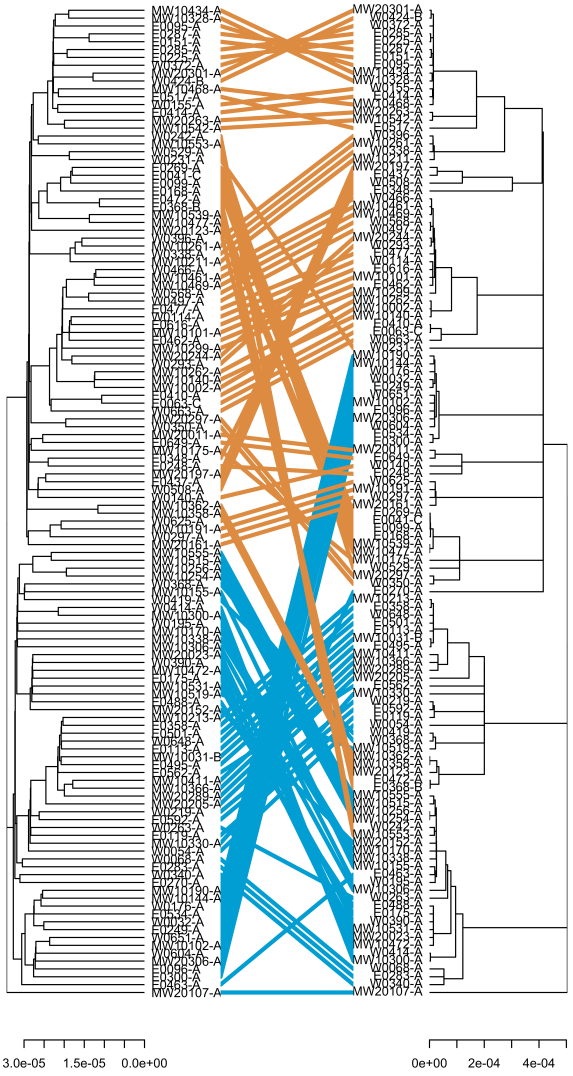

Figure B.5: Tanglegram reflecting the topological differences between a maximum-likelihood phylogeny constructed from single nucleotide variants (SNVs) with SNVs associated with recombinant events as identified by Gubbins, or phage-associated regions identified in PHASTer filtered out (3433 SNVs, left) to a maximum-likelihood phylogeny constructed from SNVs without filtering for putative recombinant sites and phage-associated regions (10,005 SNVs, right). Both phylogenies were created by mapping short reads to the EC958 reference genome. Orange corresponds to clade C2 in the filtered tree, while blue corresponds to clade C1.

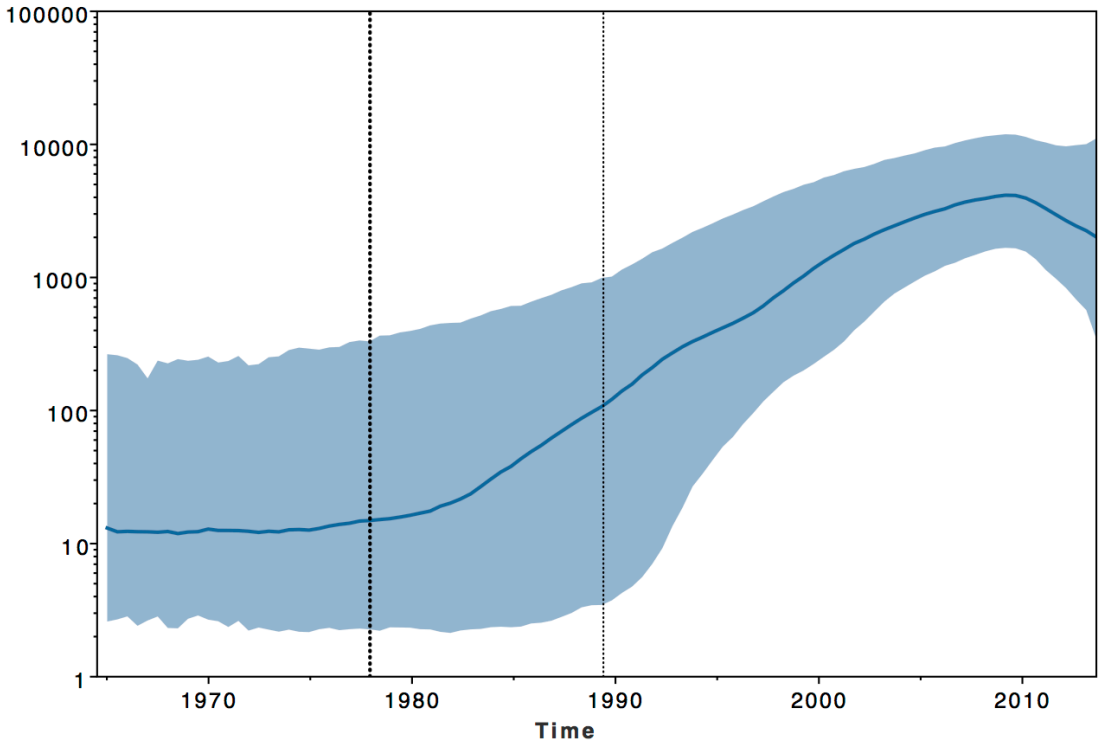

Figure B.6: This curve, output from BEAST, shows the growth line for the median population growth rate of E.coli ST131-H30 according to the temporal phylogenomic analysis conducted in this study with the selection of a strict molecular clock and a GMRF skyline demographic model. The y axis is on a log scale. The solid blue area represents the 95 percent HPD interval for the growth rate. The dotted black lines represent the 95 percent HPD for the age of the root of the phylogeny

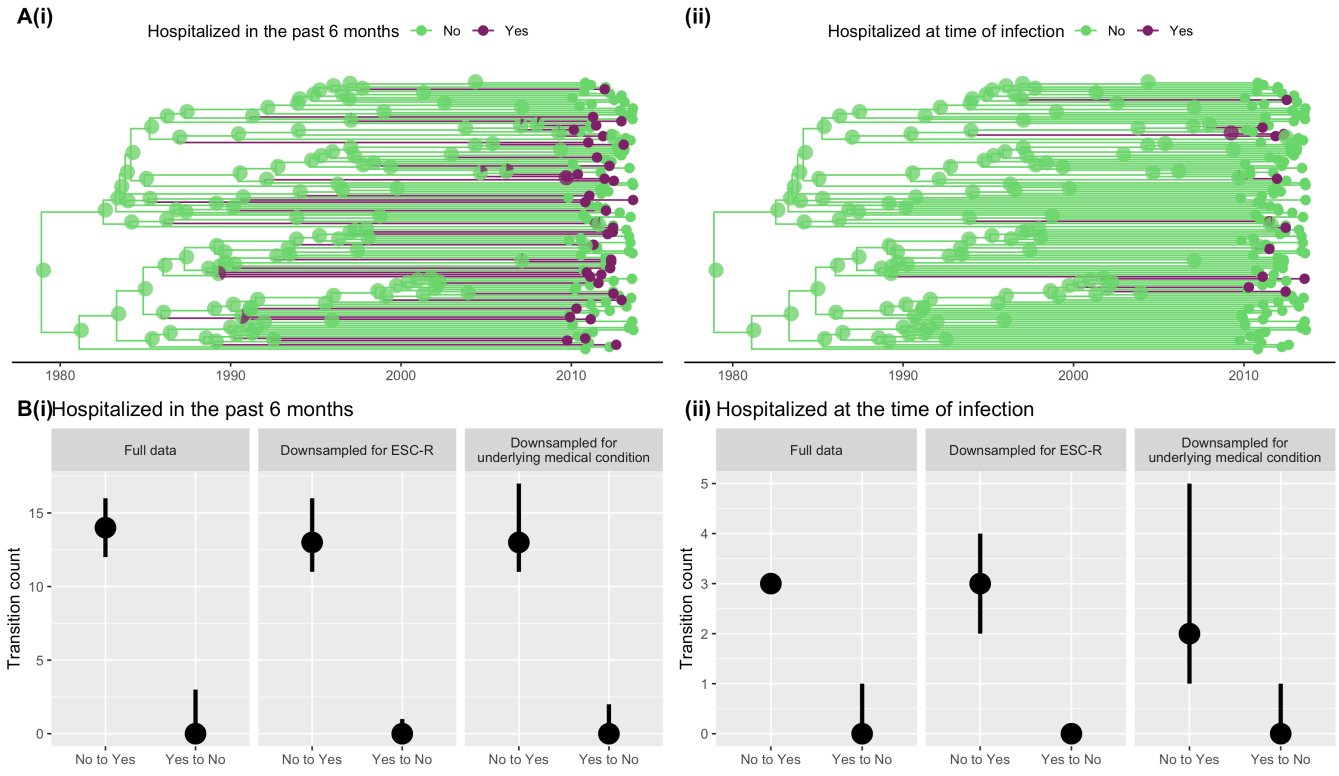

Figure B.7: Results of discrete trait mapping analyses with hospitalization in the 6 months preceding infection and hospitalization at the time of infection mapped as discrete traits. A) Results from a discrete trait mapping analysis with the complete dataset. Maximum clade credibility tree summarizing output from BEAST with time on the x-axis. Pie chart at nodes represent the posterior probability of each ancestor i) being isolated from an individual that was hospitalized in the 6 months prior to infection, ii) being isolated from an individual that was hospitalized at the time of infection B(i) Mean and 95 percent highest posterior density intervals for counts of estimated transitions between hosts that were previously hospitalized vs. not over evolutionary time. Shown for discrete trait analysis with full dataset, discrete trait analysis averaged over 9 BEAST runs on the 9 datasets of 49 isolates each where the prevalence of extended-spectrum cephalosporin isolates was fixed at 12 percent, and averaged over the 9 datasets of 55 isolates each where the prevalence of isolates collected from individuls with an underlying medical condition was fixed at 29 percent. C) Same as B except for hospitalization at the time of infection as the discrete trait

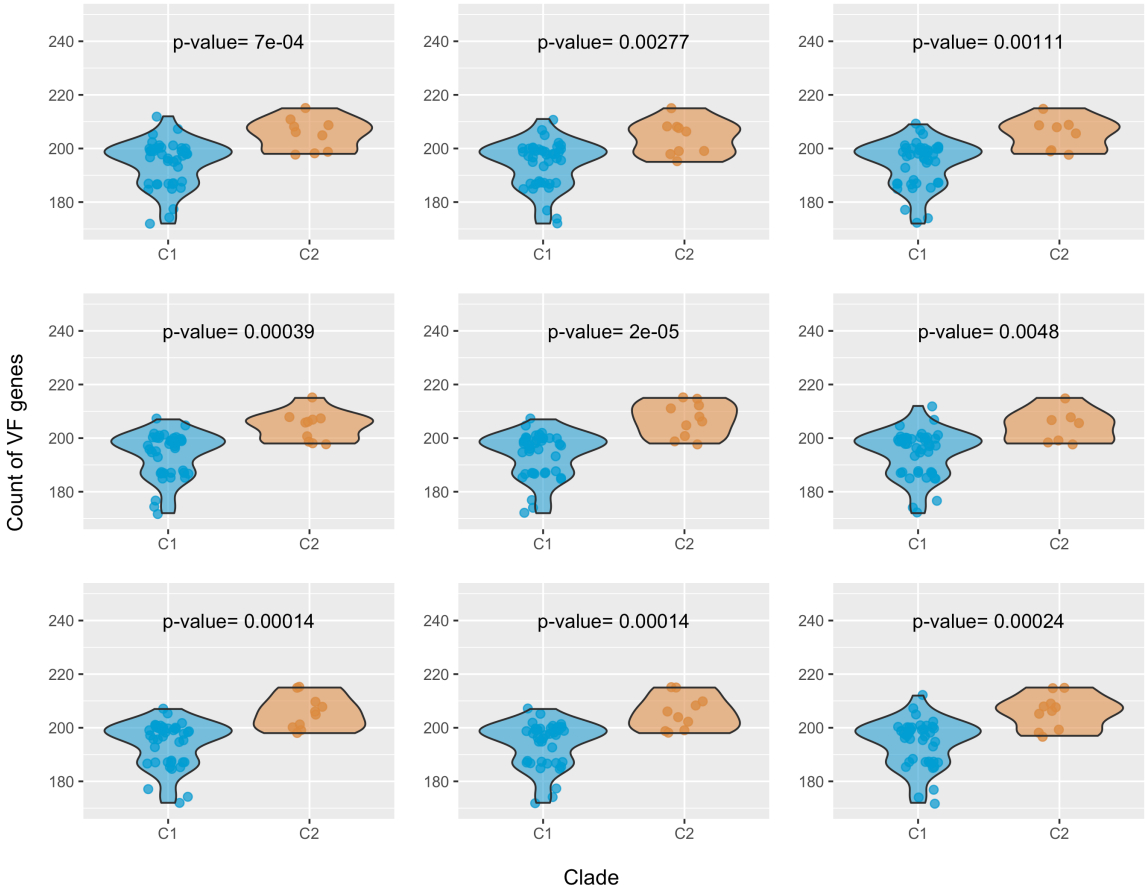

Figure B.8: Distribution of count of identified virulence factor genes as identified in the ecoli vf database by H30 cladesin 9 subsampled datasets of 49 isolates each. In each subsampled dataset, the prevalence of extended-spectrum cephalosporin-resistant isolates was down-sampled to 12 percent in order to reflect the estimated prevalence of extended-spectrum cephalosporin resistance in the general population of clinical H30 isolates in children. P-values calculated two sided student's t-tests

References

1. Kaper JB, Nataro JP, Mobley HLT. Pathogenic Escherichia coli. *Nature Reviews Microbiology*. 2004;2(2):123-140. doi:10.1038/nrmicro818

2. Gaspari RJ, Dickson E, Karlowsky J, Doern G. Antibiotic resistance trends in paediatric uropathogens. *International Journal of Antimicrobial Agents*. 2005;26(4):267-271. doi:10.1016/j.ijantimicag.2005.07.009

3. Freedman AL, the Urologic Diseases in America Project. Urologic diseases in America project: Trends in resource utilization for urinary tract infections in children. *Journal of Urology*. 2005;173(3):949-954. doi:10.1097/01.ju.0000152092.03931.9a

4. Yamamoto S, Tsukamoto T, Terai A, Kurazono H, Takeda Y, Yoshida O. Genetic evidence supporting the fecal-perineal-urethral hypothesis in cystitis caused by Escherichia coli. *Journal of Urology*. 1997;157(3):1127-1129. doi:10.1016/S0022-5347(01)65154-1

5. Russo TA, Johnson JR. Proposal for a New Inclusive Designation for Extraintestinal Pathogenic Isolates of Escherichia coli: ExPEC. *The Journal of Infectious Diseases*. 2000;181(5):1753-1754. doi:10.1086/315418

6. World Health Organization. Antimicrobial Resistance: Global Report on Surveillance 2014. World Health Organization. http://www.who.int/drugresistance/documents/surveillancereport/en/.

7. Riley L. Pandemic lineages of extraintestinal pathogenic Escherichia coli. *Clinical Microbiology and Infection*. 2014;20(5):380-390. doi:10.1111/1469-0691.12646

8. Baker S, Thomson N, Weill F-X, Holt KE. Genomic insights into the emergence and spread of antimicrobial-resistant bacterial pathogens. *Science*. 2018;360(6390):733-738. doi:10.1126/science.aar3777

9. Baquero F, Coque TM. Multilevel population genetics in antibiotic resistance. *FEMS*

Microbiology Reviews. 2011;35(5):705-706. doi:10.1111/j.1574-6976.2011.00293.x

10. Mathers AJ, Peirano G, Pitout JD. Escherichia coli ST131: The Quintessential Example of an International Multiresistant High-Risk Clone. *Adv Appl Microbiol*. 2015;90:109-154. doi:10.1016/bs.aambs.2014.09.002

11. Johnson JR, Tchesnokova V, Johnston B, et al. Abrupt emergence of a single dominant multidrug-resistant strain of Escherichia coli. *Journal of Infectious Diseases*. 2013;207(6):919-928. doi:10.1093/infdis/jis933

12. Banerjee R, Johnston B, Lohse C, et al. The clonal distribution and diversity of extraintestinal Escherichia coli isolates vary according to patient characteristics. *Antimicrobial Agents and Chemotherapy*. 2013;57(12):5912-5917. doi:10.1128/AAC.01065-13

13. Drawz SM, Porter S, Kuskowski M a., et al. Variation in resistance traits, phylogenetic background, and virulence genotypes among Escherichia coli clinical isolates from adjacent hospital campuses serving distinct patient populations. *Antimicrobial Agents and Chemotherapy*. 2015;59(9):AAC.00048-15. doi:10.1128/AAC.00048-15

14. Johnson JR, Thuras P, Johnston BD, et al. The Pandemic H 30 Subclone of Escherichia coli Sequence Type 131 (ST131) is Associated with Persistent Infections and Adverse Outcomes Independently from Its Multi-Drug Resistance and Associations with Compromised Hosts. *Clinical Infectious Diseases*. 2016;131(Cdc):ciw193. doi:10.1093/cid/ciw193

15. Price LB, Johnson JR, Aziz M, et al. The epidemic of extended-spectrum-β-lactamase-producing Escherichia coli ST131 is driven by a single highly pathogenic subclone, H30-Rx. *mBio*. 2013;4(6):e00377-13. doi:10.1128/mBio.00377-13

16. Banerjee R, Johnston B, Lohse C, Porter SB, Clabots C, Johnson JR. Escherichia coli Sequence Type 131 is a Dominant, Antimicrobial-Resistant Clonal Group Associated with Healthcare and Elderly Hosts. *Infection Control and Hospital Epidemiology*. 2013;34(4):361-369. doi:10.1086/669865

17. Burgess MJ, Johnson JR, Porter SB, et al. Long-Term Care Facilities Are Reservoirs for

Antimicrobial-Resistant Sequence Type 131 Escherichia coli. *Open forum infectious diseases*. 2015;2(1):ofv011. doi:10.1093/ofid/ofv011

18. Colpan A, Johnston B, Porter S, et al. Escherichia coli sequence type 131 (ST131) subclone h30 as an emergent multidrug-resistant pathogen among US Veterans. *Clinical Infectious Diseases*. 2013;57(9):1256-1265. doi:10.1093/cid/cit503

19. Miles-Jay A, Weissman SJ, Adler AL, et al. Epidemiology and Antimicrobial Resistance Characteristics of the Sequence Type 131-H30 Subclone among Extraintestinal Escherichia coli Collected from US Children. *Clinical Infectious Diseases*. 2018;66(3):411-419. doi:10.1093/cid/cix805

20. Johnson JR, Johnston B, Clabots C, Kuskowski M a, Castanheira M. Escherichia coli sequence type ST131 as the major cause of serious multidrug-resistant E. coli infections in the United States. *Clinical Infectious Diseases*. 2010;51(3):286-294. doi:10.1086/653932

21. Johnson JR, Nicolas-Chanoine MH, Deb Roy C, et al. Comparison of Escherichia coli ST131 pulsotypes, by epidemiologic traits, 1967-2009. *Emerging Infectious Diseases*. 2012;18(4):598-607. doi:10.3201/eid1804.111627

22. Nicolas-Chanoine M-H, Blanco J, Leflon-Guibout V, et al. Intercontinental emergence of Escherichia coli clone O25:H4-ST131 producing CTX-M-15. *Journal of Antimicrobial Chemotherapy*. 2008;61(2):273-281. doi:10.1093/jac/dkm464

23. Johnson JR, Tchesnokova V, Johnston B, et al. Abrupt emergence of a single dominant multidrug-resistant strain of Escherichia coli. *Journal of Infectious Diseases*. 2013;207(6):919-928. doi:10.1093/infdis/jis933

24. Johnson JR, Porter S, Thuras P, Castanheira M. Epidemic Emergence in the United States of Escherichia coli Sequence Type 131-H30, 2000-2009. *Antimicrobial Agents and Chemotherapy*. 2017;61(8):AAC.00732-17. doi:10.1128/AAC.00732-17

25. Petty NK, Ben Zakour NL, Stanton-Cook M, et al. Global dissemination of a multidrug resistant Escherichia coli clone. *Proceedings of the National Academy of Sciences of the United*

States of America. 2014;111(15):5694-5699. doi:10.1073/pnas.1322678111

26. Ben Zakour NL, Alsheikh-Hussain AS, Ashcroft MM, et al. Sequential acquisition of virulence and fluoroquinolone resistance has shaped the evolution of Escherichia coli ST131. *mBio*. 2016;7(May):1-12. doi:10.1101/039123

27. Matsumura Y, Pitout JD, Gomi R, et al. Global Escherichia coli Sequence Type 131 Clade with bla CTX-M-27 Gene. *Emerging Infectious Diseases*. 2016;22(11):1900-1907. doi:10.3201/eid2211.160519

28. Stoesser N, Sheppard AE, Pankhurst L, et al. Evolutionary History of the Global Emergence of the Escherichia coli Epidemic Clone ST131. *mBio*. 2016;7(2):e02162-15. doi:10.1128/mBio.02162-15

29. Baquero F, Tedim AP, Coque TM. Antibiotic resistance shaping multi-level population biology of bacteria. *Frontiers in Microbiology*. 2013;4(MAR):1-15. doi:10.3389/fmicb.2013.00015

30. Banerjee R, Johnston B, Lohse C, et al. The clonal distribution and diversity of extraintestinal Escherichia coli isolates vary according to patient characteristics. *Antimicrobial Agents and Chemotherapy*. 2013;57(12):5912-5917. doi:10.1128/AAC.01065-13

31. Banerjee R, Robicsek A, Kuskowski M a., et al. Molecular epidemiology of Escherichia coli sequence type 131 and its H30 and H30-Rx subclones among extended-spectrum-beta-lactamase-positive and -negative E. coli clinical isolates from the Chicago region, 2007 to 2010. *Antimicrobial Agents and Chemotherapy*. 2013;57(12):6385-6388. doi:10.1128/AAC.01604-13

32. Peirano G, Pitout JDD. Fluoroquinolone-resistant Escherichia coli sequence type 131 isolates causing bloodstream infections in a canadian region with a centralized laboratory system: rapid emergence of the H30-Rx sublineage. *Antimicrobial Agents and Chemotherapy*. 2014;58(5):2699-2703. doi:10.1128/AAC.00119-14

33. Jacobson SH, Eklöf O, Eriksson CG, Lins LE, Tidgren B, Winberg J. Development of hy-

pertension and uraemia after pyelonephritis in childhood: 27 year follow up. *BMJ (Clinical research ed)*. 1989;299(6701):703-706. doi:10.1136/bmj.299.6701.703

34. Shaikh N, Ewing AL, Bhatnagar S, Hoberman A. Risk of Renal Scarring in Children With a First Urinary Tract Infection: A Systematic Review. *Pediatrics*. 2010;126:1084-1091. doi:10.1542/peds.2010-0685

35. Zerr DM, Miles-Jay A, Kronman MP, et al. Previous antibiotic exposure increases risk of infection with extended spectrum beta lactamase- and AmpC-producing Escherichia coli and Klebsiella pneumoniae in pediatric patients. *Antimicrobial Agents and Chemotherapy*. 2016;60(7):4237-4243. doi:10.1128/AAC.00187-16

36. Das S, Adler AL, Miles-Jay A, et al. Antibiotic prophylaxis is associated with subsequent resistant infections in children with an initial extended-spectrum cephalosporin-resistant Enterobacteriaceae infection. *Antimicrobial Agents and Chemotherapy*. 2017;61(5):AAC.02656-16. doi:10.1128/AAC.02656-16

37. Weissman SJ, Johnson JR, Tchesnokova V, et al. High-resolution two-locus clonal typing of extraintestinal pathogenic Escherichia coli. *Applied and Environmental Microbiology*. 2012;78(5):1353-1360. doi:10.1128/AEM.06663-11

38. DiCiccio TJ, Efron B. Bootstrap Confidence Intervals. 1996;11(3):189-228. doi:doi:10.1214/ss/1032280214

39. VanderWeele TJ, Knol MJ. A Tutorial on Interaction. *Epidemiologic Methods*. 2014;3(1). doi:10.1515/em-2013-0005

40. Knol MJ, VanderWeele TJ, Groenwold RHH, Klungel OH, Rovers MM, Grobbee DE. Estimating measures of interaction on an additive scale for preventive exposures. *European Journal of Epidemiology*. 2011;26(6):433-438. doi:10.1007/s10654-011-9554-9

41. Olesen B, Frimodt-Moller J, Leihof RF, et al. Temporal Trends in Antimicrobial Resistance and Virulence-Associated Traits within the Escherichia coli Sequence Type 131 Clonal Group and Its H30 and H30-Rx Subclones, 1968 to 2012. *Antimicrobial Agents and*

Chemotherapy. 2014;58(11):6886-6895. doi:10.1128/AAC.03679-14

42. Logan LK, Hujer AM, Marshall SH, et al. Analysis of β-Lactamase Resistance Determinants in Enterobacteriaceae from Chicago Children: A Multicenter Survey. *Antimicrobial Agents and Chemotherapy*. 2016;60(March):3462-3469. doi:10.1128/AAC.00098-16

43. Bradley JS, Jackson Ma. The Use of Systemic and Topical Fluoroquinolones. *Pediatrics*. 2011;128:e1034-e1045. doi:10.1542/peds.2011-1496

44. Rogers Ba, Ingram PR, Runnegar N, et al. Sequence type 131 fimH30 and fimH41 subclones amongst Escherichia coli isolates in Australia and New Zealand. *International Journal of Antimicrobial Agents*. 2015;45:351-358. doi:10.1016/j.ijantimicag.2014.11.015

45. Logan LK, Braykov NP, Weinstein R a., Laxminarayan R. Extended-Spectrum Beta-Lactamase-Producing and Third-Generation Cephalosporin-Resistant Enterobacteriaceae in Children: Trends in the United States, 1999-2011. *Journal of the Pediatric Infectious Diseases Society*. March 2014:1-9. doi:10.1093/jpids/piu010

46. Gurnee Ea, Ndao IM, Johnson JR, et al. Gut Colonization of Healthy Children and Their Mothers With Pathogenic Ciprofloxacin-Resistant Escherichia coli. *Journal of Infectious Diseases*. 2015:1-7. doi:10.1093/infdis/jiv278

47. Blanc V, Leflon-Guibout V, Blanco J, et al. Prevalence of day-care centre children (France) with faecal CTX-M-producing Escherichia coli comprising O25b:H4 and O16:H5 ST131 strains. *Journal of Antimicrobial Chemotherapy*. 2014;69(January):1231-1237. doi:10.1093/jac/dkt519

48. Johnson JR, Davis G, Clabots C, et al. Household Clustering of Escherichia coli Sequence Type 131 Clinical and Fecal Isolates According to Whole Genome Sequence Analysis. *Open Forum Infectious Diseases*. 2016;3(3):ofw129. doi:10.1093/ofid/ofw129

49. Madigan T, Johnson JR, Clabots C, et al. Extensive Household Outbreak of Urinary Tract Infection and Intestinal Colonization due to Extended-Spectrum Beta-Lactamase-Producing Escherichia coli Sequence Type 131. *Clinical Infectious Diseases*. 2015;61(1):e5-

e12. doi:10.1093/cid/civ273

50. Banerjee R, Johnston B, Lohse C, Porter SB, Clabots C, Johnson JR. Escherichia coli Sequence Type 131 is a Dominant, Antimicrobial-Resistant Clonal Group Associated with Healthcare and Elderly Hosts. *Infection Control and Hospital Epidemiology*. 2013;34(4):361-369. doi:10.1086/669865

51. Safdar N, Maki DG. The Commonality of Risk Factors for Nosocomial Colonization and. *Annals of Internal Medicine*. 2002;136(11):834-844. doi:10.7326/0003-4819-136-11-200206040-00013

52. Russo T a., Johnson JR. Medical and economic impact of extraintestinal infections due to Escherichia coli: Focus on an increasingly important endemic problem. *Microbes and Infection*. 2003;5(5):449-456. doi:10.1016/S1286-4579(03)00049-2

53. Hilty M, Betsch BY, Bögli-Stuber K, et al. Transmission dynamics of extended-spectrum beta-lactamase-producing enterobacteriaceae in the tertiary care hospital and the household setting. *Clinical Infectious Diseases*. 2012;55:967-975. doi:10.1093/cid/cis581

54. Lartigue MF, Zinsius C, Wenger A, Bille J, Poirel L, Nordmann P. Extended-spectrum β-lactamases of the CTX-M type now in Switzerland. *Antimicrobial Agents and Chemotherapy*. 2007;51(8):2855-2860. doi:10.1128/AAC.01614-06

55. Torres E, López-Cerero L, Morales I, Navarro MD, Rodríguez-Baño J, Pascual A. Prevalence and transmission dynamics of Escherichia coli ST131 among contacts of infected community and hospitalized patients. *Clinical Microbiology and Infection*. 2018;24(6):618-623. doi:10.1016/j.cmi.2017.09.007

56. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed January 21, 2019.

57. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence

data. *Bioinformatics*. 2014;30(15):2114-2120. doi:10.1093/bioinformatics/btu170

58. Ewels P, Magnusson M, Lundin S, Kaller M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016;32(19):3047-3048. doi:10.1093/bioinformatics/btw354

59. Forde BM, Ben Zakour NL, Stanton-Cook M, et al. The complete genome sequence of escherichia coli EC958: A high quality reference sequence for the globally disseminated multidrug resistant E. coli O25b:H4-ST131 clone. *PLoS ONE*. 2014;9(8). doi:10.1371/journal.pone.0104400

60. Seemann T. *Snippy: Rapid Haploid Variant Calling and Core Genome Alignment*. https://github.com/tseemann/snippy. Accessed January 21, 2019.

61. Arndt D, Grant JR, Marcu A, et al. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic acids research*. 2016;44(W1):W16-W21. doi:10.1093/nar/gkw387

62. Croucher NJ, Page AJ, Connor TR, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research*. 2015;43(3):e15. doi:10.1093/nar/gku1196

63. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*. 2016;32(2):292-294. doi:10.1093/bioinformatics/btv566

64. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352

65. Seemann T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30(14):2068-2069. doi:10.1093/bioinformatics/btu153

66. Seemann T. *Shovill: Faster SPAdes Assembly of Illumina Reads*.; 2019. https://github.com/tseemann/shovill. Accessed January 22, 2019.

67. Carattoli A, Zankari E, Garcia-Fernandez A, et al. In Silico detection and typing of

plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrobial Agents and Chemotherapy*. 2014;58(7):3895-3903. doi:10.1128/AAC.02412-14

68. PHAC-NML. Curated virulence factors for Escherichia coli. https://github.com/phac-nml/ecoli_vf. Accessed January 22, 2019.

69. Seemann T. *Abricate: Mass Screening of Contigs for Antimicrobial and Virulence Genes*.; 2019. https://github.com/tseemann/abricate. Accessed January 21, 2019.

70. Zankari E, Hasman H, Cosentino S, et al. Identification of acquired antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy*. 2012;67(11):2640-2644. doi:10.1093/jac/dks261

71. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*. 2015;32(1):268-274. doi:10.1093/molbev/msu300

72. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. Ggtree: An R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data. *Methods in Ecology and Evolution*. 2017;8(1):28-36. doi:10.1111/2041-210X.12628

73. Freckleton RP, Harvey PH, Pagel M. Phylogenetic Analysis and Comparative Data : *The American naturalist*. 2002;160(6):712-726.

74. Seemann T. *Snp-Dists: Pairwise SNP Distance Matrix from a FASTA Sequence Alignment*. https://github.com/tseemann/snp-dists. Accessed January 21, 2019.

75. Stimson J, Gardy J, Mathema B, Crudu V, Cohen T, Colijn C. Beyond the SNP Threshold: Identifying Outbreak Clusters Using Inferred Transmissions. Leitner T, ed. *Molecular Biology and Evolution*. 2019;36(3):587-603. doi:10.1093/molbev/msy242

76. Dautzenberg MJD, Haverkate MR, Bonten MJM, Bootsma MCJ. Epidemic potential of Escherichia coli ST131 and Klebsiella pneumoniae ST258: A systematic review and meta-

analysis. *BMJ Open*. 2016;6(3):e009971. doi:10.1136/bmjopen-2015-009971

77. Drummond AJ, Suchard M a., Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*. 2012;29(8):1969-1973. doi:10.1093/molbev/mss075

78. Ender PT, Gajanana D, Johnston B, Clabots C, Tamarkin FJ, Johnson JR. Transmission of an extended-spectrum-beta-lactamase-producing Escherichia coli (sequence type ST131) strain between a father and daughter resulting in septic shock and emphysematous pyelonephritis. *Journal of Clinical Microbiology*. 2009;47(11):3780-3782. doi:10.1128/JCM.01361-09

79. Manges A. Escherichia coli and urinary tract infections: the role of poultry-meat. *Clinical Microbiology and Infection*. 2016;22(2):122-129. doi:10.1016/j.cmi.2015.11.010

80. Miles-Jay A, Weissman SJ, Adler AL, et al. Epidemiology and Antimicrobial Resistance Characteristics of the Sequence Type 131-H30 Subclone Among Extraintestinal Escherichia coli Collected From US Children. *Clinical Infectious Diseases*. 2018;66(3):411-419. doi:10.1093/cid/cix805

81. Kallonen T, Brodrick HJ, Harris SR, et al. Systematic longitudinal survey of invasive Escherichia coli in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. 2017:1-13. doi:10.1101/gr.216606.116.Freely

82. Banerjee R, Johnson JR. A new clone sweeps clean: The enigmatic emergence of Escherichia coli sequence type 131. *Antimicrobial Agents and Chemotherapy*. 2014;58(9):4997-5004. doi:10.1128/AAC.02824-14

83. Gardy JL, Johnston JC, Ho Sui SJ, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *The New England Journal of Medicine*. 2011;364(8):730-739. doi:10.1056/NEJMoa1003176

84. Grad YH, Kirkcaldy RD, Trees D, et al. Genomic epidemiology of Neisseria gonorrhoeae with reduced susceptibility to cefi xime in the USA : A retrospective observational study. *The*

Lancet Infectious Diseases. 2014;14(3):220-226. doi:10.1016/S1473-3099(13)70693-5

85. Mather a E, Reid SWJ, Maskell DJ, et al. Distinguishable epidemics of multidrug-resistant Salmonella Typhimurium DT104 in different hosts. *Science (New York, NY)*. 2013;341(2013):1514-1517. doi:10.1126/science.1240578

86. Murray GGR, Wang F, Harrison EM, et al. The effect of genetic structure on molecular dating and tests for temporal signal. Gilbert M, ed. *Methods in Ecology and Evolution*. 2016;7(1):80-89. doi:10.1111/2041-210X.12466

87. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*. 2018;4(1). doi:10.1093/ve/vey016

88. Minin VN, Bloomquist EW, Suchard MA. Smooth Skyride through a Rough Skyline: Bayesian Coalescent-Based Inference of Population Dynamics. *Mol Biol Evol*. 2008;25(7):1459-1471. doi:10.1093/molbev/msn090

89. Leaché AD, Banbury BL, Felsenstein J, De Oca ANM, Stamatakis A. Short tree, long tree, right tree, wrong tree: New acquisition bias corrections for inferring SNP phylogenies. *Systematic Biology*. 2015;64(6):1032-1047. doi:10.1093/sysbio/syv053

90. Rambaut A, Suchard M, Xie D, Drummond A. Tracer v1.6. 2014. `http://beast.bio.ed.ac.uk/Tracer`.

91. Edwards CJ, Suchard MA, Lemey P, et al. Ancient hybridization and an irish origin for the modern polar bear matriline. *Current Biology*. 2011;21(15):1251-1258. doi:10.1016/j.cub.2011.05.058

92. Lemey P, Rambaut A, Drummond AJ, Suchard M a. Bayesian phylogeography finds its roots. *PLoS Computational Biology*. 2009;5(9). doi:10.1371/journal.pcbi.1000520

93. Minin VN, Suchard M a. Fast, accurate and simulation-free stochastic mapping. *Philosophical transactions of the Royal Society of London Series B, Biological sciences*.

2008;363(1512):3985-3995. doi:10.1098/rstb.2008.0176

94. Minin VN, Suchard M a. Counting labeled transitions in continuous-time Markov models of evolution. *Journal of Mathematical Biology*. 2008;56(3):391-412. doi:10.1007/s00285-007-0120-8

95. TreeAnnotator | BEAST Documentation. http://beast.community/treeannotator. Accessed February 8, 2019.

96. Bouckaert R, Heled J, Kühnert D, et al. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. Prlic A, ed. *PLoS Computational Biology*. 2014;10(4):e1003537. doi:10.1371/journal.pcbi.1003537

97. De Maio N, Wu C-H, O'Reilly KM, Wilson D. New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. *PLOS Genetics*. 2015;11(8):e1005421. doi:10.1371/journal.pgen.1005421

98. Hsu AJ, Tamma PD. Treatment of multidrug-resistant gram-negative infections in children. *Clinical Infectious Diseases*. 2014;58:1439-1448. doi:10.1093/cid/ciu069

99. Blahna MT, Zalewski CA, Reuer J, Kahlmeter G, Foxman B, Marrs CF. The role of horizontal gene transfer in the spread of trimethoprim-sulfamethoxazole resistance among uropathogenic Escherichia coli in Europe and Canada. *Journal of Antimicrobial Chemotherapy*. 2006;57(4):666-672. doi:10.1093/jac/dkl020

100. Gray RR, Tatem AJ, Johnson JA, et al. Testing spatiotemporal hypothesis of bacterial evolution using methicillin-resistant staphylococcus aureus ST239 genome-wide data within a bayesian framework. *Molecular Biology and Evolution*. 2011;28(5):1593-1603. doi:10.1093/molbev/msq319

101. Grad YH, Lipsitch M. Epidemiologic data and pathogen genome sequences: A powerful synergy for public health. *Genome Biology*. 2014;15(11):538. doi:10.1186/s13059-014-0538-4