Genomic Insights Into Carbapenem-Resistant *Klebsiella pneumoniae* Transmission and Adaptation in the Healthcare Environment

by

Zena Lapp

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2021

Doctoral Committee:

Assistant Professor Evan Snitkin, Chair
University Professor Michael Boehnke
Assistant Professor Peter Freddolino
Associate Professor Jenna Wiens
Assistant Professor Jon Zelner

Zena Lapp

zenalapp@umich.edu

ORCID: 0000-0003-4674-2176

Dedication

To my family

&

To my undergraduate mentor

Dr. Melissa Schultz

1977-2015

Acknowledgements

First, thanks to my funding: the National Science Foundation Graduate Research Fellowship Program, the Genome Science Training Program, and Rackham Graduate School. Any opinions, findings, and conclusions or recommendations expressed in this material are mine and do not necessarily reflect the views of the National Science Foundation.

And more importantly, I am very grateful for all of the people who have helped and supported me along the way:

My undergraduate mentors at the College of Wooster who were the first to spark my excitement about science, and who taught me to think critically about every aspect of my work: Robert Woodward, Dean Fraga, Melissa Schultz, Mark Snider, and Stephanie Strand. I probably wouldn't have majored in Biochemistry and Molecular Biology if it hadn't been for the enthusiasm of Drs. Woodward and Fraga in my introductory chemistry and biology classes, respectively. Dr. Schultz taught me the importance of rigorous study design. Dr. Snider's passion for science was inspiring; I think he knew I was going to pursue a PhD before I did. Dr. Strand is one of the kindest people I have ever met, and I appreciate her encouragement to take a "gap year" after graduating from Wooster. I wouldn't be here without you all.

The PIBS, DCMB, and MSRB I administrative teams for their support and tolerance of all my questions. Particularly Michelle DiMondo, Julia Eussen, and Karrie Black. You made grad school so much easier.

My rotation mentors. Melissa Duhaime for extensive help with my NSF GRFP applica-

tion and being a role model for incorporating outreach and community engagement into all aspects of your work, and Stephen Smith for explaining the nuances of phylogenetic analysis.

Jennifer Han and Ebbing Lautenbach for feedback on, and insights into, the clinical relevance of my projects, and for the data that made my dissertation possible.

The patients and hospital staff who participated in the study that I use in this dissertation.

My committee for thoughtful insights: Michael Boehnke, Peter Freddolino, Jenna Wiens, and Jon Zelner. Particularly Jenna for guidance on everything related to machine learning.

Ali Pirani for being a bioinformatics superstar and helping me out every step of the way. If it weren't for you, I think my PhD would have taken at least a year longer.

My package development buddies: Katie Saund, Stephanie Thiede, Begüm Topçuoğlu, Kelly Sovacool, and Sophie Hoffman. I couldn't have done it without you.

Evan Snitkin for his guidance, thoughtful questions, patience for my unending questions, and assembly of an amazing group of Snitkineers. Thank you for teaching me so much and making my PhD a ton of fun!

The Snitkineers for discussions on research and life. You are such amazing, friendly, and helpful people. In particular, Stephanie Thiede for being my desk neighbor, friend, responsive rubber duck, and emotional support.

My friends from all walks of life. You have helped open my eyes to the world and have made everything so much more enjoyable and exciting.

My mom, dad, siblings Sam and Maya, extended family, and partner Alex, for their support. Regardless of how much they understand, or even care, about my research, they are always there for me. Thanks for making everything better, and for great times together.

Table of Contents

List of Figures

List of Tables

Abstract

Multidrug resistant organisms (MDROs) pose a threat to healthcare facilities worldwide due to global prevalence, limited treatment options, and high mortality rates. This thesis acts as a proof-of-principle study for using whole-genome sequencing and associated clinical metadata to provide actionable insights into MDRO infection prevention and control practices. We focus our analysis on carbapenem-resistant *Klebsiella pneumoniae* (CRKP) sequence type (ST) 258, an MDRO that is particularly prevalent in long-term acute care hospitals (LTACHs) in the United States (US). Using a comprehensive set of 417 clinical CRKP ST258 isolates collected over the course of a year in 21 US LTACHs, we investigate regional transmission, predictors of infection, and evolution of antibiotic resistance. In addition, we develop three open-source R packages that implement methods developed and applied here: `regentrans` for studying regional pathogen transmission, `mikropml` for performing machine learning, and `prewas` for preprocessing data prior to bacterial genome-wide association studies.

First, we reconstructed regional transmission pathways with genomic data and analyzed this network in the context of patient transfer data and patient-level clinical data to identify potential drivers of regional CRKP transmission. We found high regional CRKP burdens in Los Angeles area LTACHs that were due to a small number of introductions with subsequent proliferation occurring via within-facility transmission and patient transfers among healthcare facilities.

As only a subset of colonized patients develop clinical infection, we next used machine

learning to determine whether patient characteristics and CRKP genetic background can predict infection status. We found that patient and genomic features were predictive of clinical CRKP infection to similar extents. Genomic predictors of infection included presence of the ICEKp10 mobile genetic element carrying the yersiniabactin iron acquisition system and disruption of the O-antigen biosynthetic gene *kfoC* in a CRKP ST258 sublineage (clade IIB). Disrupted *kfoC* was associated with isolation from the respiratory tract, and subsequent ICEKp10 acquisition was associated with increased virulence. These results highlight the utility of machine learning to provide insight into patient clinical trajectories and ongoing within-lineage pathogen adaptation.

Finally, we investigated the evolution of CRKP resistant to the antibiotic colistin, one of the few remaining treatment options for this MDRO. Two large clusters of resistant strains in clade IIB accounted for over half of the detected colistin resistance, in stark contrast to the sporadic resistance events observed in other clades. Moreover, while resistant isolates from other clades were less fit than susceptible non-revertant isolates, clade IIB resistant isolates were more fit, underscoring the potential for continued regional spread of clade IIB colistin resistant strains.

In summary, we identified an emerging CRKP sublineage that has spread across Los Angeles area LTACHs and appears to have an increased affinity for the respiratory tract, increased transmissibility, and a decreased fitness cost of colistin resistance. Future work should continue to monitor this strain as increased prevalence could reduce the efficacy of colistin as a treatment for CRKP in this region and possibly lead to exportation to other regions. Additionally, these findings highlight the potential impact of incorporating genomic epidemiology into regional infection prevention efforts to identify high-transmission facilities and emerging strains, thereby facilitating the containment of MDROs to the greatest extent possible.

Chapter 1

Introduction

1.1 Motivation

Multidrug resistant organisms (MDROs) are a global public health threat due to rampant spread, high levels of morbidity and mortality, and continued evolution to additional antibiotics [1]. Monitoring and surveillance of MDROs requires methods for analysis of transmission, adaptation to the healthcare environment, and evolution of antibiotic resistance. This dissertation acts as a proof-of-principle study for how to integrate whole-genome sequences of clinical isolates with corresponding patient metadata to investigate endemic, yet continually evolving, public health threats. Specifically, we study transmission, infection, and antibiotic resistance evolution of the nosocomial pathogen carbapenem resistant *Klebsiella pneumoniae* (CRKP). The bioinformatic framework and corresponding tools developed here can be applied to investigate the transmission and evolution of other nosocomial pathogens.

1.2 CRKP history and population genomics

K. pneumoniae was first isolated from a patient with pneumonia in the late 1800s [2]. This Gram-negative rod-shaped bacterium can colonize the gastrointestinal tract and cause pneumonia, urinary tract infections, and bacteremia, primarily in immunocompromised individ-

uals in the healthcare setting [3]. In the past several decades, some *K. pneumoniae* have acquired antibiotic resistance elements that make them very difficult to treat. Of particular concern are CRKP, which are strains that have become resistant to carbapenems [4].

The majority of CRKP in the United States (US) harbor *K. pneumoniae* carbapenemase (KPC), which confers resistance to carbapenems [5]. KPC was first discovered in a 1996 *K. pneumoniae* isolate [6]. In 2003, outbreaks of CRKP occurred across New York healthcare facilities, and by 2005 CRKP had spread to several other countries including Israel, Italy, Colombia, the United Kingdom, and Sweden [7]. CRKP are now present in countries worldwide [5] and pose a large threat to healthcare facilities both in the US and abroad due to limited treatment options and mortality rates of up to 40% [8].

KPCs are encoded by the blaKPC gene and are harbored on the transposon Tn4401, allowing it to associate with many different plasmids and spread across many species [9]. However, the vast majority of KPC positive clinical isolates come from a handful of *K. pneumoniae* sequence types (STs), with the majority in the US being ST258 [9]. In addition to carbapenems, KPC positive ST258 isolates are often resistant to the majority of antibiotics on the market, leaving few treatment options available [4].

1.3 Methods to study CRKP regional transmission

Due to the limited treatment options for individuals with CRKP, infection prevention is a particularly important component of reducing prevalence and mortality. To improve infection prevention efforts, we must better understand where transmission is occurring. Most studies and intervention measures to date have focused on investigating and reducing transmission within a single healthcare facility, but increasing evidence suggests that regional transmission among healthcare facilities is common [10]. Therefore, it is important not only to understand transmission dynamics within a single facility, but also how and where transmission occurs

between facilities.

Efforts to study regional MDRO transmission across healthcare facilities have primarily used mathematical modeling to probe MDRO spread due to patient transfer between facilities [11, 12]. One strength of these studies is that they are able to investigate the impact of interventions on MDRO prevalence at different facilities. These studies have found that coordinated regional control efforts lead to a much lower prevalence across regional healthcare networks when compared to uncoordinated control efforts at individual facilities [13]. Simulations have also shown that targeted interventions at certain facilities, such as long-term acute care hospitals (LTACHs), also decrease regional prevalence [14]. This is particularly important as interventions to reduce transmission cannot practically be implemented at all facilities in a regional network.

While modeling studies provide important insight into putative transmission dynamics between regional healthcare facilities and allow us to probe the impact of various interventions, models are by definition simplified versions of reality that make generalized assumptions and cannot capture all of the nuances that may influence true MDRO transmission networks. As such, they do not allow us to investigate in detail actual networks of MDRO transmission, nor specific transmission events. Epidemiological studies overcome some of these limitations and are often used to investigate MDRO transmission within a given healthcare facility, and sometimes between healthcare facilities [15]. These studies usually perform contact tracing to determine what patients and healthcare workers overlapped in the same location, and thus may have transmitted the organism to one another. While these analyses are very important and can be informative, they are extremely time intensive, and it is often infeasible to determine the exact transmission pathway due to multiple overlap events between different patients or healthcare workers, particularly in endemic settings [16].

One way to overcome the limitations of modeling and epidemiological studies is to use whole-genome sequencing of isolates to gain a more nuanced understanding of the relatedness

between different patient isolates [17]. This type of analysis is particularly informative when combined with additional epidemiological information to investigate transmission. For instance, a previous study performed a retrospective analysis combining patient transfer data with whole-genome sequencing to gain a more detailed understanding of regional transmission during a CRKP outbreak [18]. Integrating genomic and epidemiologic data allowed for a more detailed understanding of directionality of transmission between facilities and identified an intermediate facility that was important for regional spread.

While whole-genome sequencing has been effectively applied to studying regional transmission in an outbreak setting, few studies have used this method to study endemic transmission both within and between healthcare facilities. One study investigated methicillin resistant *Staphylococcus aureus* transmission across the United Kingdom and found that, while within facility transmission was common, spread between facilities also regularly occurred [19]. However, non-comprehensive sampling limited their ability to compare prevalence at different facilities and quantify the extent of transmission between facilities. Furthermore, they did not incorporate patient clinical data into their transmission analysis, which could provide additional insight into patient-level drivers of regional spread.

1.4 Methods to study CRKP evolution and adaptation

In addition to investigating transmission, it is also important to understand how nosocomial pathogens adapt to the healthcare environment to better inform infection prevention and control practices. ST258 contains two clades (clade I and clade II) with distinct genetic backgrounds that have evolved independently of one another since around 2001 [20, 21, 22]. Clade I may have evolved from a clade II strain due to a recombination event in the K antigen capsule polysaccharide biosynthesis gene region. Additionally, clade I ST258 strains usually harbor a plasmid containing KPC-2, while clade II ST258 strains usually harbor a

plasmid carrying KPC-3. Furthermore, continued evolution of ST258 within the healthcare setting leads to the persistence of some sublineages over time [23]. Adaptations that increase virulence or cause antibiotic resistance are of particular clinical concern, as they may lead to adverse patient outcomes [24].

Many studies investigating healthcare pathogen evolution have used experimental methods to study pathogen adaptation in a more controlled laboratory setting, either *in vitro* or in model systems [25, 26, 27]. For instance, recent experimental results from studies in mice indicate that different global CRKP strains are differentially virulent due to mutations in, or disruption of, capsule biosynthesis genes [26]. Additionally, experimental evolution of laboratory or clinical strains can shed light on trajectories of antibiotic resistance evolution [28, 29]. These experiments provide valuable insights into a strain's virulence and resistance potential, and are critical for furthering our understanding of the molecular mechanisms underlying these phenotypes. However, they do not capture the myriad of potential selective pressures and interactions present in the healthcare setting that may influence evolutionary trajectories and strain phenotypes in the natural environment, and do not provide insight into the transmissibility of these strains among patients.

One way to overcome some of the limitations of laboratory studies is to investigate pathogen evolution using comparative genomics of global clinical isolate collections. This method identified genes associated with virulence and antibiotic resistance in global *K. pneumoniae* isolates [30]. Genome-wide association studies have also been extensively used to identify the genotypic variation underlying resistance to specific antibiotics [31]. A strength of these methods is that they derive inferences from evolutionary events that truly occurred in the healthcare setting. However, the majority of these studies to date have used a global sampling of isolates with little additional clinical metadata, limiting their ability to investigate the relationship between genomic variants, phenotypes, patient features, and strain fitness. Furthermore, a global sample of isolates by and large only captures the successful

end points of evolution. In contrast, capturing both successful and unsuccessful intermediates may yield information regarding various evolutionary trajectories, which could provide insight into how to prevent the emergence of new epidemic strains.

1.5 Dataset used in this dissertation

To study the transmission and evolution of CRKP, we use a comprehensive set of clinical isolates collected over the course of a year (August 2014 to July 2015) during a prospective observational study in 21 LTACHs across the US. For each clinical isolate, we have Illumina whole-genome sequencing data (BioProject accession no. PRJNA415194), antimicrobial susceptibility testing data for several antibiotics, and associated clinical metadata including patient comorbidities and previous antibiotic exposure. In addition, we have aggregate annual patient transfer counts between facilities from Centers for Medicare and Medicaid claims data for 2016. The vast majority (>90%) of isolates in this dataset are ST258; therefore, we subset most of our analyses in all three chapters to ST258 to gain insight into the nuances of this specific CRKP sequence type. The whole-genome sequencing data and clinical metadata were used for all three chapters, the patient transfer data was used for chapter two, and the antimicrobial susceptibility testing data was used for chapter four. More details about the data can be found in the corresponding chapters.

1.6 Dissertation outline

This dissertation investigates transmission, infection, and antibiotic resistance evolution in CRKP ST258. In chapter two I investigate how CRKP spreads within and between LTACHs. In chapter three I uncover potential genomic drivers of colonization and infection. In chapter four I identify distinct patterns of colistin resistance evolution and dissemination in different

genetic backgrounds. For chapters two through four, I also developed a corresponding open source R package to increase the ease with which other investigators can apply the methods developed here to their own analyses. In chapter five I discuss bioinformatic, genomic, and clinical implications of the findings from chapters two through four, and propose future directions based on these implications.

Chapter 2

Whole-Genome Sequencing to Identify Drivers of Carbapenem-Resistant *Klebsiella pneumoniae* Transmission Within and Between Regional Long-Term Acute-Care Hospitals

2.1 Preamble

This chapter uses whole-genome sequencing and aggregate patient transfer data to investigate intra- and inter-facility transmission of CRKP in a regional network of LTACHs. We identify several regional importation events, detect signatures of inter-facility transmission, and find that facility prevalence is driven by intra-facility transmission. This transmission analysis inspired us to create `regentrans` (https://github.com/Snitkin-Lab-Umich/regentrans), an R package for studying regional pathogen transmission.

I performed all of the data analysis and generated all off the figures for this chapter. Other co-authors performed isolate collection, detection of KPC-positive *K. pneumoniae*, antibiotic susceptibility testing, whole-genome sequencing, and single nucleotide variant identification. This work was published in Antimicrobial Agents and Chemotherapy in 2019:
Han JH, Lapp Z, Bushman F, Lautenbach E, Goldstein EJ, Mattei L, Hofstaedter CE, Kim D, Nachamkin I, Garrigan C, Jain T. Whole-genome sequencing to identify drivers of

carbapenem-resistant Klebsiella pneumoniae transmission within and between regional long-term acute-care hospitals. Antimicrobial agents and chemotherapy. 2019 Nov 1;63(11).

I developed, or helped develop, most of the methods implemented in `regentrans` and wrote draft functions for each method. `regentrans` has just been developed and does not yet have use statics; however, several public health departments involved in the Centers for Disease Control and Prevention Emerging Infections Program have expressed an interest in using the package to perform their own analyses. The manuscript corresponding to the `regentrans` package will be submitted for publication with the following co-authors:
Sophie Hoffman*, Zena Lapp*, Joyce Wang, and Evan Snitkin.
*Indicates co-first author

2.2 Introduction

Since first being reported in 2001 [6], epidemic lineages of carbapenem-resistant Klebsiella pneumoniae (CRKP) have rapidly emerged and disseminated across global healthcare systems [32]. The limited treatment options for patients with CRKP infections has created an urgent need for more effective strategies to prevent CRKP transmission [33, 34, 35, 36]. Historically, prevention of CRKP and other multi-drug resistant organisms (MDROs) has been undertaken in a siloed manner, with interventions being applied at the level of individual healthcare facilities. However, it is increasingly appreciated that healthcare facilities are highly connected to one another, with patients typically moving between different healthcare facilities as part of their treatment and recovery [10].

Support for a regional approach to infection prevention comes from the success of coordinated national interventions in controlling high-priority MDROs [37]. While a coordinated national intervention is not currently logistically feasible in the United States, the infrastruc-

ture for regional infection prevention is being advanced, particularly in collaboration with regional and state public health departments [38, 39, 40]. What is now needed to facilitate effective regional interventions are strategies to identify facilities with the highest transmission rates, methods to quantify how transmission at each facility influences colonization and infection rates across the healthcare network and insights into the clinical and epidemiologic drivers of both intra- and inter-facility transmission. With this level of understanding of regional MDRO transmission, targeted interventions can be designed that leverage limited resources to achieve the maximal possible decrease in regional MDRO prevalence [13]. In a remarkably short time, whole-genome sequencing (WGS) has fundamentally changed the resolution at which one can study the spread of bacterial pathogens [41]. Moreover, through integration of genomic data with epidemiologic meta-data, the processes driving the spread of these pathogens can be more clearly understood [42]. With respect to healthcare-associated infections, recent work has shown that by integrating patient transfer data into genomic analyses, the putative sites of transmission and the patient movements driving regional spread can be identified [18, 43]. Here, we seek to take the next step in the application of genomic epidemiology to regional infection prevention by integrating comprehensive genomic data, statewide patient transfer networks and patient-level clinical meta-data to identify drivers of the regional spread of CRKP.

2.3 Results

2.3.1 Overview of CRKP isolates from twenty long-term acute care hospitals

Due to their chronically, critically ill patient population, long-term acute care hospitals (LTACHs) are increasingly recognized as both major amplifiers and reservoirs of MDROs,

including CRKP [44, 45]. To better understand the epidemiology of CRKP in LTACHs across the United States we gathered clinical isolates and meta-data from LTACH patients in four geographic regions, including a region with a high baseline prevalence of CRKP [46]. A total of 451 unique CRKP culture episodes across the four geographic regions were identified during the 12-month study period, with the majority (n=395; 87.6%) from eleven Los Angeles area LTACHs (Table 2.S1). Demographic and clinical characteristics of patients with CRKP clinical cultures are shown in Table 2.S2. Patients with CRKP had a high rate of multiple comorbidities, and indwelling device and broad-spectrum antibiotic use was common. Antibiotic susceptibility results for CRKP isolates are shown in Table 2.S3. Nearly all of the isolates (>90%) demonstrated resistance to levofloxacin, ciprofloxacin, and tobramycin. Resistance rates to colistin or polymyxin B were about 12%.

2.3.2 Phylogeographic analysis demonstrates that CRKP burden in the Los Angeles area is largely due to endemic spread of the ST258 epidemic lineage

Inference of CRKP sequence types from WGS data revealed that greater than 90% of LTACH isolates belonged to the ST258 epidemic lineage (Table S4). To discern how circulating LTACH strains relate to ST258 clones from other regions, a phylogeographic analysis was performed using genome-wide variants identified in LTACH isolates as well as previously sequenced ST258 genomes with known geographic origins (Table 2.S5). The whole-genome phylogenetic reconstruction in Figure 2.1 shows that the common ancestor of analyzed genomes dates back to 2001, which is consistent with the timing of the first clinical observation of ST258 [47]. In addition, the phylogeny shows that cities in the Northeastern United States played a prominent role in the early dissemination of this lineage to disparate geographic areas. Focusing on the Los Angeles area LTACHs where we observe the high-

est prevalence of CRKP, the reconstruction points to a small number of introductions into the region as recently as 2012, with subsequent widespread dissemination across regional healthcare networks.

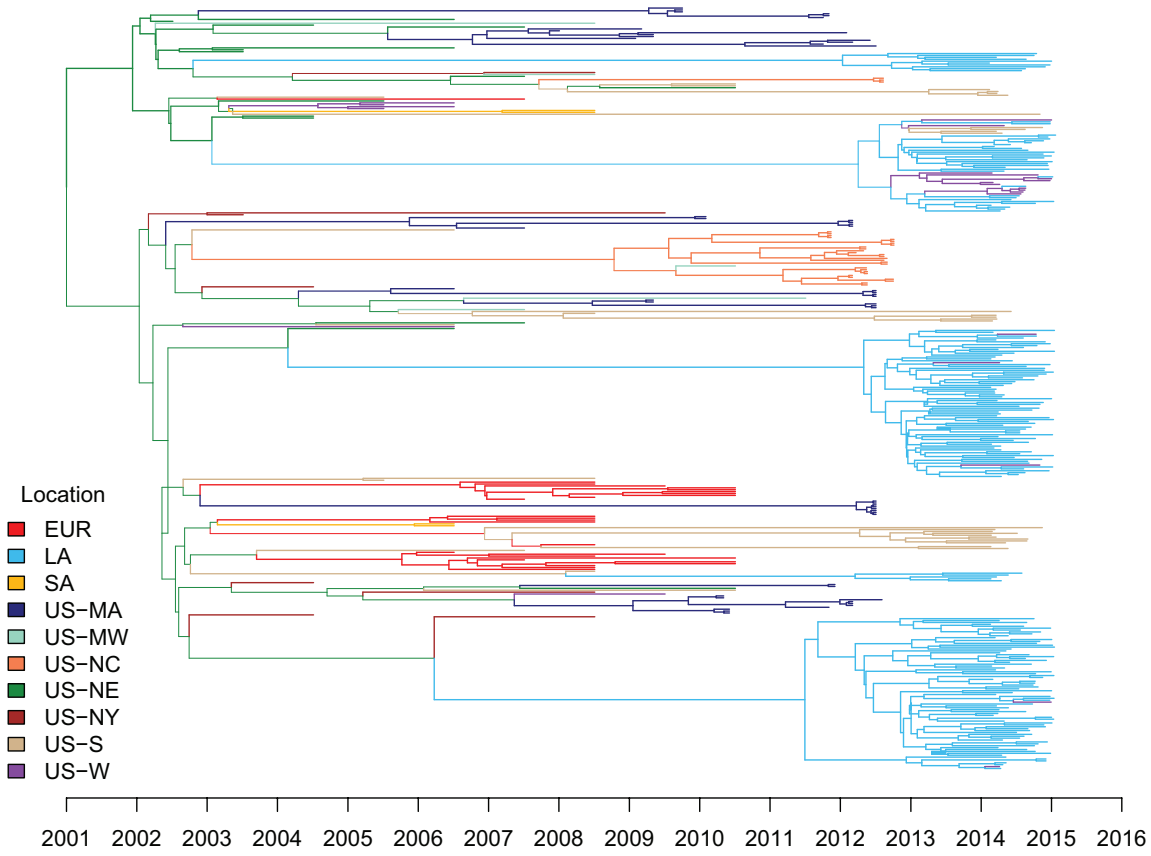

Figure 2.1: Phylogeographic reconstruction of ST258 *Klebsiella pneumoniae*.

A phylogeographic analysis was performed on ST258 isolates from the current study, along with previously sequenced ST258 genomes that had documented locations and dates of isolation (See Table 2.S3). The x-axis corresponds to predicted dates of ancestral strains and the colors of internal branches correspond to the predicted locations of ancestral strains. Abbreviations: EUR: Europe; LA: Los Angeles; SA: South America; US-MA: Massachusetts; US-MW: Midwest; US-NC: North Carolina; US-NE: North East; US-NY: New York; US-S: South; US-W: West.

2.3.3 Transmission rates between facilities are associated with patient transfer rates

To study regional transmission in more detail we next focused on Los Angeles LTACHs, for which we had the largest number of isolates and a comprehensive sampling from eleven different LTACHs. An LTACH-centric transmission map was constructed using 363 isolates that represented independent ST258 acquisitions among 328 patients (Figure 2.2A) by first performing ancestral reconstruction of LTACHs on the whole genome phylogeny. Intra- and inter-facility transmission frequencies were then extracted from the ancestral reconstruction as transitions between the same and different facilities, respectively (See Methods and Figures 2.S1 and 2.S2). Inspection of the map of inter-facility transmissions indicated higher inter-facility transmission frequencies among geographically proximate facilities, which was supported by a significant association between geographic proximity and transmission frequency (Figure 2.S3, Mantel test p = 7 x 10-5). We hypothesized that the association between transmission and geographic proximity was due to patient transfers preferentially occurring between geographically proximate healthcare facilities [19]. To test this hypothesis, we calculated the direct and indirect flow of patients between each pair of LTACHs using a California statewide patient transfer network (See Methods). This patient transfer network captures sequential utilization of different healthcare facilities based on Centers for Medicare and Medicaid billing data and thereby provides quantitative information on how healthcare facilities are connected to one another. As patients rarely transfer directly between LTACHs, the connectivity between pairs of LTACHs was quantified using the maximal patient flow path, regardless of the number of intervening facilities (See Methods). Comparison between patient transfer and genomic networks confirmed that patient flow is strongly associated with inter-facility transmission (Figure 2.S3). Of note, while the length and facility makeup of the patient-transfer paths connecting LTACHs with and without transmission are similar

(Figure 2.2B), the magnitude of patient flow is significantly higher for pairs of LTACHs with predicted genomic transmission linkages (Figure 2.2C). Moreover, the association between genomically inferred inter-facility transmission and patient flow remains significant even when there are multiple intermediate healthcare facilities connecting two LTACHs (Figure 2.2C), indicating that indirect connections mediate regional spread over short time periods.

2.3.4 Facility-level CRKP prevalence is associated with intra-facility transmission rates

With our genomic analysis revealing frequent transmission of CRKP between facilities, we next evaluated whether facilities varied with regards to the magnitude of subsequent intra-facility transmission associated with each importation event. Using the ancestrally reconstructed whole genome phylogeny (Figure 2.S1), we quantified the fraction of patient isolates attributed to intra-facility transmission versus importation (i.e., inter-facility transmission) from another LTACH, revealing extensive variation in the ratio of intra-facility transmission events per importation event (See Methods and Figure 2.S4). Moreover, those facilities with the highest intra-facility transmission to importation ratio ("transmission ratio") had the highest CRKP prevalence (colonization or infection), suggesting that higher intra-facility transmission rates drove facility-level variation in CRKP prevalence (Figure 2.3, Spearman $R = 0.75$, $p = 0.012$).

2.3.5 Differences in intra-facility transmission ratios are associated with clinical characteristics of patients with CRKP

Notably, facility G had a markedly lower ratio of intra-facility transmission per importation than all other facilities (Figure 2.3 and Figure 2.S5). While it was possible that the reduced transmission ratio in facility G could have been due to improved adherence to infection pre-

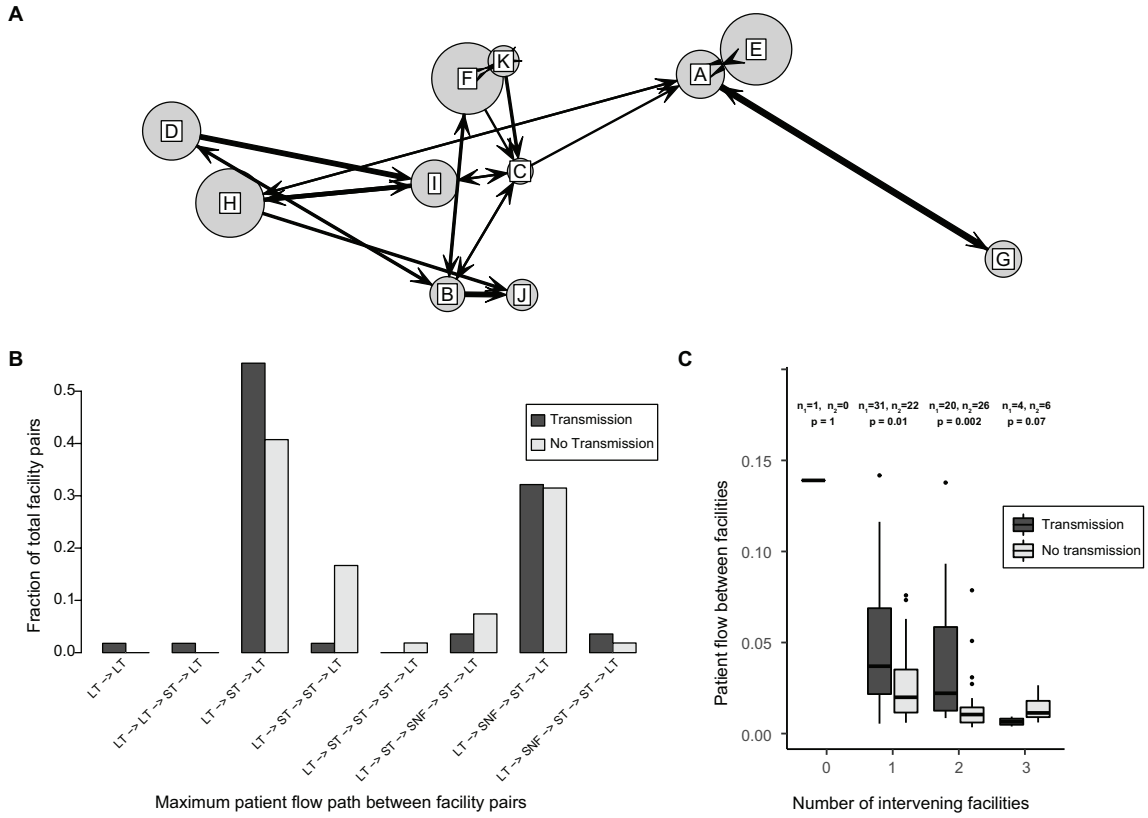

Figure 2.2: Regional transmission map for carbapenem-resistant *K. pneumoniae* (CRKP) among eleven Los Angeles area LTACHs.

(A) Each gray circle represents a single LTACH, with the size of the circle corresponding to the CRKP colonization/infection rate in that facility. The relative positions of the different circles on the graph are based on their actual longitude and latitude. Arrows between different LTACHs indicate that there was at least one predicted transmission between the pair of facilities, with the thicker lines corresponding to larger numbers of predicted transmissions. To highlight inter-facility transmissions between facilities with more linkages than expected by chance, only links between facilities whose isolates cluster on the whole genome phylogeny more than in randomly permuted phylogenies are shown ($P < 0.1$; See Methods). (B) Paths of maximum patient flow between each pair of LTACHs were extracted from the statewide patient transfer network and categorized into path motifs based on the types of facilities on the paths (LT – long-term acute care hospital, ST – short-stay hospital, SNF –skilled nursing facility). The fraction of LTACH pairs (y-axis) whose maximum flow paths were assigned to each path motif (x-axis) was determined for LTACH pairs with and without genomic transmission linkages. (C) Patient flow between LTACHs was compared for pairs of LTACHs with and without genomic transmission linkages. Pairs of LTACHs were binned into groups based upon the number of intervening facilities on the maximal patient flow path between those LTACHs (e.g. 0 – direct transfer between LTACHs, 1 – one intervening facility, etc.). Significance was assessed using a Wilcoxon rank-sum test.

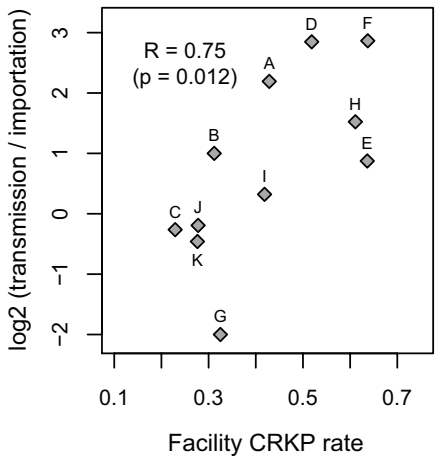

Figure 2.3: Facility colonization/infection rate associated with ratio of intra-facility transmissions per importation.

Each diamond represents one of the eleven Los Angeles area LTACHs, with the x-axis showing the colonization/infection rate within the LTACH and the y-axis showing the number of intra-facility transmissions per importation, as inferred from the genomic transmission network (See Figure 2.S3). The CRKP colonization/infection rate for a facility was calculated as the total number of colonization/infection episodes divided by the number of beds.

vention measures, we hypothesized that it may also be that the clinical characteristics of facility G's CRKP-positive patients make them less likely to transmit to other patients. To test whether the decreased transmission ratio in facility G could be due to clinical characteristics of CRKP-positive patients, we performed post hoc analyses comparing CRKP population characteristics between patients from facility G and the three facilities with the highest transmission ratios (facilities A, D, and F – Table 2.1). Patients with CRKP in the facilities with high intra-facility transmission ratios compared to those in facility G were of significantly older age (median 72 years versus 57 years, respectively; $p = 0.008$); had higher rates of certain comorbidities including chronic kidney disease and/or requirement for dialysis (43% versus 19%, respectively; $p = 0.045$), and a diagnosis of malnutrition and/or being underweight (34% versus 0%, respectively; $p = 0.003$); and had higher rates of carbapenem use in the 30 days prior to CRKP isolation (44% versus 13%, respectively; $p = 0.02$). To assess whether these differences in patient characteristics applied to patients without CRKP,

we conducted post hoc analyses of data collected from patients with carbapenem-susceptible Klebsiella pneumoniae (CSKP) infection at the same four facilities. Comparison of clinical characteristics of CSKP patients from low and high CRKP transmission facilities revealed only a difference in their rate of chronic kidney disease, but in the opposite direction as with CRKP patients (Table 2.S6). This result supported the lower transmission ratio at facility G being driven in part by clinical characteristics of putative transmitters (e.g., CRKP-positive patients), and not of the susceptible population (i.e., CSKP patients). Lastly, to determine whether the association between clinical characteristics of CRKP patients and transmission-ratio extended to all LTACHs, we compared the clinical characteristics of CRKP and CSKP patients to the CRKP transmission-ratio using Spearman rank correlations (Figures 2.S6, 2.S7, 2.S8 and 2.S9) and found consistent, although not statistically significant, trends for age (CRKP: R = 0.53, p = 0.1; CSKP R = 0.24, p = 0.49), chronic kidney disease (CRKP: R = 0.6, p = 0.056; CSKP R = -0.29, p = 0.38), malnutrition (CRKP: R = 0.58, p = 0.066; CSKP R = 0.36, p = 0.27) and carbapenem use (CRKP: R = 0.46, p = 0.15; CSKP R = -0.17, p = 0.61).

2.4 Discussion

We performed WGS on 450 CRKP isolates from patients residing in four regional networks of LTACHs. Placing these isolates in the context of previously sequenced clinical isolates from around the world suggested a small number of importation events into each region, followed by widespread dissemination across regional healthcare networks. Focusing on the Los Angeles area LTACHs, where CRKP is highly prevalent, we observed that the density of transmission between LTACHs was associated with the rates of patient flow between those facilities. However, despite inter-LTACH transmission being common, our analysis suggests that variation in CRKP prevalence across LTACHs was driven by differences in

Table 2.1: Clinical characteristics of patients with CRKP from facilities with high (facilities A, D and F) versus low (facility G) ratios of intra-facility transmissions per importation.

Variable, n (%)	**High intra-facility transmission**	**Low intra-facility transmission**	**P value**
Demographics	**A, D, F n = 163**	**G n = 16**	
Age, median (IQR)	72 (65, 82)	57 (47, 76)	**0.008 ***
Male sex	93 (57)	5 (31)	0.07
LTACH LOS prior to culture, median (IQR)	26 (7, 53)	25 (10, 32)	0.58
Comorbidities			
Congestive heart failure	32 (20)	1 (6)	0.31
Cirrhosis	12 (7)	0 (0)	0.61
Malignancy (solid or liquid)	19 (12)	1 (6)	>0.99
Brain injury	36 (22)	2 (13)	0.53
Pulmonary disease	37 (23)	1 (6)	0.2
Acute or chronic respiratory failure	71 (44)	8 (50)	0.79
Severe chronic kidney disease (stage IV or ESRD on dialysis)	70 (43)	3 (19)	**0.045***
Ventilator-dependent respiratory failure	54 (33)	2 (12)	0.16
Malnutrition/underweight	56 (34)	0 (0)	**0.003***
Stage IV/V decubitus ulcer	40 (25)	2 (13)	0.37
Antibiotic use in prior 30 days, receipt of $\geq$ 48 hours			
Carbapenem	72 (44)	2 (13)	**0.02***
Fluoroquinolones	25 (15)	3 (19)	0.72
Aminoglycosides	42 (26)	1 (6)	0.12
Cefepime	33 (20)	3 (19)	>0.99
Polymyxin/colistin	25 (15)	0 (0)	0.13
Anti-pseudomonal	104 (64)	8 (50)	0.29
Indwelling devices			
Tracheostomy	126 (77)	9 (56)	0.06
Central venous catheter	101 (62)	9 (56)	0.78
Urinary catheter	96 (59)	8 (50)	0.6

Abbreviations: CRKP, carbapenem-resistant *K. pneumoniae*; IQR, interquartile range; LTACH, long-term acute care hospital; LOS, length of stay; ESRD, end-stage renal disease.

intra-facility transmission rates. Lastly, we found that the facilities with the highest and lowest predicted rates of intra-facility transmission differed in the clinical characteristics of their CRKP patients, supporting the hypothesis that characteristics of CRKP patients may contribute to risk of transmission. In total, these findings demonstrate the capacity for WGS to reveal where regional transmission of endemic MDROs is occurring, and through integration with clinical and epidemiologic meta-data to provide insight into the external and internal forces driving MDRO colonization and infection rates within healthcare facilities.

Consistent with previous studies, we observed that patient transfer rates between healthcare facilities were associated with inter-facility transmission (i.e., importation) [19]. However, we noted that while the association between inter-facility transmission and patient flow was robust, it was imperfect, with multiple instances of high transmission rates between facilities with low patient flow and low transmission rates between facilities with high patient flow (Figure 2.S3). There are potential technical explanations for these deviations, including patients with CRKP traversing healthcare networks differently than patients from the utilized Centers for Medicare & Medicaid Services claims database as a whole, limitations in the use of billing data to capture patient movement or unsampled CRKP cases leading to missed inter-facility transmissions. Given the common practice of using patient transfer data to model regional transmission of MDROs [11, 48], it will be important to understand these discrepancies and determine whether patient transfer patterns alone can be used to model endemic spread. A potential limitation in relying solely on patient transfer patterns is a lack of consideration of uneven risk for transmission among different patient populations and healthcare facilities, both of which are suggested by our data in the form of high-risk patients and high-transmission facilities.

Previous studies have suggested that LTACHs play an important role in regional MDRO proliferation, acting as sites for uncontrolled transmission and sources of MDRO-colonized patients who move to other healthcare facilities [49, 50, 51]. Our data support these previ-

ous observations, with strains from LTACHs spreading via the patient transfer network, and CRKP transmission occurring within each LTACH. However, despite all LTACHs harboring critically ill patient populations, we found extensive variation in the amount of intra-facility transmission per importation across eleven Los Angeles area LTACHs. This result indicates that LTACHs may not all be equivalent with respect to their overall contribution to regional MDRO prevalence. An important consideration when interpreting these results is the unmeasured contribution of asymptomatically colonized patients. Given reports that the relative burden of asymptomatic carriage with CRE may be high [52], it will be important in future studies to validate these results using combinations of clinical and active surveillance cultures. However, we hypothesize that our study design and analysis strategy reduce the impact of not capturing asymptomatic carriers. In particular, previous reports indicate that frequency of asymptomatically colonized patients is markedly lower in high acuity patients [53], and our use of aggregate estimates of intra- and inter-facility transmission increases robustness to missing transmission intermediates, so long as they are not biased towards particular facilities.

Most infection prevention strategies that target already infected patients (e.g. cohorting of infected patients) assume that all carrier patients pose an equivalent risk of onward transmission to other patients. The identification of factors associated with increased risk of transmission may facilitate more targeted infection prevention strategies that allocate resources to the management of high-risk patients. Here, we identify several clinical factors that are preferentially associated with CRKP patients from high intra-facility transmission facilities. Three of the clinical characteristics associated with CRKP patients from high intra-facility transmission facilities - carbapenem use, malnutrition and older age, are associated with disruption of the gut microbiota [54, 55, 56], supporting overgrowth of colonizing CRKP in the gut as driving CRKP transmission [57]. If this finding proves generalizable, it would indicate that an effective anti-transmission therapeutic would only need to reduce

colonization density in carrier patients, as opposed to completely decolonizing them. While the relatively small number of patients in our study necessitates that future studies be undertaken to validate these specific clinical findings, the clear clinical interpretation of these patient factors provides support for the integration of genomic and clinical data as a strategy to identify drivers of uneven MDRO transmission across regional healthcare facilities.

There are several potential limitations in this study. First, we did not have access to admission screening rectal cultures for CRKP and therefore were not able to evaluate the contribution of asymptomatic gastrointestinal colonization to intra-facility transmission. While we believe future analysis should focus on the additional potential contribution of asymptomatically colonized patients to regional transmission in these LTACHs, our ability to reconstruct inter-facility transmission networks that are consistent with healthcare utilization patterns, using only clinical specimens, provides support for regional surveillance programs grounded in comprehensive clinical collections [58]. Second, we did not have access to isolates from other facilities in the region (e.g., acute care hospitals) that may play a significant role in CRKP transmission. However, restricting our analysis to LTACHs still allowed us to discern regional transmission patterns that were consistent with patient flow between facilities, and suggests that when combined with comprehensive patient transfer networks, WGS of isolates from a targeted set of high-burden facilities can allow for monitoring the regional dissemination of MDROs [59].

In conclusion, by integrating WGS with epidemiologic patient and facility data, we were able to reveal pathways and drivers of CRKP spread within and between regional LTACHs. The ability to tease apart the relative contributions of inter- and intra-facility transmission across regional healthcare facilities, and to identify drivers of these processes, is a critical first step in the development of effective regional interventions to control the spread of MDROs. We believe that this work exemplifies that the tools and resources now exist to usher in an era of precision regional infection prevention.

2.5 Methods

2.5.1 Study setting

A prospective, observational study was performed in LTACHs within the Kindred Healthcare network from August 1, 2014 to July 25, 2015. A total of 20 LTACHs from four regional networks were included: 11 LTACHs in the Los Angeles, CA area; 1 LTACH in San Diego, CA; 6 LTACHs in the Houston, TX area; and 2 LTACHs in the Tampa, FL area. These LTACHs were selected because all Kindred LTACHs in these regions are serviced by one central microbiology laboratory (Rancho, CA; Houston, TX; and Tampa, FL), thereby ensuring complete ascertainment of CRKP cultures. These regions were also selected based on variation in CRKP prevalence (e.g., LTACHs in the Los Angeles area were specifically included given their high prevalence of CRKP [59]. All study LTACHs routinely implemented contact precautions for patients with CRKP.

The study was reviewed and approved by the Institutional Review Board of the University of Pennsylvania with a waiver of informed consent.

Detection of KPC-positive *K. pneumoniae* and antibiotic susceptibility testing

All CRKP isolates during the study period were shipped to the core research laboratory of the Clinical Microbiology Laboratory at the Hospital of the University of Pennsylvania (HUP) for further testing. Confirmation that isolates were *K. pneumoniae* was performed using PCR for detection of the *mdh* gene, and through standard biochemical testing [60]. Isolates were subsequently tested for the presence of the blaKPC-1 gene using primers as previously described [6], and following the HUP Clinical Microbiology Laboratory protocol. Confirmation of carbapenemase activity was performed using the Rosco Diagnostica Rapid CARB kit (Rosco Diagnostica A/S, Tasstrup, Denmark).

Results of antibiotic susceptibility testing for all CRKP isolates were obtained from the

participating regional laboratories, including for colistin/polymyxin B. Because susceptibility testing for tigecycline was not routinely performed, testing was done at the HUP Clinical Microbiology Laboratory for all isolates using the Kirby-Bauer diffusion method and following routine laboratory protocol. Susceptibility to colistin or polymyxin B and tigecycline was defined as an MIC of ≤ 2 µg/mL, in accordance with European Committee on Antimicrobial Susceptibility Testing criteria [61].

Whole-genome sequencing and single nucleotide variant identification

WGS was performed on all study isolates at the PennCHOP Microbiome Center at the University of Pennsylvania. Genomic DNA was extracted from isolates using the DNeasy Blood & Tissue kit (Qiagen, Venlo, the Netherlands). Libraries were prepared from 1 ng extracted DNA using the NexteraXT DNA Sample Preparation Kit (Illumina, Inc., San Diego, CA). Library concentration was assessed using the Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies, Waltham, MA) and fragment size was estimated using the Fragment Analyzer (Advanced Analytical Technologies, Inc., Ankeny, IA). Libraries were pooled in equal molar quantities and sequenced on the Illumina HiSeq 2500 in Rapid Run mode to obtain 2x250 bp reads. Resulting fastq files were demultiplexed using DNAbc (https://github.com/PennChopMicrobiomeProgram/dnabc) with two allowed barcode mismatches. An average of 1.7 million read pairs were obtained per sample (Table 2.S4). Sequence data is available under Bioproject #PRJNA415194.

2.5.2 Single nucleotide variant identification

Quality of raw reads was assessed with FastQC [62], and Trimmomatic [63] was used for trimming adapter sequences and low quality bases. Variants were identified by: 1) mapping filtered reads to the finished KPNIH1 reference genome (Genbank accession no. CP008827) using the Burrows–Wheeler short-read aligner (BWA) [64], 2) discarding PCR duplicates

with Picard [65], and 3) calling variants with SAMtools and bcftools [66]. Variants were filtered from raw results using GATK's VariantFiltration (QUAL > 100, MQ > 50, > 10 reads supporting variant, FQ < 0.025) [67]. In addition, a custom Python script was used to filter out single nucleotide variants that were <5bp in proximity to indels. Lastly, for phylogenetic analyses, Gubbins was applied to filter out recombinant variants [68].

Phylogeographic analysis of CRKP spread across the United States

Study isolates were compared to publicly available previously sequenced CRKP isolates from various geographic locations to place regional CRKP populations in a broader context (Table 2.S3). The majority of analyzed CRKP isolates were from acute care hospitals in the United States (in outbreak and non-outbreak settings) and included only clinical cultures (e.g. no active surveillance cultures). Phylogeographic analyses were performed in BEASTv2.4.7 [69] using a GTR DNA substitution model, an uncorrelated lognormal (UCLN) relaxed molecular clock with a Gamma(0.001,1000) prior on the mean, and a Bayesian skyline population growth model with five groups. A strict molecular clock was rejected because zero is not in the 95% HPD of the coefficient of variation when using a UCLN clock model, and a constant population size was rejected because zero is not in the 95% HPD of the growth rate when using an exponential population model. Bayesian skyline was selected over the exponential population model using path sampling with 100 steps of 1 million iterations each. A Beta(alpha=0.3,1) distribution was used to determine the power posteriors of the steps. For each phylogeographic analysis, four independent Markov chains were run with 100 million iterations and a burn-in of 10 million. A maximum likelihood (ML) starting tree generated in RAxML [70] was used to accelerate MCMC chain convergence. Individual and combined chains were analyzed in Tracer v1.6.0 [71] to confirm convergence and adequate mixing. Maximum clade credibility (MCC) trees of the combined chains were generated using TreeAnnotator v2.4.7. These summary trees were used for all subsequent analyses.

2.5.3 Construction of regional transmission networks using maximum likelihood methods

A genomic transmission map was created among the eleven Los Angeles area LTACHs with the aim of evaluating inter- and intra-facility transmission links. First, a ML tree reconstruction was performed with IQTREE v1.5.5 [72] on variants identified by alignment to the KPNIH1 reference genome using ultrafast bootstrap with 1000 replicates (-bb 1000) [73]. ModelFinder [74] limited to ascertainment bias-corrected models (-m MFP-ASC) was used to identify the best model (TVM+ASC+R4) based on Bayesian Inference Criterion (BIC) across 10 replicates. Of the trees produced from the best-fitting model (n=2), the tree with the highest likelihood was used for all subsequent analyses. For subsequent regional transmission analyses, multiple isolates from the same patient while at the same facility were collapsed if they formed homogeneous phylogenetic sub-clades, and were retained otherwise as representing independent acquisition events.

Regional transmission networks were constructed by first calculating the relative likelihoods of unobserved ancestral outbreak strains being associated with each facility. Ancestral reconstruction was performed using the rerooting method for maximum likelihood ancestral reconstruction, as implemented in phytools [75]. Inter-facility transmissions were extracted from the ancestrally reconstructed phylogeny if a branch connected two nodes with different ancestral state assignments. Intra-facility transmissions were tabulated as branches connecting tips to an internal node with the same ancestral state. To account for importation into facilities, the number of transmissions into a given facility was subtracted from the total intra-facility transmission count.

Lastly, to quantify the strength of transmission linkages between facilities, while controlling for the total number of isolates associated with the pair of facilities, we devised a permutation test. In particular, the number of subtrees containing only isolates from a

given pair of facilities was calculated for the original phylogeny, and then compared to 1000 permuted phylogenies in order to compute an empiric p-value. To maintain the high level of intra-facility transmission in the actual data, permutation was done at the level of subtrees instead of isolates, wherein subtrees containing only isolates from a single facility were randomly swapped for one another.

2.5.4 Construction of regional transmission networks using BEAST

To validate the results from our maximum likelihood regional transmission network (Figure 2.S1), we performed a parallel analysis in BEAST (Figure 2.S2). The existence of temporal signal in the data was confirmed by using the ML tree to compare the root-to-tip regression estimate of the time to the most recent common ancestor (TMRCA) of the isolates using the true sequence isolation dates to that of randomly permuted isolation dates. A dated phylogeny of the Los Angeles area LTACH isolates was reconstructed in BEAST2 [69], with the same variants used in ML tree reconstruction, under a model averaging nucleotide substitution model (bModelTest [76]), an uncorrelated lognormal (UCLN) relaxed molecular clock [77] with a Gamma distribution prior (shape=0.1, scale=0.01) on the mean, a Bayesian skyline population model with five groups, and a discrete trait (facility). The number of invariant As, Cs, Gs, and Ts was included in the XML, and the ML tree was used as the starting tree to speed up convergence. Six 100,000,000 chain-length runs, confirmed to have reached convergence using Tracer v1.7.1 [71], were combined after removing 20% burn-in. A maximum clade credibility (MCC) tree was constructed using TreeAnnotator v2.4.7 and used for all subsequent analyses. The inter-LTACH transmission network was then extracted from the ancestrally reconstructed tree as described in the main Methods for the maximum likelihood reconstruction.

2.5.5 Analysis of LTACH CRKP rates

Evaluation of CRKP rates for the study LTACHs was performed using the following definitions:

1. Standardized CRKP rate for a specific facility was calculated as the total number of unique colonization and infection episodes in that facility divided by the number of facility beds (Table 2.S6). For this calculation, unique colonization and infection episodes were determined by collapsing multiple isolates from the same patient if they clustered on the phylogeny. Standardized CRKP rate for a specific facility includes both intra-facility and inter-facility transmission (defined below).

2. Intra-facility transmission per importation ratio for a specific facility was quantified as the number of isolates attributed to intra-facility transmission divided by the number of isolates attributed to importation (i.e., inter-facility transmission), as determined from post-hoc analysis of the ancestrally reconstructed phylogeny (See Methods above). Of note, because we only sequenced isolates from LTACHs, intra-facility transmissions could include transmissions that occurred at surrounding non-LTACH facilities, if members of transmission chains from surrounding facilities independently transferred to the LTACH.

Geographic maps including LTACH locations were created using LTACH addresses and latitude/longitude using R's network package, v.1.13.0 [78].

2.5.6 Construction and analysis of California statewide patient transfer networks

Claims data from the Centers for Medicare & Medicaid Services (CMS) were used to extract patient transfer networks connecting healthcare facilities in California from January 1, 2016

to December 31, 2016. CMS claims data captures admissions and discharges from healthcare facilities for all fee-for-service beneficiaries nationally. However, Medicare requires a qualifying stay in a hospital prior to admission to a skilled nursing facility (SNF) and will only cover the first 100 days of your SNF visit, thus not all SNF stays are included in the claims and some SNF discharges will not represent true discharges but a switch to private coverage. Claims data were used to capture direct (i.e., transfers with no more than one day between the discharge at facility 1 and admission at facility 2) and indirect transfers (i.e., readmissions with an intervening stay in the community less than a pre-specified time). Healthcare facilities included acute care hospitals, SNFs, and LTACHs, including the eleven Kindred Los Angeles area LTACHs evaluated in the present study.

This patient transfer network was considered a directed weighted graph such that nodes in the patient transfer network corresponded to healthcare facilities and edges corresponded to patient transfers from source to destination facility. Raw edge weights across the network (count of transfers from facility X to facility Y within 365 days) were normalized by dividing by the total number of outgoing patient transfers at the source facility (i.e., out strength). This normalization resulted in edge weights that represented the probability of a patient going to facility Y if he/she was discharged from facility X. The length of the directed edge from X to Y was defined as the logarithm of the inverse probability. The shortest path between each pair of LTACHs was determined using these edge lengths and Dijkstra's algorithm [79]. All network analyses and visualization were performed in R igraph v1.2.0 [80].

The maximum patient flow between each pair of LTACHs was compared to the inferred genomic transmission network and the driving distance between facilities using the Mantel test [81] in the R package vegan v2.5-2 [82]. The number of importations from source to destination facility was normalized by the total number of importations to the destination facility. For Figure 2.S3, patient transfer and genomic transmission matrices were made

undirected by adding the matrix to its transpose and dividing by two. Driving distance between facilities was determined using the gmapsdistance R package [83].

2.5.7 Analyses of patient characteristics

Comorbididty and medication data were acquired via automated queries from Kindred's electronic medical record system, ProTouch. Comorbidities were ascertained using diagnosis codes, and medications were ascertained using pharmacy administration data. Descriptive analyses were performed to summarize information of patients with CRKP, including demographics, comorbidities, and presence of indwelling devices. Bivariable analyses of CRKP and CSKP population characteristics between select facilities were performed using the Fisher's exact test for categorical variables and the Wilcoxon rank-sum test for continuous variables. For all calculations, a two-tailed P-value of ¡0.05 was considered significant. All analyses were performed using STATA v.14.0 (StataCorp, College Station, Texas).

2.6 Supplement

2.6.1 Supplementary tables and figures

For supplementary material visit https://aac.asm.org/content/63/11/e01622-19/figures-only.

Chapter 3

Patient and Microbial Genomic Factors Associated With Carbapenem-Resistant *Klebsiella pneumoniae* Extraintestinal Colonization and Infection

3.1 Preamble

This chapter explores the relationship between extraintestinal colonization and infection in clinical CRKP isolates using machine learning. We investigate the predictive power of clinical and genomic features in distinguishing between colonization and infection, and identify a sublineage of CRKP that appears to have evolved an increased affinity for colonization of the respiratory tract, followed by increased infectivity due to acquisition of a virulence factor. This analysis inspired us to create `mikropml` (https://github.com/SchlossLab/mikropml), an R package for performing machine learning analyses.

I performed all of the data analysis and generated all off the figures for this chapter. Other co-authors gathered the clinical data and defined isolates as colonization vs. infection. This work was published in mSystems in 2021:
Lapp Z, Han JH, Wiens J, Goldstein EJ, Lautenbach E, Snitkin ES. Patient and Microbial Genomic Factors Associated with Carbapenem-Resistant Klebsiella pneumoniae Extrain-

testinal Colonization and Infection. Msystems. 2021 Apr 27;6(2).

I developed the grouping aspect of `mikropml` used in this chapter, wrote the function to preprocess the data for machine learning, debugged several functions, and wrote checks and unit tests. As of April 28, 2021 `mikropml` has 2,044 downloads. The manuscript corresponding to `mikropml` was submitted for publication with the following co-authors:
Begüm Topçuoğlu*, Zena Lapp*, Kelly Sovacool*, Jenna Wiens, Evan Snitkin, and Patrick Schloss.
*Indicates co-first author

3.2 Introduction

Infections due to multidrug resistant organisms lead to hundreds of thousands of deaths worldwide each year [84]. Carbapenem-resistant Enterobacterales (CRE) are a critical-priority antibiotic resistance threat that has emerged over the past several decades, spread across the globe, and accumulated resistance to last-line antibiotic agents [5, 85]. In the United States (U.S.), CRE infections are primarily caused by the sequence type (ST) 258 strain of carbapenem resistant *Klebsiella pneumoniae* (CRKP), which has become endemic in regional healthcare networks [85, 32, 36, 86, 46]. In this background of regional endemicity the risk of patient exposure to CRKP is high, as evidenced by alarmingly high rates of colonization, especially in long-term care settings [46, 45]. However, even among critically ill patients residing in long-term care facilities, not all colonized patients develop clinical infections that require antibiotic treatment [46, 52]. Currently, our understanding of the factors that influence whether a colonized patient develops an infection is incomplete.

In addition to clinical characteristics of patients [87], the genetic background of the colonizing strain may also influence the risk of infection, as there is extensive intra-species

variation in antibiotic resistance and virulence determinants harbored by *K. pneumoniae* [85]. To date, most studies of virulence determinants have been carried out in model systems [25] or examined in human populations without considering patient characteristics or clinical context [26, 30]. One recent study investigated virulence determinants in *K. pneumoniae* clinical isolates while controlling for patient characteristics [88]. However, this was a single-site study with a focus on carbapenem-susceptible *K. pneumoniae*, thereby not addressing the impact of genomic variation in antibiotic-resistant lineages that circulate in global healthcare systems.

Here, we sought to understand the importance of both patient factors and genomic features in determining whether a patient is colonized or infected with CRKP ST258. Importantly, we restricted our comparison to patients with extraintestinal CRKP colonization versus infection. We reasoned that this comparison would reveal the patient and microbial factors that influence risk for infection when CRKP is present in an extraintestinal site; eliminating confounding by factors associated with translocation from the gastrointestinal tract. To gain an unbiased assessment of influential patient an microbial factors in a high-risk population we leveraged a comprehensive set of all clinical isolates and patient metadata collected from 21 long-term acute care hospitals (LTACHs) across the U.S. over the course of a year. Machine learning models rigorously trained and tested on these data revealed that patient and microbial factors were similarly predictive of CRKP ST258 colonization versus in infection, indicating that both contribute to infection risk. Moreover, examination of predictive genomic features revealed genetic variation within the epidemic ST258 lineage of CRKP that was associated with increased respiratory colonization and higher infection rates.

3.3 Results

Of 355 clinical CRKP isolates from 21 LTACHs across the U.S. (15), we classified 149 (42%) of the isolates as representing extraintestinal infection based on modified National Healthcare Safety Network (NHSN) criteria [89] (Figure 3.S2, Tables 3.S1-3). The rest of the isolates were classified as representing extraintestinal colonization. Stratified by anatomic site, we classified 29/29 (100%) blood isolates as infection, 69/196 (35%) respiratory isolates as infection, and 51/130 (39%) urinary isolates as infection (Table 3.S3). More than 90% of patient isolates were from the epidemic CRKP lineage ST258 (Table 3.S1). Patients harboring different sequence types of CRKP showed no significant differences in infection/colonization status or anatomic site of isolation, and no substantive differences in clinical characteristics (see supplementary material). Thus, we decided to limit our analysis to ST258 to improve our ability to discern whether genetic variation within this dominant strain is associated with infection.

3.3.1 The CRKP epidemic lineage ST258 shows evidence of sub-lineage variation in virulence and anatomic site of isolation

We next evaluated if there exist sub-lineages of ST258 with altered virulence properties by looking for clustering of isolates by infection on the whole-genome phylogeny (Figure 1; see supplementary methods) [90]. Infection status was non-randomly distributed on the phylogeny (p=0.002), supporting our hypothesis that the genetic background of CRKP influences infection. We performed a similar clustering analysis to look at potential niche-specific adaptation to certain anatomic sites (Figure 3.1), and found that respiratory (p=0.001) and urinary (p=0.013) isolates cluster on the phylogeny, but blood isolates do not (p=0.21). This analysis indicates that, in addition to patient features, intra-strain variation in virulence and adaptation to the urinary and respiratory tract might influence whether patients develop an

infection.

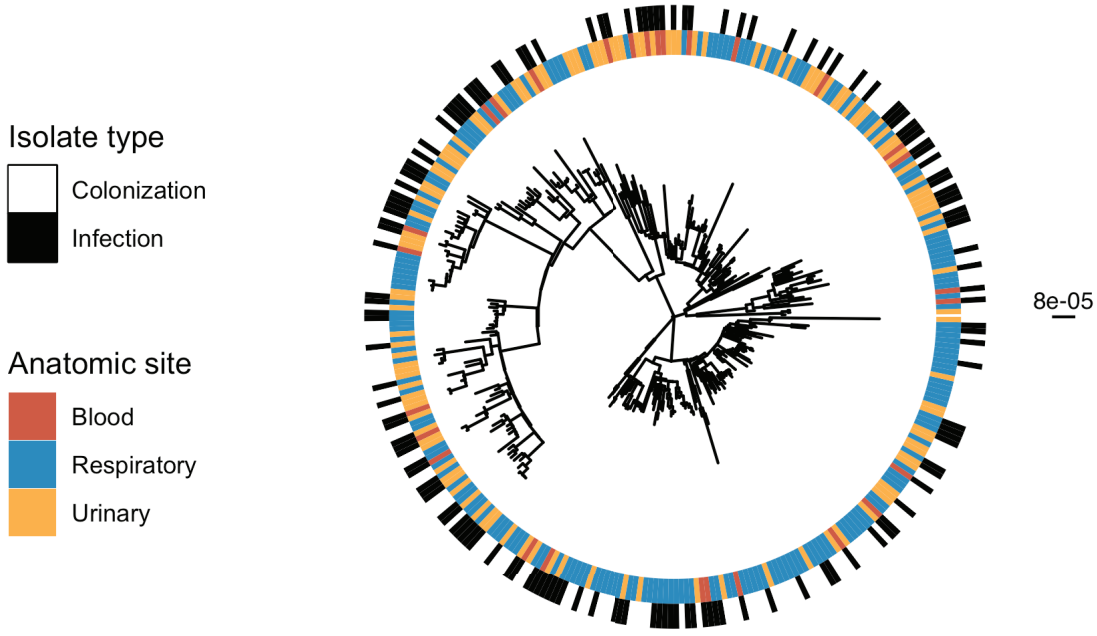

Figure 3.1: Infection and anatomic site cluster on the phylogeny.

Maximum likelihood phylogenetic tree of all isolates including infection or colonization classification for each isolate and anatomic site of isolation. The scale bar to the right of the phylogeny shows the branch length in substitutions per site. Testing for non-random distribution of isolates on the phylogeny (see supplementary methods) revealed clustering of infection, respiratory, and urinary isolates on the phylogeny, respectively.

3.3.2 Both patient and CRKP ST258 genetic characteristics are weakly predictive of infection, with relative performance being highly facility-dependent

We next performed machine learning using L2 regularized logistic regression to quantify the ability of patient and microbial genetic characteristics to predict CRKP ST258 infection (Figure 3.S1). To prevent over- or under-fitting and control for facility-level biases, we generated 100 train/test data splits, wherein a given LTACH was only included either in the

train or test set. Each LTACH occurred a median of 24 times (range 13-32) in the test data split. In this way, we were able to identify patient and CRKP ST258 strain characteristics consistently associated with infection or colonization across data splits, and thus across patient populations in different healthcare facilities.

First, we sought to understand if patient and genomic features were individually predictive of CRKP ST258 infection. To this end, we independently evaluated patient characteristics as well as three different genomic feature sets for their ability to classify colonization and infection. The three genomic feature sets were uncurated genomic (including single nucleotide variants, indels, insertion elements, and accessory genes), uncurated grouped genomic (variants grouped into genes, akin to a burden test, e.g [91]), and curated genomic (features identified using Kleborate [92]). Across the 100 different train/test splits, we observed that the average predictive performance was weak, with each of the genomic and patient feature sets predictive of infection to a similar degree (all 1st quartile area under the receiver operating characteristic curves [AUROCs] > 0.5; median range=0.55-0.68; Figure 3.2A; AUPRC: Figure 3.S3A). Additionally, no one feature set was consistently the most predictive (e.g. Figures 3.2B, 3.2C; all comparisons $p > 0.30$, see supplementary methods for p-value calculation). Furthermore, for each feature set the AUROCs were distributed such that the test AUROC ranged from below 0.5 to over 0.7, depending on how the data were split (i.e., which facilities appear in the train/test sets). This variation in model performance across different train/test sets suggests that the association of CRKP ST258 strain and patient characteristics with infection or colonization varies across facilities.

Integration of patient and CRKP strain features does not improve discriminative performance of overall or anatomic site-specific models

To determine if the predictive power of patient and genomic features is additive, and if combining these disparate feature sets improved validation on held-out facilities, we built

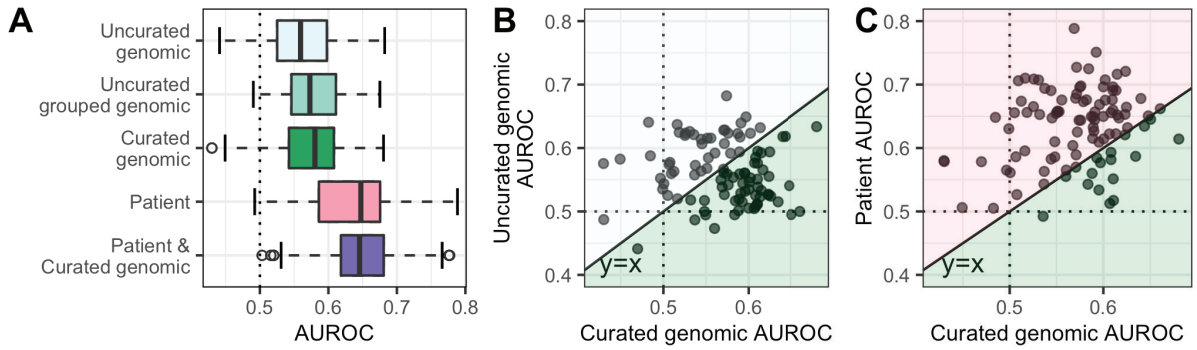

Figure 3.2: Test AUROCs for various classifiers identifying CRKP colonization vs. infection vary substantially across data splits.

(A) Test AUROCs for 100 L2 regularized logistic regression models different using train/test splits. All isolates from a given LTACH were included in either the training split or the testing split for each data split. We built models using five different feature sets, keeping the same 100 data splits. AUROCs of different feature sets were not significantly different. In the right two panels, the curated genomic feature set AUROCs are compared to: (B) the uncurated genomic feature set AUROCs, and (C) the patient feature set AUROCs. Each point is the resulting pair of AUROCs for models built with the same data split, but the two respective feature sets. The dotted lines in all 3 panels indicate the AUROC for choosing an outcome randomly (0.5); anything below the line is worse than random, and anything above the line is better than random. The solid diagonal line in the right two panels is the line y=x; points below the line correspond to a higher curated genomic AUROC for that data split, and points above the line correspond to a higher uncurated genomic AUROC (B), or patient AUROC (C), respectively. The colors in panels (B) and (C) correspond to the colors in panel (A); the points in a given colored area indicate that that feature set had the higher AUROC for that data split. In both cases, one feature set does not consistently outperform the other (p=0.4; see supplementary methods for p-value calculation). AUROC=area under the receiver operating characteristic curve.

models including both patient and curated genomic features. The discriminative performance of the models based on the combined feature set was not significantly greater than that of the individual feature sets (Figure 3.2A, all $p \geq 0.20$). Thus, despite variation in the predictive capacity of genomic and patient features across facilities (Figure 3.2C), combining the two sets did not improve overall performance. Furthermore, we found that there was no significant difference in model performance between L2 regularized logistic regression, elastic net, random forest, and support vector machines with a radial basis kernel (Figure 3.S3B, all $p > 0.1$). Focusing on anatomic site-specific L2 regularized logistic regression models revealed similar trends, where classification performances were similar for respiratory and urinary specific models, and the relative predictive capacity of patient and CRKP ST258 strain features varied across facility subsets (Figure 3.S3C, S3D).

3.3.3 Some patient and genomic features consistently discriminate colonization and infection

After evaluating the predictive capacity of models, we next sought to identify patient and CRKP ST258 strain characteristics that are most associated with infection or colonization. To this end, we identified those patient and genomic features that consistently improved model performance across the 100 different data splits (see methods). Evaluating the importance of features in this way provides insight into those characteristics that generalize across different facility subsets. This approach was taken for both overall and anatomic site-specific models to identify features predictive of different anatomic sites of infection (Figure 3.3, 3.S4).

Several patient features were consistently associated with infection in the overall analysis, including presence of a gastrostomy tube, presence of a central venous catheter, acute kidney injury, and severe chronic kidney disease (Figure 3.3), all markers of critically ill patients.

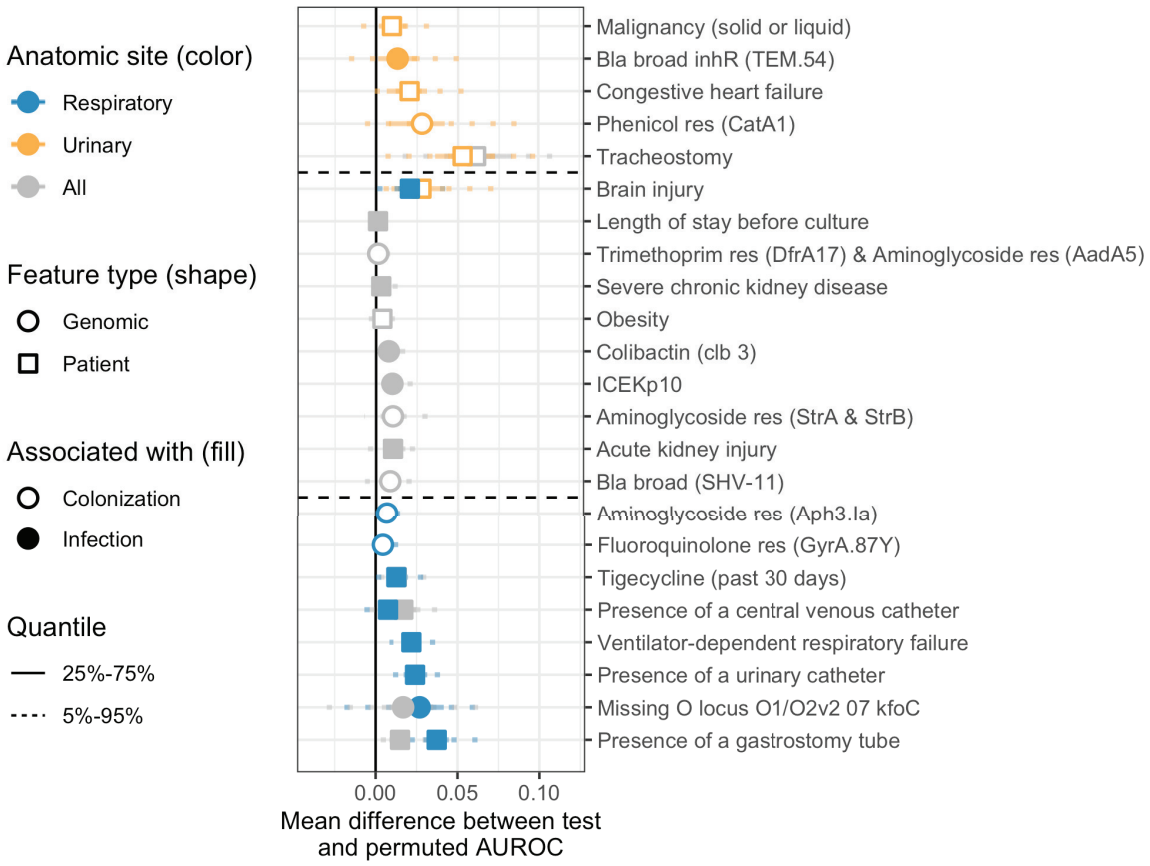

Figure 3.3: Features consistently associated with colonization or infection sometimes differ between the overall, respiratory, and urinary models.

Feature-specific improvement in model performance, measured as the mean difference between test and permuted AUROC (see methods), of features found to be consistently associated with colonization or infection in at least one of the following analyses: overall, respiratory-specific, urinary-specific. We consider features to be associated with infection/colonization if the AUROC difference was greater than zero in over 75% of the 100 data splits. Each point is surrounded by two quantiles - a solid line for 25% to 75% (the threshold used to select features) and a dotted line for 5% to 95%. The vertical solid black line indicates a difference of zero (i.e. the feature provides no improvement to model performance). Horizontal dotted lines separate features associated with urinary but not respiratory isolates (top), both urinary and respiratory (or all) isolates (middle), or respiratory but not urinary isolates (bottom). Bla=Beta lactamase, res=confers resistance to that antibiotic class. AUROC=area under the receiver operating characteristic curve.

Only a small number of genomic features were consistently associated with infection or colonization (Figure 3.3). The genomic features associated with colonization were all antibiotic resistance determinants. Conversely, all but one of the genomic features positively associated with infection (3/4) are related to virulence. The ICEKp10 element is positively associated with infection and carries colibactin and two different types of yersiniabactin, a previously identified *K. pneumoniae* virulence determinant [26]. Colibactin is a toxin [85], and yersiniabactin is an iron scavenging system that has been identified in previous animal and human studies as being associated with virulence [85, 25]. Additionally, insertion sequence-mediated disruption of the O-antigen biosynthetic gene *kfoC* was associated with respiratory infection (see methods and Figure 3.S5A for insertion sequence identification). The O-antigen of lipopolysaccharide (LPS) is a known antigenic marker, although association with a specific anatomic site has not been noted [93].

3.3.4 A sub-lineage of ST258 clade II appears to have sequentially evolved enhanced adaptation for the respiratory tract and increased virulence

We noted that *kfoC* disruption is largely confined to a sub-lineage of ST258 present across 12 LTACHs in California (Figures 3.4, 3.S5). Consistent with this feature being associated with respiratory infection, the disrupted *kfoC* lineage is enriched in respiratory isolates (82/118, 69% of isolates in the disrupted *kfoC* lineage are respiratory isolates vs. 101/213, 47% in all other isolates; Fisher's exact p=0.0001), suggesting that this lineage is associated with increased capacity for respiratory colonization. Furthermore, a subset of isolates in the disrupted *kfoC* sub-lineage harbor the ICEKp10 element containing yersiniabactin. Examination of these genetic events in the context of the whole-genome phylogeny revealed that disruption of *kfoC* occurred first, followed by at least two different acquisitions of ICEKp10

(Figure 3.4). Within the disrupted *kfoC* lineage, isolates with ICEKp10 are enriched in infection (31/55, 56% of isolates with ICEKp10 are infection isolates vs. 16/63, 25% of isolates without ICEKp10, Fisher's exact p = 0.00065), supporting an increase in virulence after acquisition of ICEKp10. It is important to note that the observed clinical associations with ICEKp10 and *kfoC* disruption do not demonstrate causality, as we cannot rule out the role of correlated genetic variation.

3.4 Discussion

There have been numerous studies aimed at identifying risk factors for healthcare-associated infections caused by prominent antibiotic-resistance threats. For the most part, these studies have found the dominant risk factors to be linked to the magnitude of exposure (e.g. length of stay or colonization pressure), use of antibiotics, and overall comorbidity [94]. What remains unclear is if in addition to these clinical features, the genetic variation in circulating resistant lineages also contributes to patient risk of infection. Here, we addressed this question for CRKP ST258 in a comprehensively sampled cohort of patients from 21 LTACHs across the U.S. Overall, we found that, while neither patient nor CRKP ST258 genetic features have high predictive accuracy on held-out test data, both feature sets were independently associated with infection, with one or the other being more predictive on different facility subsets. Moreover, the integration of clinical and genomic data led to the discovery of an emergent sub-lineage of the epidemic ST258 clone that may have increased adaptation for the respiratory tract, and is more strongly associated with infection.

One strength of our machine learning approach is that we were able to measure the variation in discriminative performance across 100 train/test iterations that differed in which facilities were included in train and test sets. We found that performance varied greatly depending on how facilities were allocated to train and test sets, highlighting how smaller

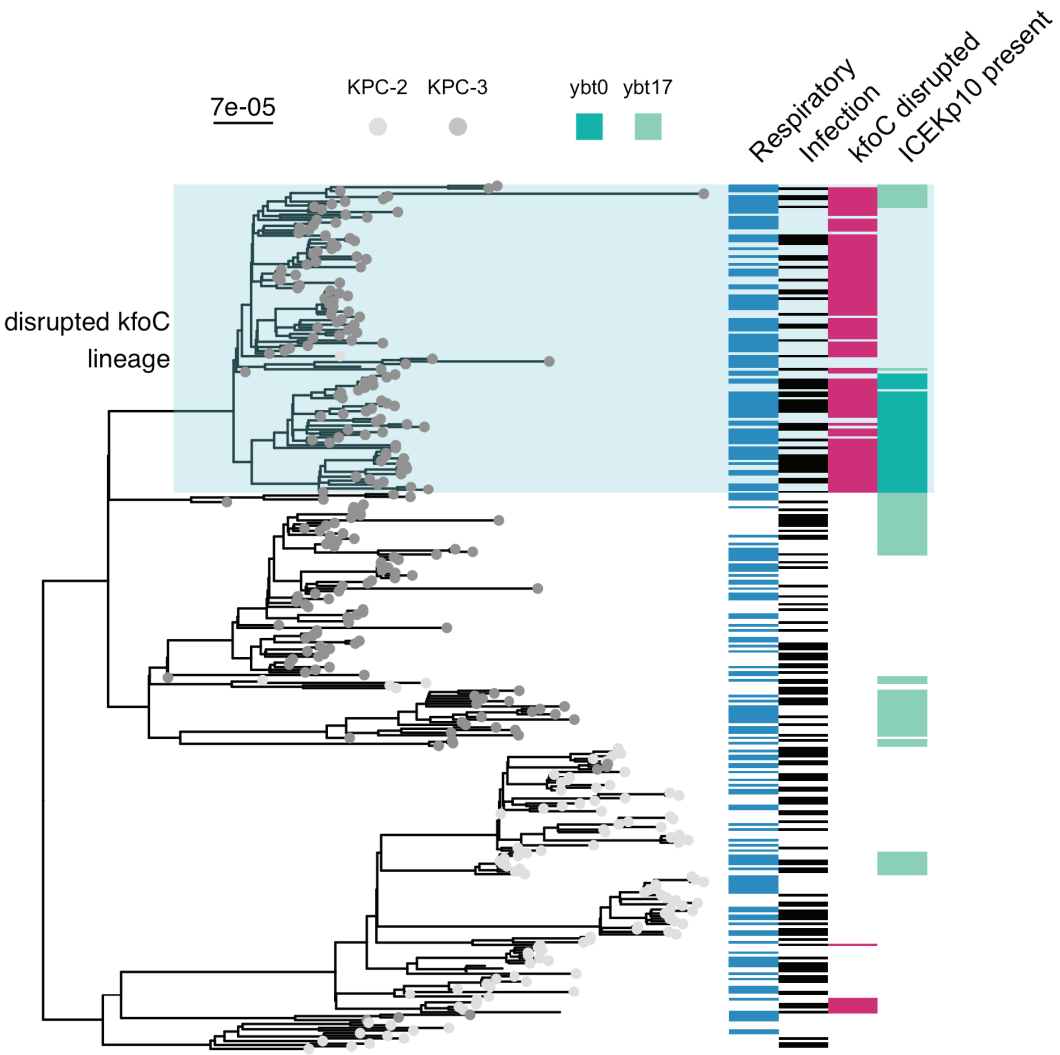

Figure 3.4: Select epidemiologic and genomic features visualized on the phylogeny indicate that a sub-clade of ST258 clade II may exhibit enhanced niche-specific adaptation and virulence.

ST258 maximum likelihood phylogeny with the tip labels colored by KPC gene. The blue box indicates the sub-lineage with apparent altered niche-specific adaptation that acquires an additional virulence locus. The heatmap beside the tree indicates information about the isolate. From left to right: if it is a respiratory isolate, if it is an infection isolate, if *kfoC* is disrupted, and if it contains ICEKp10. Disrupted *kfoC* was associated with infection in the overall and respiratory machine learning analyses and ICEKp10 presence was associated with infection in the overall analysis. Note that the majority of isolates without disrupted *kfoC* in that lineage contain a different O locus type (Figure S5A). The scale bar to the top left of the phylogeny shows the branch length in substitutions per site. ybt=Yersiniabactin; ybt0 and ybt17 are two ybtSTs defined by Kleborate.

studies could overestimate or underestimate the discriminative ability of both their model and individual features. Variation in model performance across facilities could be due to facility-level heterogeneity leading to differences in the prevalence of predictive patient or genomic features in the different test sets. For instance, certain facilities may have patient populations skewed towards individuals with characteristics that are predictive of infection. Alternatively, certain geographic regions may have CRKP ST258 strains that are more virulent than in other regions. These differences could lead to a higher predictive power for certain facilities compared to others. Another possible explanation for variation in model performance is that the critically-ill nature of LTACH patients may be such that most patients are actually highly susceptible to infection (i.e. many patients colonized with CRKP may ultimately develop an infection). However, it is noteworthy that despite these potential challenges in creating generalizable models, our analysis did yield predictors of infection and colonization consistent across test sets, and thus across LTACHs.

We built classifiers including all genomic features as well as a curated subset of features from Kleborate [92], and found that both are similarly predictive of infection. However, while the uncurated feature set presented challenges with downstream interpretation, our analyses on the curated genomic features [92] facilitated novel insights into potential evolutionary trajectories of anatomic site-specific adaptation and virulence. For example, we observed that disruption of the O-antigen biosynthetic gene, *kfoC*, is associated with isolation from the respiratory tract. While we cannot determine from our machine learning analysis if disruption of *kfoC* is directly causal, the biological plausibility of an altered O-antigen structure mediating evasion of innate immunity and/or other beneficial interactions with the host makes this a strong candidate for follow-up experiments. Supporting this hypothesis, a previous study found that absence of O-antigen is associated with decreased virulence, but not decreased intrapulmonary proliferation, in a murine model [95]. In addition, we noted that a number of antibiotic resistance determinants were associated with coloniza-

tion. We hypothesize that this observation could be a consequence of longer duration of residence being associated with increased exposure to off-target antibiotics [96]. Finally, we also saw evidence that, after acquiring the virulence factors yersiniabactin and colibactin on the ICEKp10 element, the disrupted *kfoC* subclade became more strongly associated with infection, supporting the idea that circulating ST258 sub-lineages can evolve to become both hypervirulent and multi-drug resistant [97, 23, 98, 99].

It is important to note that the machine learning method we employed does not correct for microbial population structure. We chose this method instead of alternative bacterial genome-wide association methods because our primary interest was in quantifying the overall predictive capacity of bacterial genotype in a patient population that was collected in a comprehensive and unbiased manner (i.e. all clinical isolates from 21 facilities over one year). While alternative methods controlling for population structure may yield more precise estimates of the contribution of individual variants, this would obfuscate the realized contribution in our patient population and hinder direct comparison to the predictive capacity of patient features. However, this then presented the challenge of interpreting our finding that certain sub-populations of CRKP ST258 may differ in their predilection for causing infections at different sites. For instance, the *kfoC* disruption is a lineage defining variant, and in principle other variants that define this lineage could also be causal. Here, we limited our analysis to a curated set of variants belonging to pathways known to be associated with antibiotic resistance and virulence, and found that only *kfoC* disruption was associated with increased respiratory infection, thus making it a strong candidate for follow-up in *in vitro* or *in vivo* models [25]. To identify novel loci whose role in human infection may not be appreciated both computational and experimental strategies may be employed to help prioritize putative causal versus passenger variants. Computationally, investigators may search for evidence of parallelism in genotype/phenotype associations, which would bolster confidence in causality [100]. Alternatively, high-throughput screens of genetic mutants in relevant model

systems can help prioritize candidates. Garnering further genomic or experimental support for the direct role of a specific genetic variant would in turn increase the likelihood that those genetic markers would be predictive in new strains and patient populations.

Our study also has several important limitations related to the data available. Specifically, extraintestinal CRKP colonization vs. infection for non-bloodstream isolates may be difficult to discriminate using surveillance criteria and the clinical data that were available. However, we based our definitions on established Centers for Disease Control and Prevention criteria with modifications used previously [46]. Encouragingly, we were still able to identify consistent predictors of infection, even with potential misclassifications. A second limitation is that our dataset only included one clinical culture for the majority of patients, meaning that we were unable to investigate clinical or genomic features that may be associated with progression from colonization to infection. Furthermore, we do not have CRKP rectal colonization isolates, and therefore cannot evaluate transition from rectal colonization to other body sites. However, we hypothesize that comparing rectal colonization to infection may be asking a subtly distinct question - namely bacterial genetic factors that enable translocation from the gut to other body sites. In contrast, we hypothesize that our study design is ideal to identify bacterial genetic factors associated with infection once at a given body site. Additionally, we chose to focus our analysis on ST258 due to its disproportionate presence in our dataset, but this makes it possible that our findings may not generalize to other sequence types. Nevertheless, ST258 is the dominant clone in the U.S., and the methods we employed here can be used to study other sequence types and other pathogens. Furthermore, our focus on the ST258 lineage led to the particularly notable finding that even within an established endemic multidrug resistant lineage (that emerged circa 2000 [22]), there is continued evolution that influences the manifestation and outcome of infection. This highlights the importance of performing strain-specific analyses to identify continued evolution and adaptation of hospital-associated lineages. We were also limited in

the patient data included in our model. It is likely that important differences in underlying patient conditions were not captured by the coarse clinical variables we included, and we also did not account for differences in genetic variation in the host [101]. Other limitations include that our study was restricted to LTACH patients and had non-random geographic sampling. However, our restriction to LTACHs in endemic geographic regions has the benefit of focusing on populations at disproportionate risk for CRKP infection [45].

In conclusion, we employed a machine learning approach to quantify our ability to discriminate between CRKP colonization and infection using patient and microbial genomic features. This approach highlighted the high degree of variation in predictive accuracy across different facility subsets. Furthermore, despite modest predictive power, we identified several genomic features consistently associated with infection, indicating that variation in circulating CRKP strains contributes to infection, even in the context of the critically-ill patient populations residing in LTACHs. Future work should aim to corroborate our findings with larger cohorts and follow up on strong associations to determine whether they are indeed risk factors for infection. This could ultimately help identify patients at high risk for CRKP ST258 infection and devise targeted strategies for infection prevention. Furthermore, the methods employed here can be used to study ongoing adaptation in other important MDRO lineages circulating in healthcare facilities.

3.5 Methods

3.5.1 Clinical and genomic data

We used whole-genome sequences of clinical (non-surveillance) CRKP isolates and associated patient metadata from a prospective observational study performed in 21 LTACHs from across the U.S. over the course of a year (BioProject accession no. PRJNA415194) [22]. All isolates were ordered by clinicians as part of clinical care, and clinical practice guidelines

and policies are standard across sites within the network. We included only the first clinical bloodstream, respiratory, or urinary isolate from each patient (n=355; Figure 3.S1A), and subset to only ST258 isolates for the majority of analyses (n=331; Table 3.S1; see supplementary material for reasoning). Patient metadata were obtained from electronic health records. Core genome variants were identified using a reference genome and accessory genes were identified using Roary [102]. Details about the clinical data, analysis pipeline [103], genomic data curation [22, 92, 97, 102, 104, 105, 106, 107], and phylogenetic reconstruction [66, 68, 72, 73] are provided in the supplementary material. While most clinical data cannot be shared, the deidentified patient ID, hospital of sample isolation, and isolation site are included in the Sequence Read Archive metadata for the BioProject.

3.5.2 Outcome definition

Our outcome of interest was colonization vs. clinical infection (Figure 3.S1B). Based on the U.S. Centers for Disease Control and Prevention's (CDC) established NHSN surveillance definitions, we considered all bloodstream isolates as representative of infection, and used modified definitions as in [46] to classify urinary and respiratory cultures as representative of infection versus colonization (Table 3.S2) [46, 89]. Any isolate that did not meet the criteria for infection was classified as colonization. We did not incorporate physician interpretation in applying the criteria to ensure consistency in applying the definition.

3.5.3 Feature sets

We studied the association between five different feature sets and infection/colonization in CRKP ST258 (Figure 3.S1C); the feature sets are described below. See supplementary methods for details on feature set creation and processing. Counts below are for confident features from the entire dataset prior to subsetting for different analyses.

1. Patient: Clinical features described in Han et al. (n=50; Table 3.S3) [22].

2. Uncurated genomic: Single nucleotide variants, indels, insertion sequence elements, and accessory genes (n=2447).

3. Uncurated grouped genomic: Variants grouped into genes (i.e. a burden test, e.g. [91]) and accessory genes (n=3159).

4. Curated genomic: Features identified by Kleborate [92], a tool designed to identify the presence of various genes and mutations known to be associated with either CRKP virulence or antibiotic resistance (n=91).

5. Patient & curated genomic: Patient features and curated genomic features (n=141).

3.5.4 Machine learning & model selection

We aimed to classify clinical infection (vs. colonization) using each of the different feature sets (see above); we built classifiers using the first clinical isolate from each patient for all isolates, only respiratory isolates, and only urinary isolates. We performed L2 regularized logistic regression on all feature sets using a modified version of the machine learning pipeline presented in Topçuoğlu et al. [108] using caret version 6.0-85 [109] in R version 3.6.2 [110] (Figure 3.S1D1). Furthermore, for the patient and curated genomic feature set we performed elastic net, random forest, and support vector machine with a radial basis kernel using the same method but implemented in `mikropml` version 0.0.2 [111]. We randomly split the data into 100 unique 80/20 train/test splits, keeping all isolates from each LTACH grouped in either the training set or the held-out test set to control for facility-level differences among the isolates (e.g., background of circulating strains within each facility, patient population, and clinician test ordering frequency). For valid comparison, the train/test splits were identical across models generated with different feature sets. Hyperparameters were selected via cross-

validation on the training set to maximize the average AUROC across cross-validation folds. See supplementary methods for more details.

3.5.5 Model performance

We measured model performance using the median test area under the receiver operating characteristic curve (AUROC) and area under the precision recall curve (AUPRC), as well as the interquartile range, across all 100 train/test splits (Figure 3.S1D2).

3.5.6 Features consistently associated with colonization or infection

To determine the importance of each feature in predicting colonization vs. infection, we measured how much each feature influenced model performance by calculating feature importance using a permutation test [108] (Figure 3.S1D3). For each combination of feature and data split, we randomly permuted the feature and calculated the 'permuted test AUROC' using the model generated with the training data. Features with a correlation of 1 were permuted together. Other correlation thresholds were tested and the results were very similar, so we chose to use a correlation of 1 to improve ease of interpretation. We performed this permutation test 100 times for each feature/data split pair, and obtained a mean feature importance for each data split. A mean feature importance above zero indicates that that feature improved model performance for that data split. We highlight features where the mean permuted test AUROC was above zero in at least 75% of the data splits. In this way, the permutation importance method allows us to take into account the variation we observe across the 100 models, which is not possible with standard parametric statistical tests or odds ratios.

3.5.7 Insertion sequence identification

We identified insertion sequences in the *kfoC* gene by running panISa on reads aligned to a reference genome [107, 66, 112, 113, 64]. See supplementary methods for more details.

3.5.8 Data analysis & visualization

See supplementary material for details on data analysis and visualization in R version 3.6.2 [110, 114, 115, 116, 117, 118]. All code and data that is not protected health information is on GitHub (https://github.com/Snitkin-Lab-Umich/ml-crkp-infection-manuscript).

3.6 Supplement

3.6.1 Supplementary methods

Clinical data

The majority of data elements were obtained from the electronic health record (EHR) which captures all data in the medical record across clinical sites. This EHR-based electronic database contains all demographic data, laboratory data (including microbiology), medication data, and radiographic data. ICD diagnostic and procedure codes were used to ascertain all coded primary and secondary diagnoses and procedures. The same EHR is used at all sites and is fully integrated into the EHR-based relational database. We chose not to include inflammatory markers (e.g. white blood cell count, procalcitonin, etc.) as they are highly non-specific and add little to designate a culture as representing infection vs. colonization, particularly in patients who typically have numerous comorbidities contributing to their acute illness.

Pipeline

We created a snakemake pipeline [103] to perform data pre-processing, machine learning analysis, and figure generation.

Genomic data

We used Kleborate version 0.3.0 [92] to identify sequence type, capsular types (K locus [104]) and O locus [105]), virulence factor [97], and antibiotic resistance genes. In addition, we identified single nucleotide variants (SNVs) and indels by mapping reads to the KPNIH1 reference genome (BioProject accession number PRJNA73191 [119]; https://github.com/Snitkin-Lab-Umich/variant_calling_pipeline). After variant calling, we identified the predicted functional impact of variants using SnpEff version 4.3T [106]. We also identified large insertion sequence (IS) elements using panISa version 0.1.4 with the default settings [107], and accessory genome genes using Roary version 3.12.0 with the default settings [102].

Phylogenetic tree reconstruction

We constructed a phylogenetic tree of all ST258 isolates by first creating a whole-genome alignment by mapping reads to the KPNIH1 reference genome (GenBank accession number CP008827.1) using bwa mem version 0.7.12 [64]; samtools version 1.9 [66] was used to generate binary alignment files. We then masked sites identified as recombinant by Gubbins version 2.3.2 [68] and used this masked whole-genome alignment to build a maximum likelihood phylogeny with IQ-TREE version 1.6.12 [72] using a GTR model of nucleotide substitution and ultra-fast bootstrap with 1000 replicates (-b 1000) [73]. The tree was midpoint-rooted using the midpoint.root function in phytools version 0.6-99 [75], thus splitting the tree into ST258 clades I and II [20].

Phylogenetic clustering test

To determine whether infection isolates cluster on the phylogeny more than expected by random chance, we performed a permutation test [90]. To identify the true number of isolates in a pure cluster while controlling for isolate clustering by LTACH, we enumerated isolates in pure infection or colonization subtrees that also contain isolates from more than one LTACH. Then we randomized the infection labels across isolates 1000 times and enumerated the number of isolates in a pure subtree for each randomization. We then compared the random counts of isolates in a pure cluster to the true number and calculated an empirical p-value.

Genomic feature set creation

To determine how well machine learning models perform using a targeted vs. untargeted approach, we performed machine learning with uncurated genomic features as well as curated genomic features known to be related to virulence. In an attempt to increase power, we not only performed machine learning on individual uncurated genomic features, but also using a burden test where genomic variants were grouped into genes and features were presence/absence of a variant in a gene. This may help capture variation associated with infection at the gene level that might not be captured at the variant level (e.g. different samples can have different variants in the same gene that all lead to a similar function). For the uncurated grouped genomic analysis, grouped variants included SNVs and indels identified as moderate or high impact by SnpEff and IS elements in a gene or upstream of a gene. We also grouped modifier variants identified by SnpEff into intergenic regions. In addition to core genes, we included accessory genes identified by Roary [102] as binary categorical features in both of the uncurated feature sets.

Feature set pre-processing

We used five feature sets to study the association between infection and colonization. We preprocessed all five datasets by mapping categorical features to binary variables, centering and scaling the continuous features (age and length of stay) to a mean of zero and a variance of one, and removing features present in only one sample or all but one sample. Some antibiotic features were antibiotic classes, and thus aggregates of other features (e.g. aminoglycoside, see Table 3.S3). For the uncurated genomic and uncurated grouped genomic datasets we collapsed genomic features with an identical pattern across all isolates to reduce machine learning runtime.

Machine learning rationale

We chose to perform L2 regularized logistic regression because it is easily interpretable, often performs just as well as more complicated methods when limited training data are available (18), and has a grouping effect, which means that all associated features are identified even if they are collinear [120]. L2 regularization is equivalent to including a zero-mean Gaussian prior on the weights. Thus, L2 regularized logistic regression includes implicit feature selection of sparse data by reducing the majority of feature weights to zero. Furthermore, this method allows us to maintain correlated features in the model, which aids in our interpretation of results as there may be features that together are important for understanding differences in infection and virulence. Finally, we found that L2 regularized logistic regression performs as well as other machine learning models for our dataset (Figure 3.S3), and that hyperparameter tuning for elastic net leads to a more L2-like model rather than an L1-like model.

Machine learning details

The machine learning pipeline sets at least 20% of the samples aside for testing in each train/test split. We split the training data into train and validate sets by LTACH using the groupKMultiFolds function from the caret R package (caret version 6.0-85) [109]. Within each training set, we selected hyperparameters via five-fold cross-validation 100 times, maximizing the average cross-validation AUROC.

Comparing model performance between feature sets and methods

To determine whether there was a significant difference in model performance between different feature sets and different machine learning methods, we calculated a two-sided empirical p-value for each identical train/test split using the formula 2 x min(fraction of AUROC differences ≥ 0, fraction of AUROC differences ≤ 0) [121].

Identification of IS element insertions in certain O2v2 *kfoC* genes

One of the curated genomic features we identified as associated with infection was what Kleborate called a missing gene (*kfoC*) in the O locus operon of certain O2v2 serotypes. On the Kleborate website, they indicate that a missing gene could represent a truly missing gene, or a gene that is split between two different contigs. To further investigate this, we used blastn version 2.9.0 [112] to search for *kfoC* genes in each of the samples. Additionally, we aligned the reads of all genomes to CP031810, a complete *K. pneumoniae* genome from PATRIC [113] using bwa mem version 0.7.12 [64], and samtools [66] was used to generate binary alignment files. Next, we used panISa [107] to identify IS element insertions in *kfoC*.

Data analysis & visualization

We used Fisher's exact tests for bivariable analyses with categorical variables, and Wilcoxon rank-sum tests for bivariable analyses with continuous variables. We performed all data

analysis and visualization in R version 3.6.2 [110] using the following packages: tidyverse version 1.3.0 [114], cowplot version 1.0.0 [115], ggtree version 2.0.1 [116, 117], ape version 5.3 [118], and phytools version 0.6-99 [75].

3.6.2 Supplementary results

Patients with different sequence types show no substantive differences in infection status, anatomic site of isolation, or clinical characteristics

Over 90% of the isolates in our dataset were ST258 (Table 3.S2), the dominant strain in the U.S. [85]. At the level of sequence type, we found no difference in infection prevalence (p=0.44) or anatomic site of isolation (p=0.66), indicating that at this coarse level there was not evidence of differences in strain virulence or adaptation to a certain anatomic site. Bivariable comparison of patient factors between patients with different sequence types revealed only four significant differences (unadjusted p-values < 0.05; previous use of piperacillin/tazobactam, cefepime, ciprofloxacin, or fluoroquinolones). As we found no substantive differences in infection, anatomic site of isolation, or patient variables when comparing different sequence types, we chose to focus all subsequent analyses on ST258, the dominant sequence type in our dataset. The genetic variation of isolates within ST258 is much smaller than the genetic variation between sequence types, so limiting our analyses to ST258 could improve our ability to identify associations of interest within ST258.

3.6.3 Supplementary tables and figures

For supplementary tables and figures visit https://msystems.asm.org/content/6/2/e00177-21/figures-only.

Chapter 4

Genomic and Clinical Insights Into the Emergence and Regional Spread of Colistin-Resistant *Klebsiella pneumoniae*

4.1 Preamble

This chapter investigates the origins of resistance to the antibiotic colistin in clinical CRKP isolates. We find that both *de novo* evolution and onward transmission of resistance are occurring, and that onward transmission of resistance is particularly rampant in the sublineage identified in Chapter 3. This sublineage is more fit than the other sublineages, and the fitness cost of colistin resistance in these strains appears to be diminished. The data preprocessing of variants for this analysis inspired us to create `prewas`, an R package for preprocessing data prior to downstream genomic analyses such as genome-wide association studies (https://github.com/Snitkin-Lab-Umich/prewas).

I performed all of the data analysis and generated all off the figures for this chapter. Other co-authors performed antibiotic susceptibility testing, variant calling, and phylogenetic tree reconstruction. This work will be submitted for publication with the following co-authors: Zena Lapp, Jennifer Han, Divya Choudhary, Stuart Castaneda, Ali Pirani, Ebbing Lautenbach, and Evan Snitkin.

I co-developed the methods for `prewas` with the other co-first authors. I implemented methods to find the ancestral allele, generate a binary matrix, and group variants into genes. As of April 28, 2021 `prewas` has 3,613 downloads.The manuscript corresponding to `prewas` was published in Microbial Genomics in 2020:

Saund K*, Lapp Z*, Thiede SN*, Pirani A, Snitkin ES. Prewas: Data Pre-Processing for More Informative Bacterial GWAS. Microbial genomics. 2020 May;6(5).

*Indicates co-first author

4.2 Introduction

Multidrug resistant organisms (MDROs) pose a significant threat to public health due to uncontrolled transmission and dwindling treatment options [1]. Of greatest concern are epidemic MDRO lineages that have become adapted to healthcare settings and continually gain resistance to additional antibiotics that are used to treat them [122, 123]. As resistance evolution continues to outpace our ability to develop new therapies [123], there is a critical need to improve our understanding of the forces driving the emergence and spread of resistance so that we can maintain the efficacy of effective antibiotics.

When *de novo* evolution of antibiotic resistance occurs in an individual, there is often little to no onward transmission to others due to the fitness cost of resistance in the absence of antibiotic pressure [124, 17]. However, in some cases this barrier is overcome, as evidenced by the proliferation of epidemic resistance clones [17, 36]. This may happen when the resistance variant or gene has an inherently low fitness cost and is thus maintained in the population even in the absence of antibiotic exposure [125]. Alternatively, the fitness cost of a given resistance element may depend on the genetic background in which it emerges, leading to dissemination of these elements only in certain lineages. Indeed, experimental evidence has shown that the genetic background of a strain interacts with resistance determinants to

influence the fitness of the resistant strain [126, 127], highlighting the importance of historical evolutionary events in determining the potential ability of a strain that becomes resistant to spread. While the preferential spread of resistance elements in specific strains lends support to the importance of genetic background, in most instances it is unclear what the underlying epistatic interactions are that lead to these different fitness effects [17]. Deciphering how the genetic background of a clinical strain influences transmissibility can provide insight into how fit strain/resistance combinations evolve and disseminate in the healthcare setting.

Carbapenem-resistant *Klebsiella pneumoniae* (CRKP) is an antibiotic resistance threat of critical priority as it is resistant to the majority of antibiotics on the market and has high mortality rates [1]. One of the few remaining treatment options for CRKP is the antibiotic colistin; however, colistin resistance emerges frequently during treatment with this drug [128, 129]. Despite the accessibility of these resistance variants to CRKP, their spread to other patients is rare, with most cases even within a single healthcare setting stemming from parallel evolution [129]. Here, we document the emergence and spread of multiple resistance variants in CRKP sequence type (ST) 258 across a regional healthcare network. Through integration of genomic and clinical data, we show how resistance spread was enhanced in genetic backgrounds that mitigate the fitness cost of resistance which, when combined with sustained selective pressure, enabled regional dissemination of colistin resistant lineages.

4.3 Results

4.3.1 Most resistant isolates contain variants in known resistance genes

Antimicrobial susceptibility testing revealed that 118/337 (35%) CRKP ST258 isolates were resistant to colistin, and that this resistance was present over the course of our yearlong

study across diverse long-term acute care hospitals (LTACHs; Figure 4.S1). To identify likely resistance variants, we first searched all isolates for *mcr*-containing mobile genetic elements and variants in canonical chromosomal genes known to confer resistance (*pmrA*, *pmrB*, *phoP*, *phoQ*, *crrA*, *crrB*, and *mgrB*). As expected due to the timeframe and locations in which the isolates were collected (July 2014 to August 2015 in the U.S.) [130], we did not find *mcr* genes in any of the isolates. However, the majority of colistin resistance we observed (103/118, 87%) could be explained by resistance variants in canonical resistance genes (Figure 4.S2; Table 4.S1; see methods for details on identification of resistance variants).

As not all resistance in our dataset could be explained by variants in known resistance genes, we next performed a genome-wide association study (GWAS) on the isolates with unknown resistance determinants to identify putative non-canonical resistance-conferring variants. We identified two additional putative resistance genes in this way: *qseC* and phosphotransferase system sugar transporter subunit IIB (Figure 4.S2; Table 4.S1). Resistance variants in these genes account for 6/15 (40%) of the unknown colistin resistance in our dataset (see methods for resistance variant identification). Notably, *qseC*, which explained resistance in 4 isolates, has been shown to confer colistin resistance in an experimental evolution study using clinical isolates [29]. For all subsequent analyses, we defined resistance genes as the set of canonical and GWAS-identified resistance genes.

4.3.2 Epistatic interactions appear to influence resistance in isolates with more than one variant in resistance genes

Next, we investigated the relationship between the number of variants in resistance genes and the level of colistin resistance. While the majority of isolates with one variant in a resistance gene were resistant, surprisingly, the majority of isolates with two or more variants in resistance genes were susceptible (Figure 4.1A). To better understand how the presence

of multiple variants in resistance genes influenced the extent of resistance, we explored the relationship between the number of variants in resistance genes and minimum inhibitory concentration (MIC). We found that resistant isolates with two variants in resistance genes tend to have higher MICs on average than those with one (Figure 4.1B). Moreover, susceptible isolates with one variant in a resistance gene have a higher MIC on average than those with none, suggesting that these variants increased the MIC to a level below the standard resistance cutoff. In contrast to these cases where resistance variants are associated with increased MIC, susceptible isolates with two or more variants in resistance genes are more susceptible on average than those with one or fewer. Furthermore, in most cases where susceptible isolates harbor multiple variants in resistance genes, at least one of the variants has been previously shown to confer resistance. These findings suggest that, while variants in resistance genes often increase MIC as expected, there also exist epistatic interactions among variants in these genes that influence their impact on resistance phenotypes.

4.3.3 Colistin resistance exhibits patterns of *de novo* evolution, onward dissemination, and reversion to susceptibility

To better understand the origin and fate of colistin resistance variants, we investigated the phylogenetic relationship among colistin resistant and susceptible isolates (Figures 4.2, 4.S3). Visualization of resistance on the phylogeny revealed a striking dichotomy between a clade II sublineage (defined as clade IIB) and the other ST258 sublineages (clades I and IIA). Clade IIB contains large clusters of colistin-resistant isolates, while the other two sublineages largely exhibit sporadic parallel evolution of resistance. Therefore, while 28/118 (24%) isolates acquired colistin resistance from putative *de novo* resistance evolution events, two clonal expansions in clade IIB across 11 LTACHs in California (10 in the Los Angeles area and 1 in San Diego; Figure 4.S4) accounted for over half (69/118, 58%) of resistance

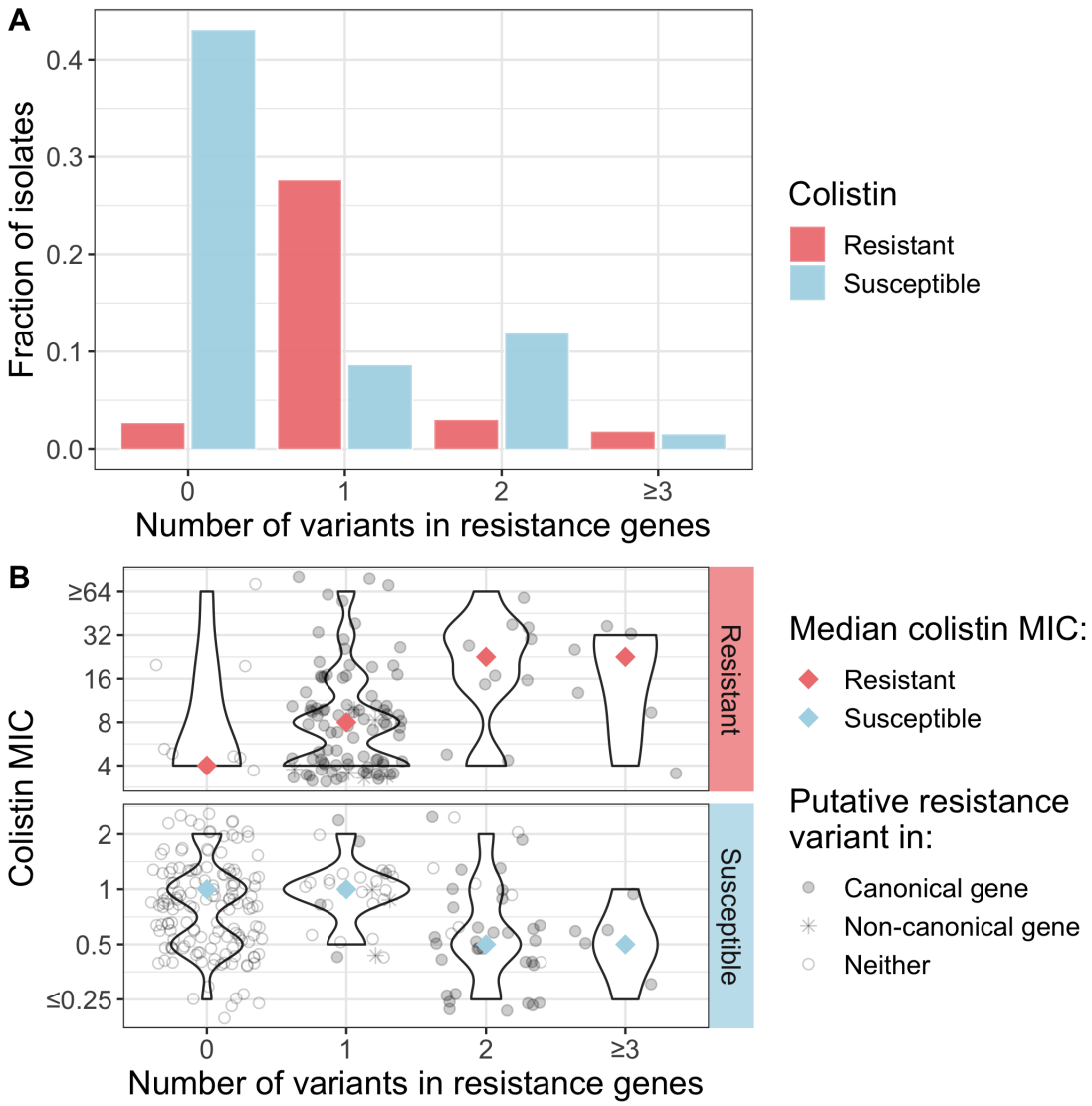

Figure 4.1: Epistatic interactions appear to influence the extent of colistin resistance in clinical CRKP isolates.

(A) Most susceptible isolates have no variants in known resistance genes, while most resistant isolates have at least one variant in a known resistance gene. Many susceptible isolates also have two variants in known resistance genes, suggesting potential reversion events. The fractions on the y axis are among all isolates. (B) On average, resistant isolates with two putative resistance variants have a higher MIC than resistance isolates with one putative variant, and susceptible isolates with two putative resistance variants have a lower MIC than susceptible isolates with one putative variant. MIC=minimum inhibitory concentration.

in this isolate collection. These two clonal expansions were due to a nonsense mutation in *mgrB* (Gln30*) and a missense mutation in *phoQ* (Thr244Asn), respectively. Notably, these same types of variants occurred in the sporadic colistin resistant isolates (Figure 4.S5), suggesting that these specific resistance variants were likely not inherently more fit. Within these clonal expansions, many of the isolates appeared to regain susceptibility to colistin via the accumulation of additional variants in resistance genes. These putative reversion mutations usually occurred within the same gene or in a downstream gene of the molecular pathway to resistance (Figures 4.2, 4.S6), and were sometimes followed by re-acquisition of resistance through a different mechanism (Figures 4.2, 4.S5).

4.3.4 Resistant strains in clade IIB are more fit than their susceptible non-revertant counterparts

Next, we were interested in understanding the relative fitness of susceptible, resistant and revertant isolates in the different sublineages. In particular, we hypothesized that the resistance variants in clade IIB became more widely disseminated due to a decreased associated fitness cost. To estimate fitness in the context of the healthcare environment we applied an analytic approach to quantify the epidemic success of isolates in our patient population. In particular, we took advantage of the comprehensive nature of our sampling and used the time-scaled haplotypic density (THD) to quantify the extent of relative spread of each genotype [131]. Applying the THD metric, we observed significant differences in fitness effects of resistance variants in clade IIB versus the other clades (Figure 4.3). In clades I and IIA we observed evidence of a significant fitness cost for resistance variants, as resistant clusters and susceptible revertants were less fit than susceptible non-revertants from those clades (R cluster: $p = 0.01$; S revertant: $p = 0.0003$). We did not observe a fitness cost for resistant singletons that did not transmit to others ($p = 0.18$), which may be because these strains

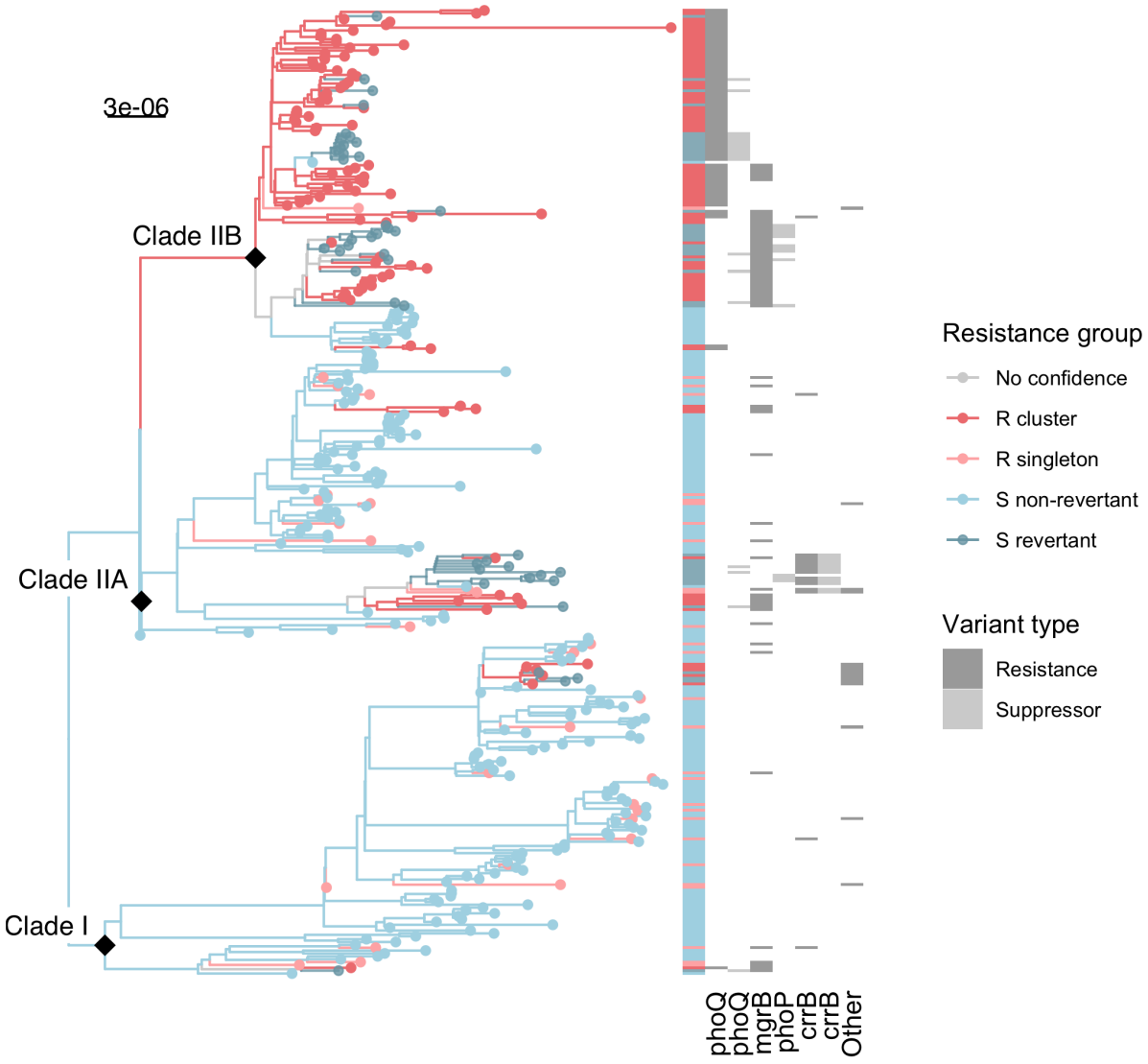

Figure 4.2: Distinct resistance evolution patterns exist in different CRKP ST258 sublineages.

Clade IIB exhibits extensive dissemination of resistance variants, as well as several putative reversion events, while the other clades contain sporadic and less transmitted resistance variants. Scale bar is in substitutions per site per year. R=resistant; S=susceptible.

very recently evolved resistance and are thus still closely related to their donor strains. In contrast, resistance variants in clade IIB were associated with a significant fitness benefit, with resistant clusters and susceptible revertants from clade IIB being more fit than susceptible non-revertants from that clade (R cluster: $p = 0.03$; S revertant: $p = 0.02$). The same findings hold when removing isolates from patients who have taken colistin in the past 30 days. Taken together, these findings suggest that resistance variants in isolates from clade IIB conferred a fitness advantage, which may account for the independent emergence and spread of two different resistance alleles in this clade. In contrast, the accumulation of variants in resistance genes in the other clades appears to be associated with a fitness cost, which is consistent with their limited clonal spread.

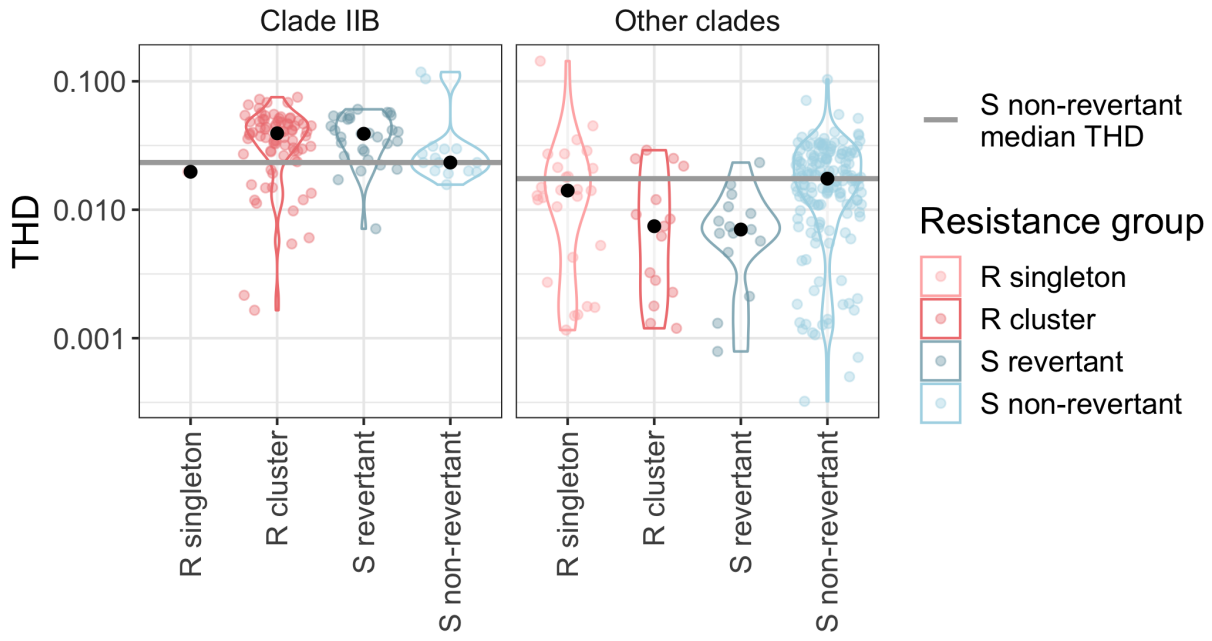

Figure 4.3: Colistin resistance does not impart a fitness cost in clade IIB strains.

In clade IIB, resistant clusters are more fit than corresponding susceptible non-revertants. On the other hand, resistant clusters in other clades are less fit than susceptible non-revertants in those clades. R=resistant; S=susceptible; THD=time-scaled haplotypic density.

4.3.5 Isolates with putative *de novo* resistance, dissemination, and reversion are each associated with exposure to colistin in the past 30 days

Lastly, we sought to gain insight into how the use of colistin in the study facilities impacted the emergence and dissemination of resistance variants. To do this, we investigated the relationship between different colistin resistance groups and the patient having taken colistin in the past 30 days. As expected based on previous studies [129], prior exposure to colistin was positively associated with putative *de novo* resistance evolution within patients (Figure 4.4). However, we unexpectedly found that prior exposure was also associated with both resistant clusters and susceptible revertants. This suggests that colistin use may contribute to the fitness benefit of resistant clusters and susceptible revertents observed in clade IIB in the THD analysis.

4.4 Discussion

The means by which colistin resistance evolves and disseminates in the healthcare setting has important implications for antibiotic stewardship and infection control. Here, we used genomic and clinical data to track the origin and fate of colistin resistance variants across a comprehensive regional sample of clinical CRKP ST258 isolates. Our analysis identified a sublineage whose genetic background appears to decrease the fitness cost of resistance, thus allowing colistin-resistant strains to spread among regional healthcare facilities.

Examination of colistin resistance variants within the comprehensive and longitudinal context of circulating CRKP strains allowed for inferences into their functional impact and clinical significance. We discovered that having multiple variants in known resistance genes appears to increase resistance in resistant isolates, but decrease resistance in susceptible

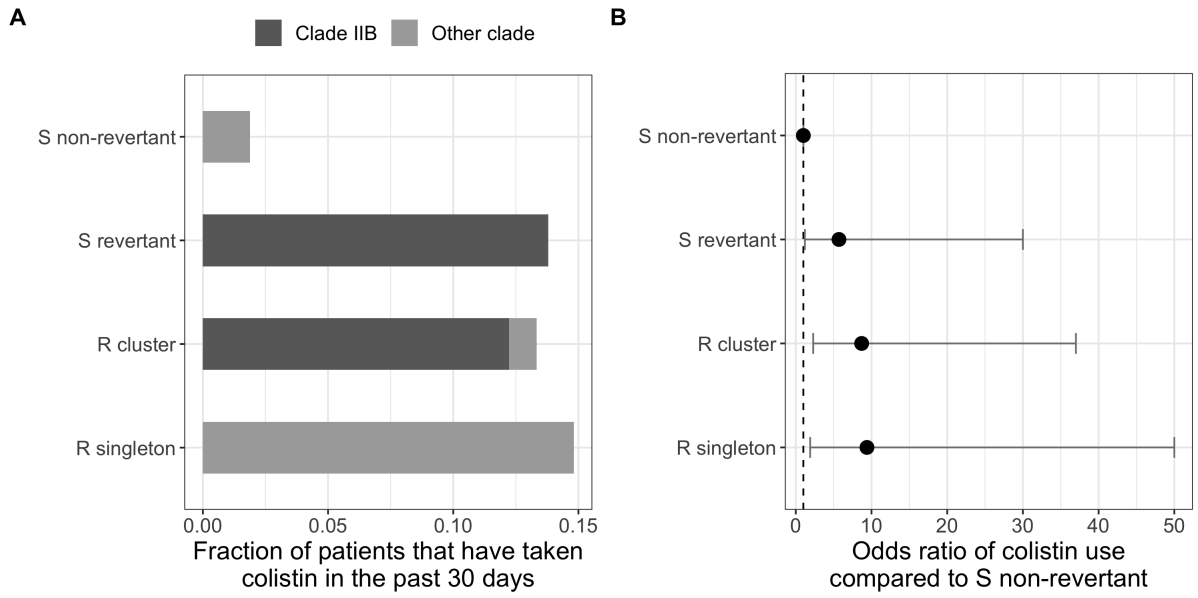

Figure 4.4: Susceptible non-revertants are disenriched in previous exposure to colistin compared to all other groups.

(A) The fraction of patients that had prior exposure to colistin was lower in susceptible non-revertants compared to all other groups. (B) The odds ratio of previous colistin exposure is higher for patients with resistant singletons, resistant clusters, and susceptible revertants when compared to susceptible non-revertants. The 95% confidence intervals of Fisher's exact odds ratios (error bars on the plot) were above one for all comparisons. R=resistant; S=susceptible.

isolates. Additionally, we identified several instances where a known resistance variant was present in susceptible isolates, likely due to a suppressor variant in either the same resistance gene or another resistance gene. This finding has two important implications. First, the presence of a colistin resistance variant cannot necessarily be equated with *in vitro* phenotypic resistance. Clinically, this suggests that testing for resistance variants may not be a substitute for testing for phenotypic resistance. In fact, while we did observe re-emergence of resistance in the genetic background of susceptible revertants, revertants may in general have more difficulty becoming resistant once more, which may lend credence to treating these patients with colistin [132]. Second, known resistance variants occurring in both resistant and susceptible isolates complicates the computational identification of novel resistance determinants using GWAS. Possible solutions include increasing sample size or removing isolates with known resistance variants.

Of particular genomic and clinical interest are two independent instances of emergence and spread of colistin resistance variants in clade IIB, in stark contrast to the sporadic emergence of resistance variants in the other clades. Notably, the types of resistance variants present in clade IIB also occurred in other clades, suggesting that the nature of the variant itself is not what allowed the clade IIB variants to spread. Instead, our analysis of fitness using THD suggests that these colistin resistance variants were able to spread in this clade because the genetic background of the clade reduced the fitness cost of the variants. Colistin is a cationic polypeptide that binds to the lipid A portion of lipopolysaccharide (LPS) and displaces divalent cations, thus disrupting the bacterial outer membrane and ultimately leading to cell death [128]. CRKP becomes resistant to colistin by adding a moiety to lipid A that increases its charge, thus decreasing colistin's ability to bind LPS [128]. However, these moieties may produce a fitness cost due to the changes they likely impart on the structure of the outer membrane. We previously showed that a defining feature of clade IIB is that the majority of isolates contain a disruption in the LPS O locus gene *kfoC*

[133], a putative glycosyltransferase. Thus, it is possible that inactivation of *kfoC* facilitates the spread of colistin resistance by altering the outer membrane in a way that reduces the fitness cost of resistance-conferring LPS modifications. This possibility has important clinical implications, as colistin resistance due to this lineage may become more prevalent over time, thus decreasing the number of patients who can be successfully treated with colistin.

Consistent with exposure to colistin driving the emergence of colistin resistance within patients, we identified a positive relationship between colistin use in the past 30 days and *de novo* resistance evolution. However, we also found a relationship between previous colistin use and resistant clusters, which may be due to the continued selective pressure to maintain resistance if colistin is being administered. Thus, although our findings support a reduced fitness cost of colistin for clade IIB strains, the prevalence of colistin resistance may partially be driven by colistin use in these facilities. Surprisingly, we also identified an association between previous colistin use and susceptible revertants. Because of this finding, we hypothesize that there may be a discrepancy between the *in vitro* MICs and clinical colistin resistance in the context of the patient [134]. If this is the case, then these additional variants may in fact be compensatory variants that reduce the fitness cost of resistance, or variants that alter the strain's interactions with the surrounding microbial community in such a way that confers resistance [135]. These findings about the relationship between previous colistin use and different resistance types highlight the power of using genomic context to investigate selective pressures imposed on these various groups in the clinic.

One strength of our study is that we have comprehensive sampling of clinical CRKP isolates from LTACHs over the course of a year. This not only allowed us to capture putative transmission events of colistin resistance, but also allowed us to estimate the fitness of different ST258 sublineages. Additionally, whole-genome sequencing of study isolates permitted us to interrogate not only variants in known resistance genes, but also investigate putative resistance-conferring variants in other genes using GWAS. Furthermore, our access to clini-

cal information about previous colistin use allowed us to identify a relationship between this clinical variable and different types of colistin susceptible and resistant isolates, providing us with insight into the selective pressures of colistin resistance evolution and spread in patients.

Our study also has several limitations. First, we do not have rectal surveillance cultures, which limits our ability to fully capture the population of colistin resistant and susceptible CRKP in the LTACHs studied. While this could bias our clinical collection if certain strains are more likely to be present and cultured at extraintestinal sites, our findings are still relevant in that they capture rampant transmission of strains that colonize and infect these sites. Also, we cannot be certain that our classification of isolates into different categories of colistin susceptibility and resistance is entirely accurate. For instance, putative *de novo* evolution of resistance in resistant singletons may not have occurred in the patient we sampled. However, our comprehensive sampling of clinical isolates from the facilities in the study provides some confidence that resistance in resistant singleton isolates did in fact evolve in that patient. Another limitation of our study is that we do not have information about colistin exposure prior to 30 days from the isolate collection date, and we do not have information about the dose or duration of colistin use. Even with this limitation, we were able to identify significant associations between colistin use and different colistin susceptible and resistant groups.

In conclusion, we observed distinct dynamics of colistin resistance evolution in different CRKP ST258 sublineages that appear to be due to differences in the fitness cost of colistin resistance dependent on the genetic background of the strain. In particular, we identified an emerging ST258 sublineage that is more fit and more amenable to maintaining colistin resistance than other ST258 strains. This is of particular concern due to the already limited treatment options for CRKP, and therefore merits further surveillance to determine the extent of spread. Furthermore, our findings highlight the importance of surveillance and monitoring of MDROs for continued evolution and adaptation to the healthcare environment,

and our method provides a framework for using genomics to study these complex evolutionary dynamics in clinical isolates.

4.5 Methods

4.5.1 Study isolates and metadata

We used whole-genome sequences of clinical CRKP isolates from a prospective longitudinal study in 21 U.S. LTACHs over the course of a year (BioProject accession no. PRJNA415194) [22]. Isolates and metadata were collected, and isolates were sequenced, as described in Han *et al.* [22]. Previous patient use of colistin in the past 30 days was extracted from the electronic health record.

4.5.2 Antibiotic susceptibility testing

To measure colistin resistance, we performed broth microdilution experiments to calculate the minimum inhibitory concentration (MIC) of each isolate to colistin using a customized Sensititre plate (MICs tested in μg/mL: 2^n where n goes from -2 to 6). To discretize colistin resistance, we use a cutoff of $\leq$2 as susceptible, and $\geq$4 as resistant, as per the Clinical & Laboratory Standards Institute guidelines. [136].

4.5.3 Isolate selection

Multi-locus sequence types were called using ARIBA [137]. Over 90% of CRKP isolates collected over the course of the study belonged to ST258; therefore, we focus all of our analyses on this sequence type. We ordered the isolates by resistance status followed by collection date. For all analyses except GWAS, only the first patient ST258 isolate from this ordered list was used, thus prioritizing resistant isolates over susceptible isolates. Sample

sizes were too small to glean insights into within-host resistance evolution.

4.5.4 Single nucleotide variant calling, indel calling, and phylogenetic tree reconstruction

Variant calling was performed with a customized variant calling pipeline (https://github.com/Snitkin-Lab-Umich/variant_calling_pipeline) as follows. The quality of sequencing reads was assessed with FastQC v0.11.9 [62], and Trimmomatic v0.39 [63] was used for trimming adapter sequences and low-quality bases. Single nucleotide variants (SNVs) were identified by (i) mapping filtered reads to the ST258 KPNIH1 reference genome (BioProject accession no. PRJNA73191) using the Burrows-Wheeler short-read aligner (bwa v0.7.17) [64], (ii) discarding polymerase chain reaction duplicates with Picard v2.24.1 [65], and (iii) calling variants with SAMtools and bcftools v1.9 [66]. Variants were filtered from raw results using VariantFiltration from GATK v4.1.9.0 [67] (QUAL $>$ 100; MQ $>$ 50; $\geq$ 10 reads supporting variant; and FQ $<$ 0.025). Indels were called using the GATK HaplotypeCaller [138] with the following filters: root mean square quality (MQ) $>$ 50.0, GATK QualbyDepth (QD) $>$ 2.0, read depth (DP) $>$ 9.0, and allele frequency (AF) $>$ 0.9. In addition, a custom Python script was used to filter out (mask) variants in the whole-genome alignment that were: (i) SNVs $<$ 5 base pairs (bp) in proximity to indels, (ii) in a recombinant region identified by Gubbins v2.3.4 [68], in a phage region identified by the Phaster web tool [139] or (iii) they resided in tandem repeats of length greater than 20bp as determined using the exact-tandem program in MUMmer v3.23 [140]. This whole-genome masked variant alignment was used to reconstruct a maximum likelihood phylogeny with IQ-TREE v1.6.12 [72] using the general time reversible model GTR+G and ultrafast bootstrap with 1000 replicates (-bb 1000) [73].

4.5.5 Insertion calling

Large insertions relative to the reference genome were called using panISa v0.1.4 [107].

4.5.6 Variant preprocessing

We preprocessed variants to include multiallelic sites and used the major allele method for variant binarization, as described in Saund *et al.* [91]. SnpEff was used to predict the functional impact of SNVs and indels (high, moderate, low, or modifier) [106]. Additionally, we considered all insertions in or upstream of genes as high impact, and those downstream of genes as moderate impact. Low impact variants were excluded from downstream analyses as they likely do not confer resistance and may induce noise into the analysis.

4.5.7 Identification of putative resistance genes

Known resistance genes

We consider the following genes known (canonical) resistance genes: *mgrB*, *phoP*, *phoQ*, *crrA*, *crrB*, *pmrA*, and *pmrB* [141]. We group variants in known resistance genes into the following categories, in order of decreasing confidence:

1. Known: experimentally confirmed resistance variants, including loss-of-function variants in *mgrB* [141, 142, 143, 144].

2. Known site: the variant occurs at a site where there is an experimentally confirmed resistance variant, but that specific amino acid change has not been experimentally confirmed [145, 146, 147, 148].

3. Putative: nonsynonymous or disruptive variants in known resistance genes where >60% of the variants are present in resistant isolates.

We define the set of all of these as resistance variants.

Genome-wide association study

We performed a burden test using treeWAS v1.0 [149], a convergence-based GWAS method, as our sample size is relatively small and resistance is a very convergent phenotype. A burden test increases power to detect resistance genes, and convergence-based methods control for the structure of the phylogeny more than mixed model methods. For GWAS, we included only isolates with no mutations in known resistance genes as we were most interested in identifying novel resistance genes, and because the entire dataset contained a number of susceptible isolates with known resistance variants that would have confounded the analysis. We used pyseer v1.3.6 [150] to calculate the number of unique patterns and determine a p-value cutoff ($p < 9.47e\text{-}5$). Putative resistance genes were considered ones that were identified as significant by treeWAS, had more than one convergence event on the phylogeny, and >60% of all the variants in the gene were found in resistant isolates (including isolates with variants in canonical resistance genes). Variants within these genes were considered putative resistance variants using the same definition as for known resistance genes. Using these requirements, we included four putative resistance variants from two putative resistance genes based on the GWAS results.

4.5.8 Identification of putative suppressor variants

We define putative suppressor variants as those in resistance genes where >60% of isolates with the variant are susceptible and contain a resistance variant.

4.5.9 Determination of isolate resistance group

We assigned each isolate into one of four categories (Figure 4.S3). Clusters were identified using the `get_clusters` function in `regentrans` (https://github.com/Snitkin-Lab-Umich/regentrans). The four categories are:

1. Resistant singletons: resistant isolates that do not cluster on the phylogeny, or that cluster but contain distinct resistance variants.

2. Resistant clusters: resistant isolates that cluster on the phylogeny and contain the same resistance variant. Additionally, if a cluster on the phylogeny has unknown resistance variants, we also defined it as a resistant cluster.

3. Susceptible revertants: susceptible isolates that contain a putative resistance-conferring variant.

4. Susceptible non-revertants: susceptible isolates that do not contain a putative resistance-conferring variant.

4.5.10 Calculation of time-scaled haplotypic density

We calculated time-scaled haplotypic density (THD) for each isolate with the R package thd v1.0.1 [131] using the KPNIH1 reference genome length (5,394,056 base pairs), a mutation rate of 1.03e-6 [21], the time scale parameter, and a timescale of one year. Pairwise single nucleotide variant distances were calculated with the dist.dna function in ape v5.4.1 [118] (model = 'N', pairwise.deletion = TRUE). Significant differences between THD of different isolates were determined using Wilcox tests.

4.5.11 Calculation of previous colistin use odds ratios

We calculated Fisher's exact odds ratios and 95% confidence intervals using the R package exact2x2 v1.6.5 [151].

4.5.12 Data visualization

We performed all data visualization in R v4.0.2 [110] using the following packages: tidyverse v1.3.0 [114], ggtree v2.2.4 [116, 117], pheatmap v1.0.12 [152], ggplotify v0.0.5 [153], and cowplot v1.1.0 [115].

4.6 Supplement

4.6.1 Figures

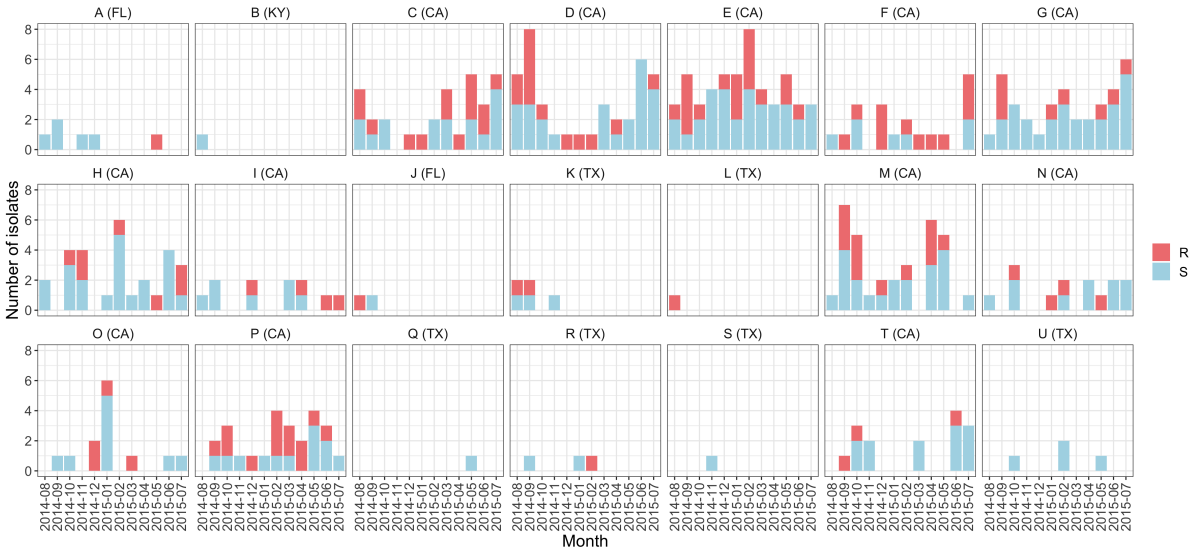

Figure 4.S1: Colistin resistance occurred across time and geography.

R=resistant, S=susceptible.

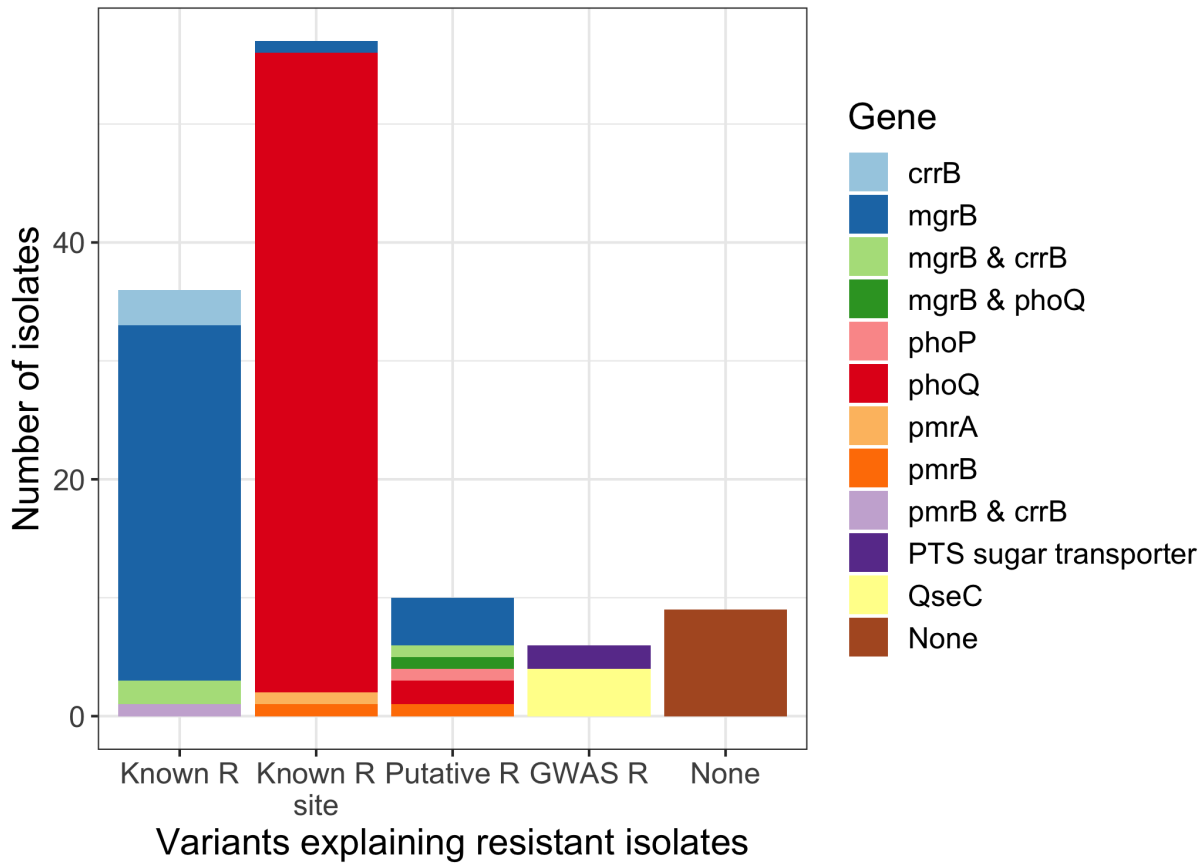

Figure 4.S2: The majority of colistin resistance can be explained by variants in known resistance genes.

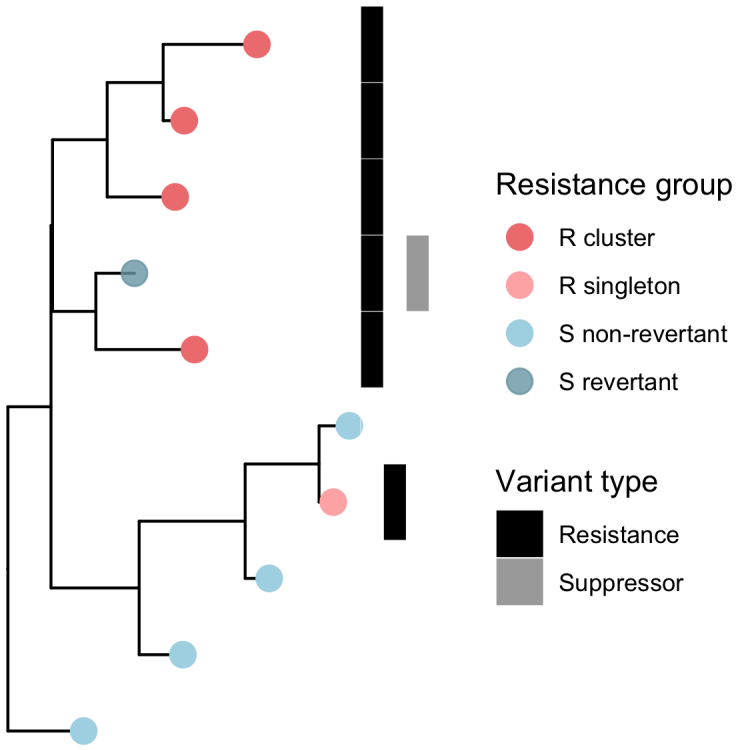

Figure 4.S3: Toy example of how resistance groups are defined.

Note that susceptible revertants are defined by the presence of a resistance variant, rather than a putative suppressor variant. See methods for more details. Heatmap columns are each a different variant found in a resistance gene. R=resistant, S=susceptible.

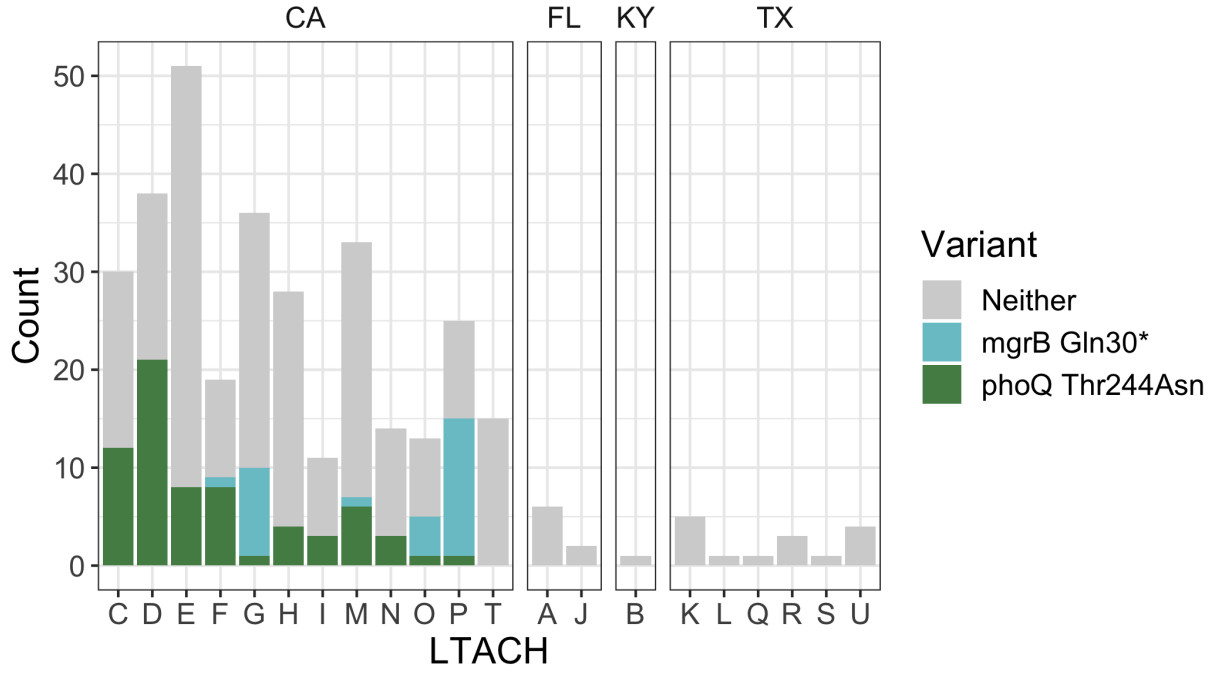

Figure 4.S4: The two clonally expanded resistance variants occur across multiple LTACHs in California.

10/11 LTACHs in the Los Angeles area have the variant, as well as one LTACH in the San Diego area.

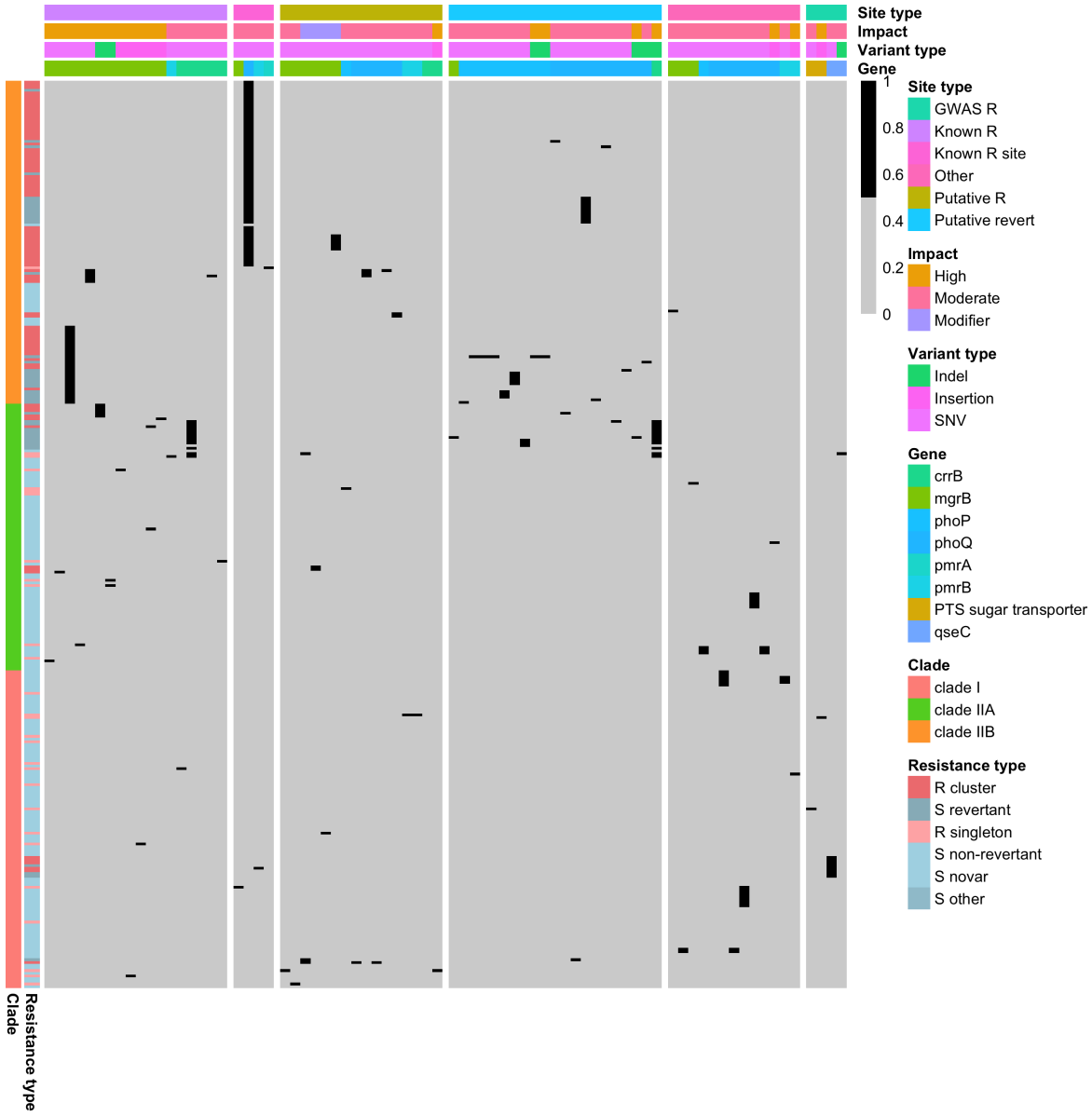

Figure 4.S5: Putative resistance and suppressor variants in canonical and non-canonical resistance genes.

Columns are variants and rows are isolates. All variants except the last panel are from canonical resistance genes. R=resistant, S=susceptible.

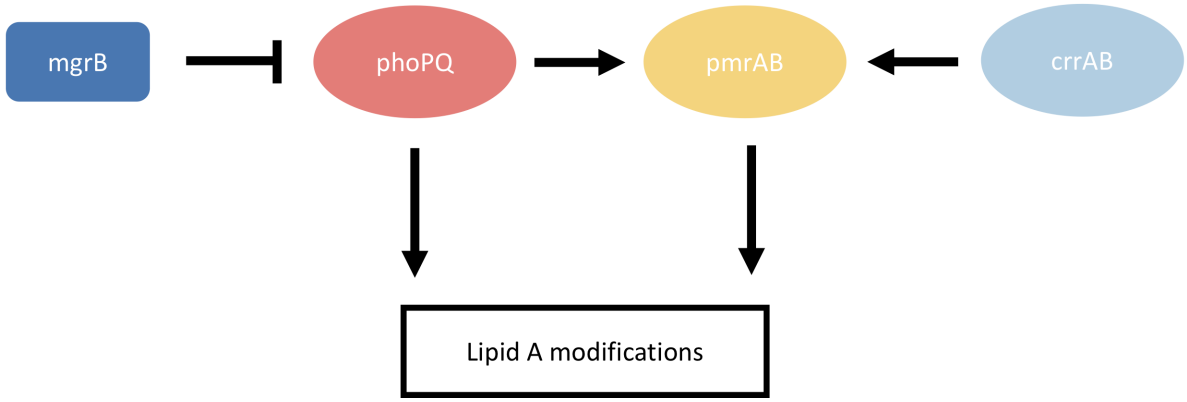

Figure 4.S6: Schematic of the molecular pathway of canonical colistin resistance genes.

4.6.2 Table

Table 4.S1: Resistance variants.

Gene	Variant type	Impact	Site type	Variant	Isolate count
mgrB	SNV	high	known	Trp47*	1
mgrB	SNV	high	known	Lys43*	1
mgrB	SNV	high	known	Gln30*	41
mgrB	SNV	high	known	Gln22*	1
mgrB	SNV	high	known	Lys3*	6
mgrB	Indel	high	known	Ile12fs	6
mgrB	Indel	high	known	Leu8fs	2
mgrB	Insertion	high	known	NA	1
mgrB	Insertion	high	known	NA	1
mgrB	Insertion	high	known	NA	1
mgrB	Insertion	high	known	NA	2
mgrB	Insertion	high	known	NA	1
pmrB	SNV	moderate	known	Leu82Arg	1
crrB	SNV	moderate	known	Gln10Leu	1
crrB	SNV	moderate	known	Leu94Met	16
crrB	SNV	moderate	known	Pro151Ser	2
crrB	SNV	moderate	known	Pro151Ser	1
crrB	SNV	moderate	known	Gly183Val	1
mgrB	SNV	moderate	known_site	Asp31Asn	1
phoQ	SNV	moderate	known_site	Thr244Asn	85
pmrB	SNV	moderate	known_site	Pro158Gln	1
pmrA	SNV	moderate	known_site	Gly53Ser	1

mgrB	SNV	moderate	putative_r	Ile41Asn	1
mgrB	SNV	moderate	putative_r	Asn25Thr	1
mgrB	SNV	modifier	putative_r		3
mgrB	SNV	modifier	putative_r		3
mgrB	SNV	modifier	putative_r		1
mgrB	SNV	modifier	putative_r		9
phoP	SNV	moderate	putative_r	Gly53Val	1
phoQ	SNV	moderate	putative_r	Leu308Gln	1
phoQ	SNV	moderate	putative_r	Leu257Gln	4
phoQ	SNV	moderate	putative_r	Thr248Pro	1
phoQ	SNV	moderate	putative_r	Leu105Gln	2
phoQ	SNV	moderate	putative_r	Leu62Pro	2
pmrB	SNV	moderate	putative_r	Leu262Pro	1
pmrB	SNV	moderate	putative_r	Ser203Pro	1
crrB	SNV	moderate	putative_r	Ala200Pro	1
crrB	Insertion	high	putative_r	NA	1
mgrB	SNV	moderate	putative_revert	Thr21Pro	1
phoP	SNV	moderate	putative_revert	Gln202Pro	3
phoP	SNV	moderate	putative_revert	Thr189Pro	1
phoP	SNV	moderate	putative_revert	His188Gln	1
phoP	SNV	moderate	putative_revert	His188Arg	1
phoP	SNV	moderate	putative_revert	Arg185Leu	3
phoP	SNV	moderate	putative_revert	Arg111His	11
phoP	SNV	moderate	putative_revert	Val18Ala	4
phoP	Indel	high	putative_revert	Thr189fs	1
phoP	Indel	high	putative_revert	His188fs	1
phoQ	SNV	moderate	putative_revert	Gln481Leu	1
phoQ	SNV	moderate	putative_revert	Ala451Thr	1
phoQ	SNV	moderate	putative_revert	Gly443Ser	1
phoQ	SNV	moderate	putative_revert	Arg433Leu	13
phoQ	SNV	moderate	putative_revert	Asp417Val	2
phoQ	SNV	moderate	putative_revert	Leu322Met	1
phoQ	SNV	moderate	putative_revert	Glu123Lys	2
phoQ	SNV	moderate	putative_revert	Ala36Thr	1
phoQ	Indel	high	putative_revert	Leu344fs	1
phoQ	Indel	moderate	putative_revert	Thr248_Leu254del	1
crrB	Indel	high	putative_revert	Ser338fs	16
mgrB	SNV	moderate	other	Phe44Cys	1
mgrB	SNV	moderate	other	Leu24Pro	3
mgrB	SNV	moderate	other	Leu19Gln	1
phoP	SNV	moderate	other	Ala80Gly	3
phoQ	SNV	moderate	other	Ile462Ser	1

phoQ	SNV	moderate	other	Leu365His	8
phoQ	SNV	moderate	other	Pro343Leu	3
phoQ	SNV	moderate	other	Thr332Pro	9
phoQ	SNV	moderate	other	Arg50Leu	7
phoQ	SNV	moderate	other	Gly33Ser	3
phoQ	Insertion	high	other	NA	1
pmrB	SNV	moderate	other	Pro95Gln	3
pmrB	Insertion	high	other	NA	1
PTS sugar transporter	SNV	moderate	gwas_r	Ser72Ile	1
PTS sugar transporter	Insertion	high	gwas_r	NA	1
qseC	SNV	moderate	gwas_r	Val180Ala	9
qseC	Indel	moderate	gwas_r	Arg12_Leu13del	1

Abbreviations: SNV, single nucleotide variant; PTS; phosphotransferase system; known, known resistance variant; known_r, known_site, known resistance variant site; putative_r, putative resistance variant in a known resistance gene; putative_revert, putative suppressor variant; gwas_r, putative resistance variant from GWAS.

Chapter 5

Discussion

5.1 Major dissertation contributions

This dissertation integrates genomic and clinical data to gain insight into transmission and adaptation of CRKP in the healthcare environment and provides a framework and corresponding tools for applying the methods developed here to other MDROs. The following sections summarize the methods, insights, and bioinformatic tools that this dissertation contributes.

5.1.1 Regional pathogen transmission

Methods

In chapter two, whole-genome sequencing facilitated the investigation of endemic CRKP ST258 transmission within and between regional healthcare facilities. First, we demonstrate the utility of using public whole-genome sequences to investigate the extent of importation into a region. Additionally, our comprehensive sampling of clinical isolates allowed us to estimate the magnitude of within- and between-facility transmission and compare the extent of between-facility transmission to patient transfer events. Finally, we combined the results of our transmission analysis with patient factors to interrogate potential drivers of CRKP

transmission within healthcare facilities.

Results and implications

We found that, while transmission between facilities is common, within-facility transmission drives prevalence of CRKP at a given facility. Furthermore, we identified certain patient factors that may influence the extent of within-facility transmission. These findings have several implications for public health efforts to reduce CRKP spread in this healthcare network. First, since within-facility transmission drives prevalence, focusing infection control efforts on reducing intra-facility transmission at high-transmission facilities might reduce CRKP prevalence more than coordinated efforts between facilities in this region. Second, future epidemiological studies can investigate the potential patient risk factors associated with transmission identified here [154]. Moreover, while these findings about how CRKP spreads among LTACHs in the Los Angeles area may not be generalizable to other MDROs, regions, or healthcare settings, the methods applied here can be used to investigate transmission in other scenarios.

5.1.2 Pathogen evolution and adaptation

Methods

In chapters three and four, genomic analysis provided insight into the continued evolution and adaptation of CRKP ST258 to the healthcare setting. In chapter three, we demonstrate the utility of using machine learning to identify pathogen sublineages related to a certain clinical factor. In chapter four, we integrate phenotypic, phylogenetic, genomic, and clinical features to gain insight into antibiotic resistance evolution and further probe the evolutionary dynamics of different ST258 subclades. Furthermore, our comprehensive sampling enabled us to compare fitness differences between resistant and susceptible isolates in

different sublineages, and corresponding clinical metadata allowed us to gain insight into additional potential selective pressures related to antibiotic resistance evolution.

Results and implications

We identified an ST258 subclade (clade IIB) present in California LTACHs that appears to have adapted to the respiratory tract and subsequently evolved increased virulence. Concerningly, clade IIB includes large colistin resistant sublineages that have arisen from clonal dissemination of resistance, in stark contrast to the sporadic resistance we observe in other sublineages. Moreover, these clusters of resistant strains seem to be more transmissible than susceptible non-revertant strains in the subclade, unlike the colistin resistant strains in other subclades. This finding has important public health implications. First, this emergent and possibly more virulent sublineage may continue to spread across facilities in the region and could possibly be exported out of the region. This is particularly concerning as colistin resistance appears to impart less of a fitness cost in clade IIB strains, indicating that this strain may continue to spread and gain a foothold in the region, after which this antibiotic may no longer be a viable treatment option for the many patients with these strains. Second, our identification of an emerging and potentially more transmissible and virulent sublineage highlights the importance of continued monitoring and surveillance of MDROs using integrated methods such as whole-genome sequencing combined with patient and phenotypic metadata to identify emerging strains.

5.1.3 Bioinformatic contributions

This dissertation led to the development of three open-source R packages. The `regentrans` package implements methods to study within- and between-facility transmission in a regional network of healthcare facilities and includes several of the methods used in chapter two. The `mikropml` package implements the robust machine learning framework used in chapter three

to make it more accessible for others to use in their own research. The `prewas` package implements the method used in chapter four to preprocess bacterial genomic variants from whole-genome sequencing for downstream analysis. These packages will allow other investigators to apply the methods developed and used here to their own work.

5.2 Future work

The comprehensive dataset used in this dissertation allowed us to investigate many questions related to CRKP regional transmission and evolution. Future work should aim to build upon these analyses with additional studies that overcome some of the limitations of the dataset used here. Limitations of this dataset include that we do not have rectal surveillance cultures, multiple isolates per patient, longitudinal sampling of patients, or samples from other regional healthcare facilities. While we were able to glean many insights into CRKP transmission and adaptation solely using a dataset from LTACHs that includes a single clinical isolate from each patient included in the study, this additional data would allow us to gain an even more nuanced understanding of transmission and adaptation within and among patients and healthcare facilities. The following sections describe future work that could be performed using this extended dataset, and potential insights gleaned from this work.

5.2.1 Surveillance culturing paired with clinical samples

The vast majority of patients colonized with CRKP remain undetected when using clinical cultures alone [155, 156, 157]. Performing rectal surveillance culturing of patients in addition to collecting clinical cultures would allow us to investigate the hidden portion of the colonization iceberg that is not captured by clinical cultures, as well as differences between gastrointestinal and extraintestinal isolates.

Transmission

Collecting surveillance cultures in addition to clinical cultures would allow us to investigate the full CRKP transmission network. We could probe the connections between patients with gastrointestinal and extraintestinal colonization (or infection), and compare inferences made between transmission networks generated with clinical cultures alone to those generated with surveillance cultures. If most transmission events occur between patients with clinical cultures, infection control efforts could focus on those patients rather than patients that are only gastrointestinally colonized. However, if transmission between these two groups of patients patients is common, surveillance culturing could provide important information for reducing CRKP transmission.

Evolution and adaptation

Rectal surveillance culturing would also allow us to compare differences in CRKP isolates from gastrointestinal and extraintestinal colonization. One question of particular interest is whether clade IIB is enriched in extraintestinal colonization compared to the other clades, which could indicate that strains from this clade are more likely to migrate from the gastrointestinal tract to the respiratory tract, and thus may be more likely to cause infections. Furthermore, surveillance culturing could provide insight into the extent of undetected colistin resistance in CRKP, which may have implications for when and how to use colistin in the clinic. More generally, consistent surveillance culturing paired with clinical cultures allows for the early identification of more fit sublineages, potentially yielding insight into how they arise, and providing the opportunity to attempt to curtail their spread.

5.2.2 Multiple isolates per patient and longitudinal sampling of patients

Patients can be multiply colonized with different CRKP strains or sequence types at the same or multiple body sites, and patients can be colonized for extended periods of time, leading to variation between strains from the same patient due to within-host evolution [158]. Additionally, CRKP colonization is a risk factor for future infection [159]. Collection of multiple isolates per patient at multiple points in time would allow us to investigate transmission and within-host evolution in more detail.

Transmission

Including only a single isolate from each patient when investigating transmission limits the resolution of the transmission network if many patients are co-colonized with multiple strains, either at a single time point or longitudinally. Therefore, having multiple and longitudinal isolates per patient would allow for a higher resolution analysis of patient-to-patient transmission. Paired with surveillance cultures, this data would allow us to investigate how incorporating within-host diversity into transmission analyses yields additional insight into pathways of transmission both within and between healthcare facilities.

Evolution and adaptation

Longitudinal sampling of patients would also allow us to capture within-host evolution. Using surveillance cultures plus clinical isolates at one or multiple body sites for each patient, we could investigate genomic and clinical features related to the transition of pathogens from the gastrointestinal tract to other body sites, and subsequent evolution at those body sites. This may provide insight into the biological basis of translocation, and predictors of this event. Of specific interest is the trajectory by which patients become colonized with clade IIB

isolates in the respiratory tract to further probe the basis of this association. Longitudinal samples would also allow us to capture within-host evolution of virulence and antibiotic resistance, providing more direct evidence of genomic and clinical features associated with these phenotypes. These investigations paired with experimental follow-up studies would provide valuable insight into pathogen within-host evolution related to adaptation in the healthcare environment.

5.2.3 Samples from other regional healthcare facilities

There is substantial evidence that LTACHs are hotspots of CRKP transmission [45], and that interventions in these facilities reduce overall CRKP prevalence across the healthcare network [49]. For this reason, our analysis focused on transmission and evolution in LTACHs. However, acquisition of CRKP samples from other facilities in the healthcare network would allow for a more nuanced understanding of inter-facility transmission dynamics. While we have strong evidence that patient transfer often drives inter-facility transmission, direct transfers between LTACHs are rare. In chapter two, we identified paths of maximum patient flow between LTACHs that usually included at least one intermediate facility, and these paths correlated well with the extent of transmission between LTACHs. It would be interesting to determine whether these identified paths of maximum patient flow are the true transmission paths, and whether there are features of the intermediate facilities, or patients at these facilities, that drive between-facility transmission. These insights could provide valuable information for regional infection prevention by highlighting facilities or facility-level factors that amplify or reduce inter-facility transmission.

5.2.4 Bioinformatic tools

In this dissertation, we took the approach of developing methods to answer specific questions of interest, and subsequently creating open-source tools for these methods. This allowed us to find gaps in the field and build tools to fill those gaps. The future work described above should continue to follow this framework of developing user-friendly open source methods and tools for genomic epidemiology. Additionally, these tools must be maintained and potentially even expanded so that they remain useful to users over time.

5.3 Moving towards real-time genomic epidemiology

While this dissertation acts as a proof-of-principle study for investigating pathogen transmission and adaptation in the healthcare environment, the ultimate goal is to integrate the methods developed, and insights gained, here and elsewhere into real-time genomic epidemiology in the public health system. This requires a coordinated surveillance and response framework, increased bioinformatic analysis capacity at public health institutions, and clear communication of findings between researchers, healthcare workers, and the public.

5.3.1 Building a coordinated surveillance and response framework

As we move from retrospective to real-time analyses, we will need a coordinated surveillance and response framework for practical, yet informative and timely, sample collection, sequencing, and analysis. This process could be facilitated by a unified healthcare system that allows for streamlined access to patient clinical data and clinical cultures across regions. Importantly, this pipeline must include collaboration and communication between clinicians, epidemiologists, bioinformaticians, microbiologists, and public officials, as well as clear and timely communication among experts and the community. Lessons learned from this dissertation that may be incorporated into this pipeline include the use of clinical cultures from

LTACHs to investigate regional MDRO transmission, and the utility of clinical cultures for identifying emerging clones and studying the dynamics of antibiotic resistance evolution.

5.3.2 Increasing bioinformatic analysis capacity at public health institutions

In this dissertation we developed R packages that, paired with other pre-existing tools, can be used by those with relatively little coding experience to study pathogen transmission and adaptation. This is a very small contribution towards the ultimate development and implementation of tools that people with little to no programming expertise can use to analyze data. Empowering individuals at local and regional healthcare institutions to perform preliminary analysis of data collected from surveillance systems is key to establishing the ability to perform local surveillance nationwide. However, collaborations and communication with bioinformaticians and researchers with more area-specific expertise is also necessary to follow up on findings that may require more complex analysis, as well as stay up to date on cutting edge methods and findings. These insights should then be incorporated into public health standards and policy. The importance of collaborations between public health and research organizations highlights the interdisciplinary nature of this work, and the importance of clear communication between these different individuals and entities.

5.4 Conclusion

The work in this dissertation was motivated by the goal of reducing nosocomial transmission, morbidity, and mortality. Here, we focus on CRKP due to global prevalence, limited treatment options, and high mortality rates. We develop and apply methods to study regional transmission of endemic MDROs using whole-genome sequencing data and demonstrate that these tools can provide actionable insights into where transmission is occurring. We also il-

lustrate the utility of using machine learning for investigating patient outcomes, identify different patterns of antibiotic resistance evolution in different genomic backgrounds, and highlight the importance of continual high-resolution monitoring of MDROs for novel potentially high-risk strains. These contributions represent a step in the direction toward our long-term goal of real-time genomic epidemiology that integrates whole-genome sequencing, clinical metadata, and phenotypic data to monitor and control endemic and emerging pathogens.

Bibliography

[1] Evelina Tacconelli et al. "Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis". In: *The Lancet Infectious Diseases* (Dec. 2017). ISSN: 1473-3099. DOI: 10.1016/S1473-3099(17)30753-3. URL: http://www.sciencedirect.com/science/article/pii/S1473309917307533.

[2] C. Friedlaender. "Ueber die Schizomyceten bei der acuten fibrösen Pneumonie". de. In: *Archiv f. pathol. Anat.* 87.2 (Feb. 1882), pp. 319–324. ISSN: 1432-2307. DOI: 10.1007/BF01880516. URL: https://doi.org/10.1007/BF01880516.

[3] R. Podschun and U. Ullmann. "Klebsiella spp. as Nosocomial Pathogens: Epidemiology, Taxonomy, Typing Methods, and Pathogenicity Factors". en. In: *Clinical Microbiology Reviews* 11.4 (Oct. 1998). Publisher: American Society for Microbiology Journals Section: ARTICLE, pp. 589–603. ISSN: 0893-8512, 1098-6618. DOI: 10.1128/CMR.11.4.589. URL: http://cmr.asm.org/content/11/4/589.

[4] Shiri Navon-Venezia, Kira Kondratyeva, and Alessandra Carattoli. "Klebsiella pneumoniae: a major worldwide source and shuttle for antibiotic resistance". In: *FEMS Microbiology Reviews* 41.3 (May 2017), pp. 252–275. ISSN: 0168-6445. DOI: 10.1093/femsre/fux013. URL: https://doi.org/10.1093/femsre/fux013.

[5] L Silvia Munoz-Price et al. "Clinical epidemiology of the global expansion of Klebsiella pneumoniae carbapenemases". In: *Lancet Infect Dis* 13.9 (Sept. 2013), pp. 785–796. ISSN: 1473-3099. DOI: 10.1016/S1473-3099(13)70190-7. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4673667/.

[6] H. Yigit et al. "Novel carbapenem-hydrolyzing beta-lactamase, KPC-1, from a carbapenem-resistant strain of Klebsiella pneumoniae". eng. In: *Antimicrob Agents Chemother* 45.4 (Apr. 2001), pp. 1151–1161. ISSN: 0066-4804. DOI: 10.1128/AAC.45.4.1151-1161.2001.

[7] Maryn McKenna. "Antibiotic resistance: The last resort". en. In: *Nature News* 499.7459 (July 2013). Section: News Feature, p. 394. DOI: 10.1038/499394a. URL: http://www.nature.com/news/antibiotic-resistance-the-last-resort-1.13426.

[8] Liangfei Xu, Xiaoxi Sun, and Xiaoling Ma. "Systematic review and meta-analysis of mortality of patients infected with carbapenem-resistant Klebsiella pneumoniae". In:

Ann Clin Microbiol Antimicrob 16 (Mar. 2017). ISSN: 1476-0711. DOI: 10.1186/s12941-017-0191-3. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5371217/.

[9] Ryan S. Arnold et al. "Emergence of Klebsiella pneumoniae Carbapenemase (KPC)-Producing Bacteria". In: *South Med J* 104.1 (Jan. 2011), pp. 40–45. ISSN: 0038-4348. DOI: 10.1097/SMJ.0b013e3181fd7d5a. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3075864/.

[10] Mariano Ciccolini et al. "Infection prevention in a connected world: The case for a regional approach". In: *International Journal of Medical Microbiology*. Special Issue Antibiotic Resistance 303.6–7 (Aug. 2013), pp. 380–387. ISSN: 1438-4221. DOI: 10.1016/j.ijmm.2013.02.003. URL: http://www.sciencedirect.com/science/article/pii/S1438422113000180.

[11] Bruce Y. Lee et al. "The Regional Healthcare Ecosystem Analyst (RHEA): a simulation modeling tool to assist infectious disease control in a health system". eng. In: *J Am Med Inform Assoc* 20.e1 (June 2013), e139–146. ISSN: 1527-974X. DOI: 10.1136/amiajnl-2012-001107.

[12] Prabasaj Paul et al. "Modeling Regional Transmission and Containment of a Healthcare-associated Multidrug-resistant Organism". In: *Clinical Infectious Diseases* 70.3 (Jan. 2020), pp. 388–394. ISSN: 1058-4838. DOI: 10.1093/cid/ciz248. URL: https://doi.org/10.1093/cid/ciz248.

[13] Rachel B. Slayton et al. "Vital signs: Estimated effects of a coordinated approach for action to reduce antibiotic-resistant infections in health care facilities - United States". English (US). In: *American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons* 15.11 (Nov. 2015). Publisher: Wiley-Blackwell, pp. 3002–3007. ISSN: 1600-6135. DOI: 10.1111/ajt.13555. URL: https://jhu.pure.elsevier.com/en/publications/vital-signs-estimated-effects-of-a-coordinated-approach-for-actio-6.

[14] Bruce Y. Lee et al. "How to Choose Target Facilities in a Region to Implement Carbapenem-resistant Enterobacteriaceae Control Measures". en. In: *Clin Infect Dis* 72.3 (Feb. 2021). Publisher: Oxford Academic, pp. 438–447. ISSN: 1058-4838. DOI: 10.1093/cid/ciaa072. URL: http://academic.oup.com/cid/article/72/3/438/5714274.

[15] Sarah Y. Won et al. "Emergence and rapid regional spread of Klebsiella pneumoniae carbapenemase-producing Enterobacteriaceae". eng. In: *Clin Infect Dis* 53.6 (Sept. 2011), pp. 532–540. ISSN: 1537-6591. DOI: 10.1093/cid/cir482.

[16] Shawn E. Hawken et al. "A novel threshold-independent approach to genomic cluster analysis discloses persistent routes of KPC+ Klebsiella pneumoniae transmission in a long-term acute care hospital". en. In: *medRxiv* (Sept. 2020). Publisher: Cold Spring

Harbor Laboratory Press, p. 2020.09.26.20200097. DOI: 10.1101/2020.09.26.20200097. URL: https://www.medrxiv.org/content/10.1101/2020.09.26.20200097v1.

[17] Stephen Baker et al. "Genomic insights into the emergence and spread of antimicrobial-resistant bacterial pathogens". en. In: *Science* 360.6390 (May 2018). Publisher: American Association for the Advancement of Science Section: Review, pp. 733–738. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aar3777. URL: http://science.sciencemag.org/content/360/6390/733.

[18] Evan S. Snitkin et al. "Integrated genomic and interfacility patient-transfer data reveal the transmission pathways of multidrug-resistant Klebsiella pneumoniae in a regional outbreak". eng. In: *Sci Transl Med* 9.417 (Nov. 2017). ISSN: 1946-6242. DOI: 10.1126/scitranslmed.aan0093.

[19] Tjibbe Donker et al. "Population genetic structuring of methicillin-resistant Staphylococcus aureus clone EMRSA-15 within UK reflects patient referral patterns". eng. In: *Microb Genom* 3.7 (July 2017), e000113. ISSN: 2057-5858. DOI: 10.1099/mgen.0.000113.

[20] Frank R. DeLeo et al. "Molecular dissection of the evolution of carbapenem-resistant multilocus sequence type 258 Klebsiella pneumoniae". en. In: *PNAS* 111.13 (Apr. 2014). Publisher: National Academy of Sciences Section: Biological Sciences, pp. 4988–4993. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1321364111. URL: https://www.pnas.org/content/111/13/4988.

[21] Jolene R. Bowers et al. "Genomic Analysis of the Emergence and Rapid Global Dissemination of the Clonal Group 258 Klebsiella pneumoniae Pandemic". eng. In: *PLoS ONE* 10.7 (2015), e0133727. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0133727.

[22] Jennifer H. Han et al. "Whole-Genome Sequencing To Identify Drivers of Carbapenem-Resistant Klebsiella pneumoniae Transmission within and between Regional Long-Term Acute-Care Hospitals". en. In: *Antimicrobial Agents and Chemotherapy* 63.11 (Nov. 2019). ISSN: 0066-4804, 1098-6596. DOI: 10.1128/AAC.01622-19. URL: https://aac.asm.org/content/63/11/e01622-19.

[23] Jane W. Marsh et al. "Evolution of Outbreak-Causing Carbapenem-Resistant Klebsiella pneumoniae ST258 at a Tertiary Care Hospital over 8 Years". en. In: *mBio* 10.5 (Oct. 2019), e01945–19. ISSN: 2150-7511. DOI: 10.1128/mBio.01945-19. URL: https://mbio.asm.org/content/10/5/e01945-19.

[24] Angela Gomez-Simmonds and Anne-Catrin Uhlemann. "Clinical Implications of Genomic Adaptation and Evolution of Carbapenem-Resistant Klebsiella pneumoniae". en. In: *J Infect Dis* 215.suppl_1 (Feb. 2017), S18–S27. ISSN: 0022-1899. DOI: 10.1093/infdis/jiw378. URL: https://academic.oup.com/jid/article/215/suppl_1/S18/3092086.

[25] Michael A. Bachman et al. "Klebsiella pneumoniae yersiniabactin promotes respiratory tract infection through evasion of lipocalin 2". eng. In: *Infection and Immunity* 79.8 (Aug. 2011), pp. 3309–3316. ISSN: 1098-5522. DOI: 10.1128/IAI.05114-11.

[26] Christoph M. Ernst et al. "Adaptive evolution of virulence and persistence in carbapenem-resistant Klebsiella pneumoniae". en. In: *Nature Medicine* 26.5 (May 2020). Number: 5 Publisher: Nature Publishing Group, pp. 705–711. ISSN: 1546-170X. DOI: 10.1038/s41591-020-0825-4. URL: http://www.nature.com/articles/s41591-020-0825-4.

[27] Gabriel G Perron, Michael Zasloff, and Graham Bell. "Experimental evolution of resistance to an antimicrobial peptide". In: *Proceedings of the Royal Society B: Biological Sciences* 273.1583 (Jan. 2006). Publisher: Royal Society, pp. 251–256. DOI: 10.1098/rspb.2005.3301. URL: http://royalsocietypublishing.org/doi/full/10.1098/rspb.2005.3301.

[28] Adam C. Palmer and Roy Kishony. "Understanding, predicting and manipulating the genotypic evolution of antibiotic resistance". en. In: *Nature Reviews Genetics* 14.4 (Apr. 2013). Number: 4 Publisher: Nature Publishing Group, pp. 243–248. ISSN: 1471-0064. DOI: 10.1038/nrg3351. URL: http://www.nature.com/articles/nrg3351.

[29] Miranda E Pitt et al. "Octapeptin C4 and polymyxin resistance occur via distinct pathways in an epidemic XDR Klebsiella pneumoniae ST258 isolate". In: *Journal of Antimicrobial Chemotherapy* 74.3 (Mar. 2019), pp. 582–593. ISSN: 0305-7453. DOI: 10.1093/jac/dky458. URL: https://doi.org/10.1093/jac/dky458.

[30] Kathryn E. Holt et al. "Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health". en. In: *PNAS* 112.27 (July 2015), E3574–E3581. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1501049112. URL: https://www.pnas.org/content/112/27/E3574.

[31] Robert A. Power, Julian Parkhill, and Tulio de Oliveira. "Microbial genome-wide association studies: lessons from human GWAS". en. In: *Nature Reviews Genetics* 18.1 (Jan. 2017), pp. 41–50. ISSN: 1471-0064. DOI: 10.1038/nrg.2016.132. URL: https://www.nature.com/articles/nrg.2016.132.

[32] Uzma Ansari et al. "Molecular Characterization of Carbapenem-Resistant Enterobacteriaceae in the USA, 2011–2015". en. In: *Open Forum Infect Dis* 4.suppl_1 (Oct. 2017), S179–S179. DOI: 10.1093/ofid/ofx163.328. URL: https://academic.oup.com/ofid/article/4/suppl_1/S179/4294266.

[33] *Tackling drug-resistant infections globally : final report and recommendations / the Review on Antimicrobial Resistance chaired by Jim O'Neill.* en. 2016. URL: https://wellcomecollection.org/works/thvwsuba.

[34] Centers for Disease Control and Prevention (U.S.) *Antibiotic resistance threats in the United States, 2019*. en. Tech. rep. Centers for Disease Control and Prevention (U.S.), Nov. 2019. DOI: 10.15620/cdc:82532. URL: https://stacks.cdc.gov/view/cdc/82532.

[35] *Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics*. Publisher: World Health Organization. 2017. URL: http://www.who.int/medicines/publications/global-priority-list-antibiotic-resistant-bacteria/en/.

[36] Latania K. Logan and Robert A. Weinstein. "The Epidemiology of Carbapenem-Resistant Enterobacteriaceae: The Impact and Evolution of a Global Menace". In: *J Infect Dis* 215.Suppl 1 (Feb. 2017), S28–S36. ISSN: 0022-1899. DOI: 10.1093/infdis/jiw282. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5853342/.

[37] Mitchell J. Schwaber et al. "Containment of a Country-wide Outbreak of Carbapenem-Resistant Klebsiella pneumoniae in Israeli Hospitals via a Nationally Implemented Intervention". en. In: *Clin Infect Dis.* 52.7 (Apr. 2011), pp. 848–855. ISSN: 1058-4838, 1537-6591. DOI: 10.1093/cid/cir025. URL: http://cid.oxfordjournals.org/content/52/7/848.

[38] Kate Russell Woodworth. "Vital Signs: Containment of Novel Multidrug-Resistant Organisms and Resistance Mechanisms — United States, 2006–2017". en-us. In: *MMWR Morb Mortal Wkly Rep* 67 (2018). ISSN: 0149-21951545-861X. DOI: 10.15585/mmwr.mm6713e1. URL: https://www.facebook.com/cdcmmwr.

[39] William E. Trick et al. "Electronic Public Health Registry of Extensively Drug-Resistant Organisms, Illinois, USA". In: *Emerg Infect Dis* 21.10 (Oct. 2015), pp. 1725–1732. ISSN: 1080-6040. DOI: 10.3201/eid2110.150538. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4593443/.

[40] Raveena D Singh et al. "The CDC SHIELD Orange County Project – Baseline Multi Drug-Resistant Organism (MDRO) Prevalence in a Southern California Region". In: *Open Forum Infect Dis* 4.Suppl 1 (Oct. 2017), S46–S47. ISSN: 2328-8957. DOI: 10.1093/ofid/ofx162.109. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5631751/.

[41] Shawn E Hawken and Evan S Snitkin. "Genomic epidemiology of multidrug-resistant Gram-negative organisms". In: *Ann N Y Acad Sci* 1435.1 (Jan. 2019), pp. 39–56. ISSN: 0077-8923. DOI: 10.1111/nyas.13672. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6167210/.

[42] Yonatan H. Grad and Marc Lipsitch. "Epidemiologic data and pathogen genome sequences: a powerful synergy for public health". eng. In: *Genome Biol* 15.11 (Nov. 2014), p. 538. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0538-4.

[43] Francesc Coll et al. "Longitudinal genomic surveillance of MRSA in the UK reveals transmission patterns in hospitals and the community". eng. In: *Sci Transl Med* 9.413 (Oct. 2017). ISSN: 1946-6242. DOI: 10.1126/scitranslmed.aak9745.

[44] Carolyn V. Gould, Richard Rothenberg, and James P. Steinberg. "Antibiotic resistance in long-term acute care hospitals: the perfect storm". eng. In: *Infect Control Hosp Epidemiol* 27.9 (Sept. 2006), pp. 920–925. ISSN: 0899-823X. DOI: 10.1086/507280.

[45] Michael Y. Lin et al. "The Importance of Long-term Acute Care Hospitals in the Regional Epidemiology of Klebsiella pneumoniae Carbapenemase–Producing Enterobacteriaceae". en. In: *Clin Infect Dis* 57.9 (Nov. 2013), pp. 1246–1252. ISSN: 1058-4838. DOI: 10.1093/cid/cit500. URL: https://academic.oup.com/cid/article/57/9/1246/488369.

[46] Jennifer H. Han et al. "Epidemiology of Carbapenem-Resistant Klebsiella pneumoniae in a Network of Long-Term Acute Care Hospitals". en. In: *Clin Infect Dis* 64.7 (Apr. 2017), pp. 839–844. ISSN: 1058-4838. DOI: 10.1093/cid/ciw856. URL: https://academic.oup.com/cid/article/64/7/839/2738659.

[47] Neil Woodford et al. "Outbreak of Klebsiella pneumoniae producing a new carbapenem-hydrolyzing class A beta-lactamase, KPC-3, in a New York Medical Center". eng. In: *Antimicrob Agents Chemother* 48.12 (Dec. 2004), pp. 4793–4799. ISSN: 0066-4804. DOI: 10.1128/AAC.48.12.4793-4799.2004.

[48] Tjibbe Donker, Jacco Wallinga, and Hajo Grundmann. "Patient Referral Patterns and the Spread of Hospital-Acquired Infections through National Health Care Networks". In: *PLoS Comput Biol* 6.3 (Mar. 2010). ISSN: 1553-734X. DOI: 10.1371/journal.pcbi.1000715. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2841613/.

[49] Damon J. A. Toth et al. "The Potential for Interventions in a Long-term Acute Care Hospital to Reduce Transmission of Carbapenem-Resistant Enterobacteriaceae in Affiliated Healthcare Facilities". eng. In: *Clin Infect Dis* 65.4 (Aug. 2017), pp. 581–587. ISSN: 1537-6591. DOI: 10.1093/cid/cix370.

[50] Mary K. Hayden et al. "Prevention of colonization and infection by Klebsiella pneumoniae carbapenemase-producing enterobacteriaceae in long-term acute-care hospitals". eng. In: *Clin Infect Dis* 60.8 (Apr. 2015), pp. 1153–1161. ISSN: 1537-6591. DOI: 10.1093/cid/ciu1173.

[51] Michael J. Ray et al. "Spread of Carbapenem-Resistant Enterobacteriaceae Among Illinois Healthcare Facilities: The Role of Patient Sharing". eng. In: *Clin Infect Dis* 63.7 (Oct. 2016), pp. 889–893. ISSN: 1537-6591. DOI: 10.1093/cid/ciw461.

[52] Bruce Y. Lee et al. "Tracking the spread of carbapenem-resistant Enterobacteriaceae (CRE) through clinical cultures alone underestimates the spread of CRE even more than anticipated". en. In: *Infection Control & Hospital Epidemiology* 40.6 (June 2019). Publisher: Cambridge University Press, pp. 731–734. ISSN: 0899-823X, 1559-6834. DOI: 10.1017/ice.2019.61. URL: http://www.cambridge.org/core/journals/infection-control-and-hospital-epidemiology/article/tracking-the-spread-of-carbapenemresistant-enterobacteriaceae-cre-through-clinical-cultures-alone-underestimates-the-spread-of-cre-even-more-than-anticipated/F1F4A7E472795D8BFCA63CBD1BA934A6/core-reader.

[53] Debby Ben-David et al. "Potential role of active surveillance in the control of a hospital-wide outbreak of carbapenem-resistant Klebsiella pneumoniae infection". eng. In: *Infect Control Hosp Epidemiol* 31.6 (June 2010), pp. 620–626. ISSN: 1559-6834. DOI: 10.1086/652528.

[54] Teppei Shimasaki et al. "Increased Relative Abundance of Klebsiella pneumoniae Carbapenemase-producing Klebsiella pneumoniae Within the Gut Microbiota Is Associated With Risk of Bloodstream Infection in Long-term Acute Care Hospital Patients". eng. In: *Clin Infect Dis* 68.12 (May 2019), pp. 2053–2059. ISSN: 1537-6591. DOI: 10.1093/cid/ciy796.

[55] Ian B. Jeffery, Denise B. Lynch, and Paul W. O'Toole. "Composition and temporal stability of the gut microbiota in older persons". eng. In: *ISME J* 10.1 (Jan. 2016), pp. 170–182. ISSN: 1751-7370. DOI: 10.1038/ismej.2015.88.

[56] Paul W. O'Toole and Ian B. Jeffery. "Gut microbiota and aging". eng. In: *Science* 350.6265 (Dec. 2015), pp. 1214–1215. ISSN: 1095-9203. DOI: 10.1126/science.aac8469.

[57] C. J. Donskey et al. "Effect of antibiotic therapy on the density of vancomycin-resistant enterococci in the stool of colonized patients". eng. In: *N Engl J Med* 343.26 (Dec. 2000), pp. 1925–1932. ISSN: 0028-4793. DOI: 10.1056/NEJM200012283432604.

[58] Sandra Reuter et al. "Building a genomic framework for prospective MRSA surveillance in the United Kingdom and the Republic of Ireland". eng. In: *Genome Res* 26.2 (Feb. 2016), pp. 263–270. ISSN: 1549-5469. DOI: 10.1101/gr.196709.115.

[59] Centers for Disease Control and Prevention. *Facility guidance for control of Carbapenem-resistant Enterobacteriaceae (CRE) : November 2015 update - CRE toolkit*. 2017. URL: https://stacks.cdc.gov/view/cdc/79104.

[60] K. L. Thong et al. "Simultaneous detection of methicillin-resistant Staphylococcus aureus, Acinetobacter baumannii, Escherichia coli, Klebsiella pneumoniae and Pseudomonas aeruginosa by multiplex PCR". eng. In: *Trop Biomed* 28.1 (Apr. 2011), pp. 21–31. ISSN: 2521-9855.

[61] *EUCAST: Clinical breakpoints and dosing of antibiotics*. 2017. URL: https://www.eucast.org/clinical_breakpoints/.

[62] Simon Andrews. *s-andrews/FastQC*. original-date: 2017-12-21T11:48:51Z. Mar. 2021. URL: https://github.com/s-andrews/FastQC.

[63] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data". In: *Bioinformatics* 30.15 (Aug. 2014), pp. 2114–2120. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu170. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103590/.

[64] Heng Li and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform". en. In: *Bioinformatics* 25.14 (July 2009). Publisher: Oxford Academic, pp. 1754–1760. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp324. URL: https://academic.oup.com/bioinformatics/article/25/14/1754/225615.

[65] *broadinstitute/picard*. original-date: 2014-03-28T20:43:35Z. Feb. 2021. URL: https://github.com/broadinstitute/picard.

[66] Heng Li et al. "The Sequence Alignment/Map format and SAMtools". en. In: *Bioinformatics* 25.16 (Aug. 2009), pp. 2078–2079. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp352. URL: https://academic.oup.com/bioinformatics/article/25/16/2078/204688.

[67] Aaron McKenna et al. "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data". eng. In: *Genome Res* 20.9 (Sept. 2010), pp. 1297–1303. ISSN: 1549-5469. DOI: 10.1101/gr.107524.110.

[68] Nicholas J. Croucher et al. "Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins". en. In: *Nucleic Acids Res* 43.3 (Feb. 2015), e15–e15. ISSN: 0305-1048. DOI: 10.1093/nar/gku1196. URL: https://academic.oup.com/nar/article/43/3/e15/2410982.

[69] Remco Bouckaert et al. "BEAST 2: A Software Platform for Bayesian Evolutionary Analysis". In: *PLOS Computational Biology* 10.4 (Apr. 2014), e1003537. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003537. URL: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003537.

[70] Alexandros Stamatakis. "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies". In: *Bioinformatics* 30.9 (May 2014), pp. 1312–1313. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu033. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3998144/.

[71] Andrew Rambaut et al. "Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7". eng. In: *Syst Biol* 67.5 (Sept. 2018), pp. 901–904. ISSN: 1076-836X. DOI: 10.1093/sysbio/syy032.

[72] Lam-Tung Nguyen et al. "IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies". In: *Mol Biol Evol* 32.1 (Jan. 2015), pp. 268–274. ISSN: 0737-4038. DOI: 10.1093/molbev/msu300. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4271533/.

[73] Bui Quang Minh, Minh Anh Thi Nguyen, and Arndt von Haeseler. "Ultrafast Approximation for Phylogenetic Bootstrap". en. In: *Mol Biol Evol* 30.5 (May 2013), pp. 1188–1195. ISSN: 0737-4038. DOI: 10.1093/molbev/mst024. URL: https://academic.oup.com/mbe/article/30/5/1188/997508.

[74] S. Kalyaanamoorthy et al. "ModelFinder: fast model selection for accurate phylogenetic estimates., ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates". eng. In: *Nat Methods* 14, 14.6, 6 (June 2017), pp. 587, 587–589. ISSN: 1548-7091. DOI: 10.1038/nmeth.4285,10.1038/nmeth.4285. URL: https://www.ncbi.nlm.nih.govhttp://europepmc.org/articles/PMC5453245/.

[75] Liam J. Revell. "phytools: an R package for phylogenetic comparative biology (and other things)". en. In: *Methods in Ecology and Evolution* 3.2 (2012), pp. 217–223. ISSN: 2041-210X. DOI: 10.1111/j.2041-210X.2011.00169.x. URL: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.2041-210X.2011.00169.x.

[76] Remco R. Bouckaert and Alexei J. Drummond. "bModelTest: Bayesian phylogenetic site model averaging and model comparison". In: *BMC Evolutionary Biology* 17.1 (Feb. 2017), p. 42. ISSN: 1471-2148. DOI: 10.1186/s12862-017-0890-6. URL: https://doi.org/10.1186/s12862-017-0890-6.

[77] Alexei J Drummond et al. "Relaxed Phylogenetics and Dating with Confidence". In: *PLoS Biol* 4.5 (May 2006). ISSN: 1544-9173. DOI: 10.1371/journal.pbio.0040088. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1395354/.

[78] Carter T. Butts. "**network** : A Package for Managing Relational Data in *R*". en. In: *J. Stat. Soft.* 24.2 (2008). ISSN: 1548-7660. DOI: 10.18637/jss.v024.i02. URL: http://www.jstatsoft.org/v24/i02/.

[79] Douglas Brent West. *Introduction to Graph Theory*. English. Subsequent edition. Upper Saddle River, N.J: Pearson College Div, Aug. 2000. ISBN: 978-0-13-014400-3.

[80] Gabor Csardi and Tamas Nepusz. "The Igraph Software Package for Complex Network Research". In: *InterJournal* Complex Systems (Nov. 2005), p. 1695.

[81] P. Legendre and Louis Legendre. *Numerical Ecology*. English. 3rd edition. Amsterdam: Elsevier, Aug. 2012. ISBN: 978-0-444-53868-0.

[82] Jari Oksanen et al. *vegan: Community Ecology Package*. Nov. 2020. URL: https://CRAN.R-project.org/package=vegan.

[83] Rodrigo Azuero Melo & & Demetrio Rodriguez T. & David Zarruk. *gmapsdistance: Distance and Travel Time Between Two Points from Google Maps*. Aug. 2018. URL: https://CRAN.R-project.org/package=gmapsdistance.

[84] World Health Organization. *Antimicrobial Resistance: Global Report on Surveillance*. en. Geneva: World Health Organization, 2014. ISBN: 978-92-4-156474-8.

[85] Kelly L. Wyres, Margaret M. C. Lam, and Kathryn E. Holt. "Population genomics of Klebsiella pneumoniae". en. In: *Nature Reviews Microbiology* (Feb. 2020), pp. 1–16. ISSN: 1740-1534. DOI: 10.1038/s41579-019-0315-1. URL: https://www.nature.com/articles/s41579-019-0315-1.

[86] Chang-Ro Lee et al. "Global Dissemination of Carbapenemase-Producing Klebsiella pneumoniae: Epidemiology, Genetic Context, Treatment Options, and Detection Methods". English. In: *Front. Microbiol.* 7 (2016). ISSN: 1664-302X. DOI: 10.3389/fmicb.2016.00895. URL: https://www.frontiersin.org/articles/10.3389/fmicb.2016.00895/full?utm_source=FWEB&utm_medium=NBLOG&utm_campaign=CIT-2018_FMICB_20180809.

[87] Angela Cano et al. "Risks of Infection and Mortality Among Patients Colonized With Klebsiella pneumoniae Carbapenemase–Producing K. pneumoniae: Validation of Scores and Proposal for Management". In: *Clinical Infectious Diseases* 66.8 (Apr. 2018), pp. 1204–1210. ISSN: 1058-4838. DOI: 10.1093/cid/cix991. URL: https://doi.org/10.1093/cid/cix991.

[88] Rebekah M. Martin et al. "Identification of Pathogenicity-Associated Loci in Klebsiella pneumoniae from Hospitalized Patients". en. In: *mSystems* 3.3 (June 2018). Publisher: American Society for Microbiology Journals Section: Research Article. ISSN: 2379-5077. DOI: 10.1128/mSystems.00015-18. URL: https://msystems.asm.org/content/3/3/e00015-18.

[89] "2020 NHSN Patient Safety Component Manual". en. In: (2020), p. 434.

[90] Kyle J. Popovich et al. "Genomic and Epidemiological Evidence for Community Origins of Hospital-Onset Methicillin-Resistant Staphylococcus aureus Bloodstream Infections". eng. In: *J. Infect. Dis.* 215.11 (2017), pp. 1640–1647. ISSN: 1537-6613. DOI: 10.1093/infdis/jiw647.

[91] Katie Saund et al. "prewas: data pre-processing for more informative bacterial GWAS". In: *Microbial Genomics,* 6.5 (2020). Publisher: Microbiology Society, e000368. ISSN: , DOI: 10.1099/mgen.0.000368. URL: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000368.

[92] Kat Holt. *katholt/Kleborate*. original-date: 2016-12-12T06:13:03Z. Apr. 2020. URL: https://github.com/katholt/Kleborate.

[93] Rainer Follador et al. "The diversity of Klebsiella pneumoniae surface polysaccharides". In: *Microb Genom* 2.8 (Aug. 2016). ISSN: 2057-5858. DOI: 10.1099/mgen.0.000073. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5320592/.

[94] Pin Liu et al. "Risk Factors for Carbapenem-Resistant Klebsiella pneumoniae Infection: A Meta-Analysis". In: *Microbial Drug Resistance* 24.2 (July 2017), pp. 190–198. ISSN: 1076-6294. DOI: 10.1089/mdr.2017.0061. URL: https://www.liebertpub.com/doi/full/10.1089/mdr.2017.0061.

[95] Sunita Shankar-Sinha et al. "The Klebsiella pneumoniae O Antigen Contributes to Bacteremia and Lethality during Murine Pneumonia". en. In: *Infection and Immunity* 72.3 (Mar. 2004). Publisher: American Society for Microbiology Journals Section: MOLECULAR PATHOGENESIS, pp. 1423–1430. ISSN: 0019-9567, 1098-5522. DOI: 10.1128/IAI.72.3.1423-1430.2004. URL: http://iai.asm.org/content/72/3/1423.

[96] Christine Tedijanto et al. "Estimating the proportion of bystander selection for antibiotic resistance among potentially pathogenic bacterial flora". en. In: *Proceedings of the National Academy of Sciences* 115.51 (Dec. 2018), E11988–E11995. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1810840115. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1810840115.

[97] Margaret M. C. Lam et al. "Genetic diversity, mobilisation and spread of the yersiniabactin-encoding mobile element ICEKp in Klebsiella pneumoniae populations". In: *Microbial Genomics,* 4.9 (2018), e000196. ISSN: , DOI: 10.1099/mgen.0.000196. URL: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000196.

[98] Kai Zhou et al. "Novel Subclone of Carbapenem-Resistant Klebsiella pneumoniae Sequence Type 11 with Enhanced Virulence and Transmissibility, China - Volume 26, Number 2—February 2020 - Emerging Infectious Diseases journal - CDC". en-us. In: (). DOI: 10.3201/eid2602.190594. URL: https://wwwnc.cdc.gov/eid/article/26/2/19-0594_article.

[99] Danxia Gu et al. "A fatal outbreak of ST11 carbapenem-resistant hypervirulent Klebsiella pneumoniae in a Chinese hospital: a molecular epidemiological study". English. In: *The Lancet Infectious Diseases* 18.1 (Jan. 2018), pp. 37–46. ISSN: 1473-3099, 1474-

4457. DOI: 10.1016/S1473-3099(17)30489-9. URL: http://www.thelancet.com/journals/laninf/article/PIIS1473-3099(17)30489-9/abstract.

[100] Katie Saund and Evan S. Snitkin. "Hogwash: three methods for genome-wide association studies in bacteria". In: *Microb Genom* 6.11 (Nov. 2020). ISSN: 2057-5858. DOI: 10.1099/mgen.0.000469. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7725327/.

[101] Stephen J. Chapman and Adrian V. S. Hill. "Human genetic susceptibility to infectious disease". en. In: *Nat Rev Genet* 13.3 (Mar. 2012). Number: 3 Publisher: Nature Publishing Group, pp. 175–188. ISSN: 1471-0064. DOI: 10.1038/nrg3114. URL: http://www.nature.com/articles/nrg3114.

[102] Andrew J. Page et al. "Roary: rapid large-scale prokaryote pan genome analysis". en. In: *Bioinformatics* 31.22 (Nov. 2015), pp. 3691–3693. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv421. URL: https://academic.oup.com/bioinformatics/article/31/22/3691/240757.

[103] Johannes Köster and Sven Rahmann. "Snakemake—a scalable bioinformatics workflow engine". en. In: *Bioinformatics* 28.19 (Oct. 2012). Publisher: Oxford Academic, pp. 2520–2522. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts480. URL: http://academic.oup.com/bioinformatics/article/28/19/2520/290322.

[104] Kelly L. Wyres et al. "Identification of Klebsiella capsule synthesis loci from whole genome data". In: *Microbial Genomics,* 2.12 (2016), e000102. ISSN: , DOI: 10.1099/mgen.0.000102. URL: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000102.

[105] Ryan R. Wick et al. "Kaptive Web: User-Friendly Capsule and Lipopolysaccharide Serotype Prediction for Klebsiella Genomes". en. In: *Journal of Clinical Microbiology* 56.6 (June 2018). ISSN: 0095-1137, 1098-660X. DOI: 10.1128/JCM.00197-18. URL: https://jcm.asm.org/content/56/6/e00197-18.

[106] Pablo Cingolani et al. "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff". In: *Fly* 6.2 (Apr. 2012), pp. 80–92. ISSN: 1933-6934. DOI: 10.4161/fly.19695. URL: https://doi.org/10.4161/fly.19695.

[107] Panisa Treepong et al. "panISa: ab initio detection of insertion sequences in bacterial genomes from short read sequence data". en. In: *Bioinformatics* 34.22 (Nov. 2018), pp. 3795–3800. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty479. URL: https://academic.oup.com/bioinformatics/article/34/22/3795/5040324.

[108] Begüm D. Topçuoğlu et al. "A Framework for Effective Application of Machine Learning to Microbiome-Based Classification Problems". en. In: *mBio* 11.3 (June 2020).

Publisher: American Society for Microbiology Section: Research Article. ISSN: 2150-7511. DOI: 10.1128/mBio.00434-20. URL: https://mbio.asm.org/content/11/3/e00434-20.

[109] Max Kuhn. "Building Predictive Models in R Using the caret Package". en. In: *Journal of Statistical Software* 28.1 (Nov. 2008). Number: 1, pp. 1–26. ISSN: 1548-7660. DOI: 10.18637/jss.v028.i05. URL: https://www.jstatsoft.org/index.php/jss/article/view/v028i05.

[110] *R: The R Project for Statistical Computing*. URL: https://www.r-project.org/.

[111] Begüm Topçuoğlu et al. *mikropml: User-Friendly R Package for Supervised Machine Learning Pipelines*. Dec. 2020. URL: https://CRAN.R-project.org/package=mikropml.

[112] Christiam Camacho et al. "BLAST+: architecture and applications". In: *BMC Bioinformatics* 10.1 (Dec. 2009), p. 421. ISSN: 1471-2105. DOI: 10.1186/1471-2105-10-421. URL: https://doi.org/10.1186/1471-2105-10-421.

[113] Alice R. Wattam et al. "Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center". eng. In: *Nucleic Acids Res.* 45.D1 (2017), pp. D535–D542. ISSN: 1362-4962. DOI: 10.1093/nar/gkw1017.

[114] Hadley Wickham et al. "Welcome to the Tidyverse". en. In: *Journal of Open Source Software* 4.43 (Nov. 2019), p. 1686. ISSN: 2475-9066. DOI: 10.21105/joss.01686. URL: https://joss.theoj.org/papers/10.21105/joss.01686.

[115] Claus O. Wilke. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. July 2019. URL: https://CRAN.R-project.org/package=cowplot.

[116] Guangchuang Yu et al. "ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data". en. In: *Methods in Ecology and Evolution* 8.1 (2017), pp. 28–36. ISSN: 2041-210X. DOI: 10.1111/2041-210X.12628. URL: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12628.

[117] Guangchuang Yu et al. "Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree". In: *Mol Biol Evol* 35.12 (Dec. 2018), pp. 3041–3043. ISSN: 0737-4038. DOI: 10.1093/molbev/msy194. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6278858/.

[118] Emmanuel Paradis and Klaus Schliep. "ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R". en. In: *Bioinformatics* 35.3 (Feb. 2019). Publisher: Oxford Academic, pp. 526–528. ISSN: 1367-4803. DOI: 10.1093/bioinforma

tics/bty633. URL: https://academic.oup.com/bioinformatics/article/35/3/526/5055127.

[119] Sean Conlan et al. "Plasmid Dynamics in KPC-Positive Klebsiella pneumoniae during Long-Term Patient Colonization". en. In: *mBio* 7.3 (July 2016). ISSN: 2150-7511. DOI: 10.1128/mBio.00742-16. URL: https://mbio.asm.org/content/7/3/e00742-16.

[120] Deanna N. Schreiber-Gregory, Jennifer Waller, and Tyler Smith. "Ridge Regression and multicollinearity: An in-depth review". In: *Model Assisted Statistics & Applications* 13.4 (Oct. 2018), pp. 359–365. ISSN: 15741699. DOI: 10.3233/MAS-180446. URL: http://proxy.lib.umich.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=aci&AN=132899398&site=ehost-live&scope=site.

[121] Begüm D. Topçuoğlu et al. "Effective application of machine learning to microbiome-based classification problems". en. In: *bioRxiv* (Oct. 2019), p. 816090. DOI: 10.1101/816090. URL: https://www.biorxiv.org/content/10.1101/816090v1.

[122] R. Craig MacLean and Alvaro San Millan. "The evolution of antibiotic resistance". en. In: *Science* 365.6458 (Sept. 2019). Publisher: American Association for the Advancement of Science Section: Perspective, pp. 1082–1083. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aax3879. URL: http://science.sciencemag.org/content/365/6458/1082.

[123] World Health Organization, ed. *Antimicrobial resistance: global report on surveillance*. en. OCLC: ocn880847527. Geneva, Switzerland: World Health Organization, 2014. ISBN: 978-92-4-156474-8.

[124] Johanna Björkman and Dan I. Andersson. "The cost of antibiotic resistance from a bacterial perspective". eng. In: *Drug Resist Updat* 3.4 (Aug. 2000), pp. 237–245. ISSN: 1532-2084. DOI: 10.1054/drup.2000.0147.

[125] Anita H Melnyk, Alex Wong, and Rees Kassen. "The fitness costs of antibiotic resistance mutations". In: *Evol Appl* 8.3 (Mar. 2015), pp. 273–283. ISSN: 1752-4571. DOI: 10.1111/eva.12196. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4380921/.

[126] T. Vogwill, M. Kojadinovic, and R. C. MacLean. "Epistasis between antibiotic resistance mutations and genetic background shape the fitness effect of resistance across species of Pseudomonas". In: *Proceedings of the Royal Society B: Biological Sciences* 283.1830 (May 2016). Publisher: Royal Society, p. 20160151. DOI: 10.1098/rspb.2016.0151. URL: http://royalsocietypublishing.org/doi/full/10.1098/rspb.2016.0151.

[127] Tom Vogwill et al. "Testing the Role of Genetic Background in Parallel Evolution Using the Comparative Experimental Evolution of Antibiotic Resistance". In: *Molec-*

ular Biology and Evolution 31.12 (Dec. 2014), pp. 3314–3323. ISSN: 0737-4038. DOI: 10.1093/molbev/msu262. URL: https://doi.org/10.1093/molbev/msu262.

[128] Laurent Poirel, Aurélie Jayol, and Patrice Nordmann. "Polymyxins: Antibacterial Activity, Susceptibility Testing, and Resistance Mechanisms Encoded by Plasmids or Chromosomes". en. In: *Clin. Microbiol. Rev.* 30.2 (Apr. 2017), pp. 557–596. ISSN: 0893-8512, 1098-6618. DOI: 10.1128/CMR.00064-16. URL: http://cmr.asm.org/content/30/2/557.

[129] Nenad Macesic et al. "Emergence of Polymyxin Resistance in Clinical Klebsiella pneumoniae Through Diverse Genetic Adaptations: A Genomic, Retrospective Cohort Study". In: *Clinical Infectious Diseases* 70.10 (May 2020), pp. 2084–2091. ISSN: 1058-4838. DOI: 10.1093/cid/ciz623. URL: https://doi.org/10.1093/cid/ciz623.

[130] Robert L. Skov and Dominique L. Monnet. "Plasmid-mediated colistin resistance (mcr-1 gene): three months later, the story unfolds". en. In: *Eurosurveillance* 21.9 (Mar. 2016). Publisher: European Centre for Disease Prevention and Control, p. 30155. ISSN: 1560-7917. DOI: 10.2807/1560-7917.ES.2016.21.9.30155. URL: https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2016.21.9.30155?crawler=true.

[131] Jean-Philippe Rasigade et al. "Strain-specific estimation of epidemic success provides insights into the transmission dynamics of tuberculosis". en. In: *Scientific Reports* 7.1 (Mar. 2017). Number: 1 Publisher: Nature Publishing Group, p. 45326. ISSN: 2045-2322. DOI: 10.1038/srep45326. URL: http://www.nature.com/articles/srep45326.

[132] Evan S. Snitkin et al. "Genomic insights into the fate of colistin resistance and Acinetobacter baumannii during patient treatment". In: *Genome Res* 23.7 (July 2013), pp. 1155–1162. ISSN: 1088-9051. DOI: 10.1101/gr.154328.112. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3698508/.

[133] Zena Lapp et al. "Patient and Microbial Genomic Factors Associated with Carbapenem-Resistant Klebsiella pneumoniae Extraintestinal Colonization and Infection". en. In: *mSystems* 6.2 (Apr. 2021). Publisher: American Society for Microbiology Journals Section: Research Article. ISSN: 2379-5077. DOI: 10.1128/mSystems.00177-21. URL: https://msystems.asm.org/content/6/2/e00177-21.

[134] Selvi C. Ersoy et al. "Correcting a Fundamental Flaw in the Paradigm for Antimicrobial Susceptibility Testing". en. In: *EBioMedicine* 20 (June 2017), pp. 173–181. ISSN: 2352-3964. DOI: 10.1016/j.ebiom.2017.05.026. URL: https://www.sciencedirect.com/science/article/pii/S2352396417302244.

[135] Michael J. Bottery, Jonathan W. Pitchford, and Ville-Petri Friman. "Ecology and evolution of antimicrobial resistance in bacterial communities". en. In: *The ISME Journal* 15.4 (Apr. 2021). Number: 4 Publisher: Nature Publishing Group, pp. 939–

948. ISSN: 1751-7370. DOI: 10.1038/s41396-020-00832-7. URL: http://www.nature.com/articles/s41396-020-00832-7.

[136] Melvin P Weinstein. *M100-performance standards for antimicrobial susceptibility testing, 30th edition*. 2020. URL: http://em100.edaptivedocs.net/GetDoc.aspx?doc=CLSI%20M100%20ED30:2020&sbssok=CLSI%20M100%20ED30:2020%20TABLE%202A&format=HTML&hl=colistin%20klebsiella.

[137] Martin Hunt et al. "ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads". In: *Microb Genom* 3.10 (Sept. 2017). ISSN: 2057-5858. DOI: 10.1099/mgen.0.000131. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5695208/.

[138] Ryan Poplin et al. "Scaling accurate genetic variant discovery to tens of thousands of samples". en. In: *bioRxiv* (July 2018). Publisher: Cold Spring Harbor Laboratory Section: New Results, p. 201178. DOI: 10.1101/201178. URL: https://www.biorxiv.org/content/10.1101/201178v3.

[139] David Arndt et al. "PHASTER: a better, faster version of the PHAST phage search tool". In: *Nucleic Acids Res* 44.Web Server issue (July 2016), W16–W21. ISSN: 0305-1048. DOI: 10.1093/nar/gkw387. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4987931/.

[140] Stefan Kurtz et al. "Versatile and open software for comparing large genomes". In: *Genome Biology* 5.2 (Jan. 2004), R12. ISSN: 1474-760X. DOI: 10.1186/gb-2004-5-2-r12. URL: https://doi.org/10.1186/gb-2004-5-2-r12.

[141] Björn Berglund. "Acquired Resistance to Colistin via Chromosomal And Plasmid-Mediated Mechanisms in Klebsiella pneumoniae". en-US. In: *Infectious Microbes & Diseases* 1.1 (Sept. 2019), pp. 10–19. ISSN: 2641-5917. DOI: 10.1097/IM9.0000000000000002. URL: http://journals.lww.com/imd/fulltext/2019/09000/acquired_resistance_to_colistin_via_chromosomal.3.aspx.

[142] Abiola O. Olaitan, Serge Morand, and Jean-Marc Rolain. "Mechanisms of polymyxin resistance: acquired and intrinsic resistance in bacteria". English. In: *Front. Microbiol.* 5 (2014). ISSN: 1664-302X. DOI: 10.3389/fmicb.2014.00643. URL: https://www.frontiersin.org/articles/10.3389/fmicb.2014.00643/full.

[143] Aurélie Jayol et al. "High-Level Resistance to Colistin Mediated by Various Mutations in the crrB Gene among Carbapenemase-Producing Klebsiella pneumoniae". In: *Antimicrob Agents Chemother* 61.11 (Oct. 2017). ISSN: 0066-4804. DOI: 10.1128/AAC.01423-17. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5655078/.

[144] Yi-Hsiang Cheng et al. "Amino Acid Substitutions of CrrB Responsible for Resistance to Colistin through CrrC in Klebsiella pneumoniae". en. In: *Antimicrobial Agents*

and Chemotherapy 60.6 (June 2016). Publisher: American Society for Microbiology Journals Section: Mechanisms of Resistance, pp. 3709–3716. ISSN: 0066-4804, 1098-6596. DOI: 10.1128/AAC.00009-16. URL: http://aac.asm.org/content/60/6/3709.

[145] Susana Matamouros, Kyle R. Hager, and Samuel I. Miller. "HAMP Domain Rotation and Tilting Movements Associated with Signal Transduction in the PhoQ Sensor Kinase". en. In: *mBio* 6.3 (July 2015). Publisher: American Society for Microbiology Section: Research Article. ISSN: 2150-7511. DOI: 10.1128/mBio.00616-15. URL: https://mbio.asm.org/content/6/3/e00616-15.

[146] Srujana S. Yadavalli et al. "Functional Determinants of a Small Protein Controlling a Broadly Conserved Bacterial Sensor Kinase". en. In: *Journal of Bacteriology* 202.16 (July 2020). Publisher: American Society for Microbiology Journals Section: Research Article. ISSN: 0021-9193, 1098-5530. DOI: 10.1128/JB.00305-20. URL: http://jb.asm.org/content/202/16/e00305-20.

[147] Miranda E. Pitt et al. "Multifactorial chromosomal variants regulate polymyxin resistance in extensively drug-resistant Klebsiella pneumoniae". In: *Microb Genom* 4.3 (Feb. 2018). ISSN: 2057-5858. DOI: 10.1099/mgen.0.000158. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5885010/.

[148] Abiola Olumuyiwa Olaitan et al. "Worldwide emergence of colistin resistance in Klebsiella pneumoniae from healthy humans and patients in Lao PDR, Thailand, Israel, Nigeria and France owing to inactivation of the PhoP/PhoQ regulator mgrB: an epidemiological and molecular study". en. In: *International Journal of Antimicrobial Agents* 44.6 (Dec. 2014), pp. 500–507. ISSN: 0924-8579. DOI: 10.1016/j.ijantimicag.2014.07.020. URL: https://www.sciencedirect.com/science/article/pii/S0924857914002581.

[149] Caitlin Collins and Xavier Didelot. "A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination". en. In: *PLOS Computational Biology* 14.2 (Feb. 2018), e1005958. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005958. URL: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005958.

[150] John A Lees et al. "pyseer: a comprehensive tool for microbial pangenome-wide association studies". In: *Bioinformatics* 34.24 (Dec. 2018), pp. 4310–4312. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty539. URL: https://doi.org/10.1093/bioinformatics/bty539.

[151] Michael P. Fay. "Confidence intervals that match Fisher's exact or Blaker's exact tests". eng. In: *Biostatistics* 11.2 (Apr. 2010), pp. 373–374. ISSN: 1468-4357. DOI: 10.1093/biostatistics/kxp050.

[152] Raivo Kolde. *pheatmap: Pretty Heatmaps*. Jan. 2019. URL: https://CRAN.R-project.org/package=pheatmap.

[153] Guangchuang Yu. *ggplotify: Convert Plot to 'grob' or 'ggplot' Object*. Mar. 2020. URL: https://CRAN.R-project.org/package=ggplotify.

[154] Kerri A. Thom et al. "Factors leading to transmission risk of Acinetobacter baumannii". In: *Crit Care Med* 45.7 (July 2017), e633–e639. ISSN: 0090-3493. DOI: 10.1097/CCM.0000000000002318. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5474153/.

[155] Larissa M. Pisney et al. "Carbapenem-Resistant Enterobacteriaceae Rectal Screening during an Outbreak of New Delhi Metallo--Lactamase–Producing Klebsiella pneumoniae at an Acute Care Hospital". en. In: *Infection Control & Hospital Epidemiology* 35.4 (Apr. 2014). Publisher: Cambridge University Press, pp. 434–436. ISSN: 0899-823X, 1559-6834. DOI: 10.1086/675597. URL: http://www.cambridge.org/core/journals/infection-control-and-hospital-epidemiology/article/carbapenemresistant-enterobacteriaceae-rectal-screening-during-an-outbreak-of-new-delhi-metallolactamaseproducing-klebsiella-pneumoniae-at-an-acute-care-hospital/78B9E7550DF9FC39CE82ACA991F4A209.

[156] James A. McKinnell et al. "High Prevalence of Multidrug-Resistant Organism Colonization in 28 Nursing Homes: An "Iceberg Effect"". en. In: *Journal of the American Medical Directors Association* 21.12 (Dec. 2020), 1937–1943.e2. ISSN: 1525-8610. DOI: 10.1016/j.jamda.2020.04.007. URL: https://www.sciencedirect.com/science/article/pii/S1525861020303261.

[157] Claire L. Gorrie et al. "Gastrointestinal Carriage Is a Major Reservoir of Klebsiella pneumoniae Infection in Intensive Care Patients". In: *Clinical Infectious Diseases* 65.2 (July 2017), pp. 208–215. ISSN: 1058-4838. DOI: 10.1093/cid/cix270. URL: https://doi.org/10.1093/cid/cix270.

[158] Xavier Didelot et al. "Within-host evolution of bacterial pathogens". In: *Nat Rev Microbiol* 14.3 (Mar. 2016), pp. 150–162. ISSN: 1740-1526. DOI: 10.1038/nrmicro.2015.13. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5053366/.

[159] Jessica Tischendorf, Rafael Almeida de Avila, and Nasia Safdar. "Risk of infection following colonization with carbapenem-resistant Enterobactericeae: A systematic review". en. In: *American Journal of Infection Control* 44.5 (May 2016), pp. 539–543. ISSN: 01966553. DOI: 10.1016/j.ajic.2015.12.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S0196655315012353.