

Washington University in St. Louis

Washington University Open Scholarship

Arts & Sciences Electronic Theses and
Dissertations

Arts & Sciences

Summer 8-15-2020

Genomic analysis of diverse bacterial pathogens

Robert Potter

Washington University in St. Louis

Follow this and additional works at: https://openscholarship.wustl.edu/art_sci_etds



Part of the [Microbiology Commons](#)

Recommended Citation

Potter, Robert, "Genomic analysis of diverse bacterial pathogens" (2020). *Arts & Sciences Electronic Theses and Dissertations*. 2336.

https://openscholarship.wustl.edu/art_sci_etds/2336

This Dissertation is brought to you for free and open access by the Arts & Sciences at Washington University Open Scholarship. It has been accepted for inclusion in Arts & Sciences Electronic Theses and Dissertations by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.

WASHINGTON UNIVERSITY IN ST. LOUIS
Division of Biology and Biomedical Sciences
Molecular Microbiology & Microbial Pathogenesis

Dissertation Examination Committee:

Gautam Dantas, Chair

Megan Baldrige

Carey-Ann Burnham

Mario Feldman

Jim Fleckenstein

Stephanie Fritz

Andy Kau

Genomic Analysis of Diverse Bacterial Pathogens

By

Robert F. Potter

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2020

St. Louis, Missouri

© 2020, Robert Potter

Table of Contents

List of Figures.....	vi
List of Tables.....	viii
Acknowledgments	ix
Abstract.....	x
Chapter 1: Introduction.....	1
Chapter 2: <i>bla</i> _{IMP-27} on transferable plasmids in <i>Proteus mirabilis</i> and <i>Providencia rettgeri</i>	6
2.1 Abstract.....	6
2.2 Introduction.....	7
2.3 Results.....	7
2.3.1 Southern blot confirmation of <i>bla</i> _{IMP-27} in transconjugants.....	7
2.3.2 Conjugation effects on phenotypic resistance in <i>E. coli</i> J53.....	8
2.3.3 Complete <i>bla</i> _{IMP-27} positive plasmids sequence.....	9
2.4 Discussion.....	12
2.5 Materials and Methods.....	13
2.5.1 Bacterial isolates and culturing.....	13
2.5.2 Broth Conjugation.....	14
2.5.3 Susceptibility testing.....	14
2.5.4 Southern blot.....	14
2.5.5 Plasmid assembly and annotation.....	15
2.6 Acknowledgments.....	15
2.7 References.....	16
Chapter 3: Population Structure, Antibiotic Resistance, and Uropathogenicity of <i>Klebsiella variicola</i>	20
3.1 Abstract.....	20
3.2 Introduction.....	21
3.3 Results.....	22
3.3.1 Average nucleotide identity can delineate <i>Klebsiella variicola</i> from related species.....	22
3.3.2 <i>Klebsiella variicola</i> is composed of two distantly related lineages.....	25
3.3.3 Acquired ARGs and VGs are not restricted to any <i>K. variicola</i> cluster.....	28
3.3.4 WUSM <i>K. variicola</i> cohort are susceptible to most antibiotics.....	30
3.3.5 Changes in <i>fim</i> operon are associated with uropathogenicity in a murine UTI model.....	33
3.3.6 <i>K. variicola</i> contains both conserved and novel usher genes.....	36
3.4 Discussion.....	39
3.5 Materials and Methods.....	45
3.5.1 Clinical <i>Klebsiella</i> collection.....	45

3.5.2 Illumina whole-genome sequencing and publicly available <i>Klebsiella</i> genomes.....	46
3.5.3 Antimicrobial susceptibility testing.....	47
3.5.4 Mouse Urinary tract infection model.....	48
3.5.5 Phase assays.....	49
3.5.6 FimA and GroEL immunoblots.....	49
3.5.7 Statistics.....	49
3.6 Acknowledgments.....	50
3.7 References.....	51
Chapter 4: Spatiotemporal dynamics of multidrug resistant bacteria on intensive care unit surfaces.....	63
4.1 Abstract.....	63
4.2 Introduction.....	63
4.3 Results.....	67
4.3.1 PAK-H ICU surfaces had high bacterial burden.....	67
4.3.2 Sequence based bacterial identification outperformed MALDI-TOF MS.....	68
4.3.3 Single lineages dominated <i>A. baumannii</i> and <i>E. faecium</i> populations.....	73
4.3.4 Spatiotemporal distance identifies relevant epidemiologic groups.....	76
4.3.5 PAK-H isolates have high genotypic and phenotypic resistance.....	78
4.3.6 ARGs against almost all antimicrobials are shared between species.....	81
4.3.7 <i>A. baumannii</i> and <i>E. faecium</i> have synergistic biofilm interactions.....	84
4.4 Discussion.....	87
4.5 Materials and Methods.....	94
4.5.1 Sample collection and culturing.....	94
4.5.2 Antibiotic susceptibility testing.....	95
4.5.3 Illumina Whole Genome Sequencing.....	96
4.5.4 Taxonomic assignment.....	97
4.5.5 Core genome alignment.....	97
4.5.6 Clonality analysis.....	98
4.5.7 Calculate temporal and spatial distances for variant cliques.....	98
4.5.8 ARG identification.....	99
4.5.9 <i>bla</i> _{NDM-1} loci annotation and comparison.....	99
4.5.10 <i>A. baumannii</i> and <i>E. faecium</i> co-association permutation testing.....	100
4.5.11 Biofilm assays.....	101
4.5.12 Statistics.....	102
4.5.13 Data availability.....	103
4.6 Acknowledgments.....	103
4.7 References.....	103

Chapter 5: In Silico Analysis of <i>Gardnerella</i> Genomespecies Detected in the Setting of Bacterial Vaginosis.....	124
5.1 Abstract.....	124
5.2 Introduction.....	125
5.3 Results.....	126
5.3.1 <i>In silico</i> tool-dependent classification of <i>G. vaginalis</i> into eight to fourteen genomespecies.....	126
5.3.2 Core-genome alignment support relatedness of the genomespecies into 8 clades.....	129
5.3.3 <i>Gardnerella</i> genomespecies have distinct accessory gene repertoires.....	132
5.3.4 Taxonomic signatures of novel genomespecies during BV.....	135
5.3.5 Expression of translation machinery and putative virulence factors by <i>Gardnerella</i> during BV.....	138
5.4 Discussion.....	140
5.5 Materials and Methods.....	144
5.5.1 Publicly available genomes and reads.....	144
5.5.2 <i>In silico</i> taxonomic analysis.....	145
5.5.3 Core-genome analysis.....	145
5.5.4 Accessory genome analysis.....	146
5.5.5 Cluster of orthologous groups (COGs) and gene of interest quantification.....	146
5.5.6 Taxonomic Metatranscriptome analysis.....	146
5.5.7 Metatranscriptome functional analysis.....	147
5.5.8 Statistical Analysis.....	148
5.6 Acknowledgments.....	148
5.7 References.....	149
Chapter 6: Phenotypic and genotypic characterization of linezolid-resistant <i>Enterococcus faecium</i> from the USA and Pakistan.....	158
6.1 Abstract.....	158
6.2 Introduction.....	158
6.3 Results.....	160
6.3.1 Acquired linezolid resistance genes (<i>optrA</i> , <i>poxxA</i> and <i>cfr</i> -like) were found exclusively in the <i>E. faecium</i> isolates recovered from Pakistan, regardless of clade.....	160
6.3.2 Linezolid resistance differs by genes present, not by mechanism.....	164
6.3.3 Different genetic platforms of <i>optrA</i> in linezolid-resistant <i>E. faecium</i> from Pakistan.....	165
6.4 Discussion.....	167

6.5 Materials and Methods.....	171
6.5.1 Linezolid-non-susceptible <i>E. faecium</i> cohort.....	171
6.5.2 Illumina WGS and genomic analysis.....	172
6.5.3 Antibiotic susceptibility testing.....	173
6.5.4 <i>In silico</i> Oxazolidinone resistance determinant identification.....	174
6.5.5 Data availability	175
6.6 Acknowledgments.....	175
Chapter 7: Pleiotropic effects of <i>pgsA2</i> mediated daptomycin resistance in <i>Corynebacterium</i> .	185
7.1 Abstract.....	185
7.2 Introduction.....	186
7.3 Results.....	187
7.3.1 <i>In silico</i> species identification.....	188
7.3.2 Resistant mutation mapping.....	189
7.3.3 BiOLOG Chemical Sensitivity Screen.....	190
7.3.4 Proteomic identification of impaired nitrate reductase levels and anaerobic growth assessment.....	191
7.4 Discussion.....	195
7.5 Materials and Methods.....	197
7.5.1 Clinical and computational cohort.....	197
7.5.2 Proteomic characterization.....	198
7.5.3 BioLOG chemical sensitivity assay.....	199
7.5.4 Anaerobic growth.....	199
7.6 Acknowledgments.....	199
7.7 References.....	200
Chapter 8: General Conclusion.....	204

List of Figures

Figure 2.3.1 Annotated plasmid diagram from DNAPlotter of pPM187 and pPR1.....	8
Figure 3.3.1 Pairwise Average Nucleotide Identity Clustermap of WUSM and NCBI <i>Klebsiella</i>	24
Figure 3.3.2 Population structure of <i>K. variicola</i> genomes.....	26
Figure 3.3.3 Distribution of acquired antibiotic resistance and virulence genes in the <i>K. variicola</i> cohort.	29
Figure 3.3.4 WUSM <i>K. variicola</i> strains have a low burden of ARGs and are generally susceptible to antibiotic.....	33
Figure 3.3.5 Changes in <i>fim</i> operon are associated with outcomes in mouse UTI model.....	35
Figure 3.3.6 <i>K. variicola</i> carries both conserved and novel usher genes.....	38
Figure 4.3.1 Bacterial isolate taxonomic identification and location.....	67
Figure 4.3.2 MALDI-TOF Identification and distribution.....	69
Figure 4.3.3 Phylogenetic trees of high abundance species from core genome alignments.....	72
Figure 4.3.4 Relationship of core genome SNP groups to spatial and temporal distance.....	74
Figure 4.3.5 Genotypic antibiotic resistance in major species.....	79
Figure 4.3.6 Shared antibiotic resistance genes across diverse taxonomic groups.....	82
Figure 4.3.7 Synergistic biofilm interactions for <i>A. baumannii</i> and <i>E. faecium</i> predicted by surface collections.....	85
Figure 5.3.1 Different in silico taxonomic tools produce 8 to 14 <i>Gardnerella</i> genomospecies..	127
Figure 5.3.2 Core genome phylogenetic analysis shows the genomospecies fall into 9 distinct clusters.....	131
Figure 5.3.3 Accessory gene burden is different between the major genomospecies.....	133
Figure 5.3.4 Newly elucidated genomospecies are identified in BV metatranscriptome samples.....	136
Figure 5.3.5 <i>Gardnerella</i> translation machinery and vaginolysin expression during BV.....	139
Figure 6.3.1 Recombination-free phylogenetic tree including MLST, country, source, resistance, resistance gene and mutation data. Linezolid resistance in US isolates was attributed solely to the G2576T mutation of the 23S rRNA gene sequence.....	161
Figure 6.3.2 Linezolid and tedizolid MICs and comparisons by basis of resistance mechanism.	164
Figure 6.3.3 Genetic context of <i>optrA</i> in isolates that harbor <i>optrA</i> , <i>cfr</i> -like and <i>poxA</i> genes.....	166

Figure 7.3.1 ANI heatmap for entire cohort.....	188
Figure 7.3.2 PCA of peptide fragments from proteomics.....	192
Figure 7.3.3 Volcano plot of differential abundant proteins.	193
Figure 7.3.4 Four quadrant streak of PS (a), PR (b), and IR (c) under anerobic conditions after 96 hours.....	196

List of Tables

Table 2.3.1 Phenotypic resistance of <i>bla</i> _{IMP-27} positive isolates, <i>E. coli</i> J53, and transconjugants..	9
Table 7.3.1 SNP analysis of susceptible-resistant pairs.	187
Table 7.3.2 Structure and description of top BiOLOG hits that had differential activity against daptomycin resistant <i>C. striatum</i> compared to susceptible.....	190
Table 7.3.3 Proteins that are commonly downregulated in PR and IR when compared against PS.....	192
Table 7.3.4 Proteins that are commonly upregulated in PR and IR when compared against PS.....	192

Acknowledgments

I would like to thank my lab-mates, collaborators, mentors, friends, and family.

Robert Potter

Washington University in St. Louis

August 2020

ABSTRACT OF THE DISSERTATION

Genomic Analysis of Diverse Bacterial Pathogens

for Arts & Sciences Graduate Students

by

Robert Potter

Doctor of Philosophy in Biology and Biomedical Sciences

Molecular Microbiology & Microbial Pathogenesis

Washington University in St. Louis, 2020

Professor Gautam Dantas, Chair

Bacterial pathogens have been a historical scourge for the entirety of human existence but have been significantly thwarted since the 20th century due to the development of antibiotics. However, owing to the large selection pressure of antibiotics on bacterial populations, phenotypic antibiotic resistance from the development of vertically transmitted mutations and horizontally acquired antibiotic resistance genes (ARGs) is increasing. The sum has produced multidrug resistant organisms (MDROs) which have extremely limited treatment options. Epidemiological studies have determined that carbapenem resistant *Enterobacteriaceae* (CRE), *Acinetobacter baumannii*, and vancomycin resistant *Enterococcus* (VRE) are some of the most problematic MDRO infections.

The advent of cost-effective and accurate next-generation sequencing has resulted in a proliferation of bacterial genomes available. ARGs, antibiotic resistance conferring single nucleotide polymorphism (SNPs), and virulence genes can be identified within an assembled genome by comparison to known databases. The

combination of the genetic information encoded within the genome of an isolate along with metadata related to important phenotypes or clinical context can be used to identify trends in ARG carriage, evolution over time, and differences in gene burden. This information can also be used in understanding the effects of antibiotic treatment on multi-organism infections such as bacterial vaginosis.

My thesis intends to investigate features related to natural populations of bacterial isolates in the *Enterobacteriaceae* family and *Acinetobacter baumannii* in Chapters 2, 3, 4 and the Gram-positive organisms *Enterococcus faecium*, *Gardnerella*, and *Corynebacterium* in Chapters 5, 6, and 7.

In Chapter 2 we identify the carbapenem resistance gene *bla*_{IMP-27} in a clinical isolate of carbapenem resistance *Providencia rettgeri*. We then acquired two *bla*_{IMP-27} bearing *Proteus mirabilis* and determine that one isolate (PM187) also has it on a plasmid. We were able to completely close the *bla*_{IMP-27} bearing plasmids pPR1 and pPM187 and determine that the local genetic context was similar but the background of the plasmids were different. In Chapter 3 we collect a cohort of longitudinally antibiotic resistant organisms recovered from hospital surfaces in the United States and Pakistan. We compare the phenotypic identification with the genomic identification to determine that several isolates represent novel taxonomic groups, we identify a severe degree of phenotypic antibiotic resistance in the collected important human pathogens and elucidate a network of ARGs common amongst the bacteria. Importantly, we demonstrate that *E. faecium* and *A. baumannii* co-occur greater than predicted by chance alone and that laboratory strains of these organisms are capable of forming synergistic growth in biofilms. In Chapter 4 we collect a cohort of *Klebsiella variicola*

from Washington University and use whole-genome sequencing to determine the population structure of all publicly available *K. variicola* genomes and identify genes relevant for infection related phenotypes. We show that these differences may have a functional consequence as some *K. variicola* strains can be more competent uropathogens than *Klebsiella pneumoniae*.

In Chapter 5 we compare linezolid resistance mechanisms within a cohort to VRE from the United States and Pakistan to determine that all of the US isolates were resistant due to SNPs in the 23S rRNA sequence, but the Pakistan isolates all had acquired ARGs. Two of these ARGs were the limited scope efflux pumps *optrA* and *poxxA* but the other ARGs are novel variants of the *cfp* family. In Chapter 6 we analyze a set of publicly available *Gardnerella vaginalis* genomes and metatranscriptomes of women with bacterial vaginosis to determine that what is commonly considered a single species can be interpreted as 9 different species with differences in accessory genome function and varying presence in bacterial vaginosis cases. Different genomospecies are present at varying abundance and putative virulence genes have high expression values during infection. Finally, in chapter 7 we determine the effects of acquired daptomycin resistance on the biology of *Corynebacterium striatum*. In summation this work provides novel insights on the relatedness of important human pathogens to one another and the content of their genes relevant toward infection across a wide range of species.

Chapter 1: Introduction

1.1 Bacterial pathogens and antibiotic resistance

While bacteria are ubiquitous in all studied environments, they are perhaps best known for their capacity to cause disease in humans. Evidence of their damaging effects on human civilizations have been documented since antiquity but it was not until the development of microscopes and the germ theory of disease in the 17th-19th century that the culprits of disease were identified(1). With the serendipitous discovery of penicillin by Alexander Fleming and the development of arsenic based compounds by Paul Ehrlich in the early 20th century humans finally became equipped to fight back(2). The combination of knowledge that many diseases were bacterial in origin and that small molecules can be developed or identified that could selectively kill microorganisms while leaving humans unscathed led to the golden age of antibiotic development(2). However, soon after implementation of these new drug modalities into clinical use it became apparent that more treatment failures were occurring(3). Unfortunately, this increased development of resistance coincided with a drop in the development of new antibiotics during the later part of the 20th century, leading to the current crisis of global antibiotic resistance threatening one of modern medicines greatest achievements(3).

Through analysis of bacterial DNA it has been established that bacteria can gain antibiotic resistance through alteration of antibiotic targets (ie daptomycin resistance in *Corynebacterium* occurring through loss of phosphatidylglycerol in the cell membrane), increased efflux of antibiotics via pumps (ie presence of *optrA* or *poxtA* in *Enterococcus* conferring linezolid resistance), modification of the antibiotic

(carbapenemases in Enterobacteriaceae and *Acinetobacter baumannii* able to cleave carbapenems), or decreased penetration of the antibiotic due to porin loss (multiple bacteria)(3). In 2013 and updated in 2019, the CDC analyzed epidemiological trends and the ability of our current arsenal to stave off infection to create a list of the most urgent and serious antibiotic resistance threats(3). These include carbapenem resistant Enterobacteriaceae, multidrug resistant *Acinetobacter baumannii*, extended spectrum β -lactamase producing Enterobacteriaceae, and vancomycin resistant *Enterococcus*(3).

In addition to these clearly delineated threats, there is a need for constant surveillance of possible future problems related to antibiotic treatment failure. Bacterial vaginosis is a common infectious disease of women that is caused by several bacteria including *Gardnerella vaginalis*(4). Metronidazole is an anaerobic bacteria targeting antibiotic however in 1/3-1/2 of bacterial vaginosis cases it is not capable of clearing the infection, confounding factors the mechanism of resistance by *G. vaginalis* or other vaginal bacteria is not known(4). An additional emerging complication is the rapid development of daptomycin resistance in *Corynebacterium striatum* due to loss of function mutations in phosphatidylglycerol synthase leading to a depletion of phosphatidylglycerol in the membrane(5). This resistance development has been demonstrated to occur in multiple *C. striatum* isolates and overnight(5).

1.2 Whole-genome sequencing, the bacterial species concept, and microbial taxonomy

Since the development of accurate and cost-effective next-generation sequencing technologies there has been an explosion of bacterial genomes available(3). Essentially, bacteria DNA can be isolated from purified cultures and used as input for sequencing libraries. Currently the most common platform is the second generation short read Illumina technology and the longer read PacBio and Oxford Nanopore systems(3). Following completion of the sequencing run the reads can be processed to remove artificial adapters and low-quality signals. These reads can then be assembled into scaffolds or contigs representing the bacterial genome(3). From comparison of many different bacterial genomes, we can identify genes relevant for virulence or antibiotic resistance by comparison against known databased, construct core-genome phylogenies to examine relatedness of isolates and use as input for average nucleotide identity analysis to determine if two or more bacterial genomes are from the same species(6).

Since first identified by Antoine van Leeuwenhoek in the 17th century, bacteria isolates have been categorized into species alongside known plants and animals. Given the lack of knowledge on DNA at the time, historical species delineation in bacteria was accomplished through analysis of phenotypic traits such as enzyme activity(7). It was later determined that certain molecules such as fatty acids and respiratory quinolones have efficacy in differentiating bacteria from one another. A breakthrough in understanding the relatedness of bacteria occurred in the 1980's when Carl Woese determined that the 16S rRNA sequence provides discriminatory resolution for analysis of most bacterial genera and some species(7). Since then, incorporation of the bacterial genome has been one of the most useful metrics for delineation of bacterial species. In

the pre-genomic era this was accomplished using laborious DNA-DNA hybridization assays. Currently, average nucleotide identity has become the gold standard for in silico differentiation of bacterial species(6). This has resulted in a proliferation of the number of new species and a revision of some medically relevant species such as *Klebsiella pneumoniae* into *Klebsiella variicola* and *Klebsiella quasipneumoniae*(8). Currently, it has been established that bacterial species may exist as a mono phyletic group with a high amount of similarity to one another regarding their core-genome(7). Horizontal gene transfer from other taxa may complicate this matter, which is why the increased number of bacterial genomes due to advances in sequencing technology provide an ideal way to study bacterial species(7) .

1.3 References

1. Achtman M. How old are bacterial pathogens? Proc Biol Sci. 2016;283(1836). doi: 10.1098/rspb.2016.0990. PubMed PMID: 27534956; PMCID: PMC5013766.
2. Peterson E, Kaur P. Antibiotic Resistance Mechanisms in Bacteria: Relationships Between Resistance Determinants of Antibiotic Producers, Environmental Bacteria, and Clinical Pathogens. Front Microbiol. 2018;9:2928. doi: 10.3389/fmicb.2018.02928. PubMed PMID: 30555448; PMCID: PMC6283892.
3. Boolchandani M, D'Souza AW, Dantas G. Sequencing-based methods and resources to study antimicrobial resistance. Nat Rev Genet. 2019;20(6):356-70. doi: 10.1038/s41576-019-0108-4. PubMed PMID: 30886350; PMCID: PMC6525649.
4. Bagnall P, Rizzolo D. Bacterial vaginosis: A practical review. JAAPA. 2017;30(12):15-21. doi: 10.1097/01.JAA.0000526770.60197.fa. PubMed PMID: 29135564.

5. Goldner NK, Bulow C, Cho K, Wallace M, Hsu FF, Patti GJ, Burnham CA, Schlesinger P, Dantas G. Mechanism of High-Level Daptomycin Resistance in *Corynebacterium striatum*. *mSphere*. 2018;3(4). doi: 10.1128/mSphereDirect.00371-18. PubMed PMID: 30089649; PMCID: PMC6083094.
6. Richter M, Rossello-Mora R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A*. 2009;106(45):19126-31. doi: 10.1073/pnas.0906412106. PubMed PMID: 19855009; PMCID: PMC2776425.
7. Riley MA, Lizotte-Waniewski M. Population genomics and the bacterial species concept. *Methods Mol Biol*. 2009;532:367-77. doi: 10.1007/978-1-60327-853-9_21. PubMed PMID: 19271196; PMCID: PMC2842946.
8. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, Jenney A, Connor TR, Hsu LY, Severin J, Brisse S, Cao H, Wilksch J, Gorrie C, Schultz MB, Edwards DJ, Nguyen KV, Nguyen TV, Dao TT, Mensink M, Minh VL, Nhu NT, Schultz C, Kuntaman K, Newton PN, Moore CE, Strugnell RA, Thomson NR. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. *Proc Natl Acad Sci U S A*. 2015;112(27):E3574-81. doi: 10.1073/pnas.1501049112. PubMed PMID: 26100894; PMCID: PMC4500264.

Chapter 2: *bla*_{IMP-27} on transferable plasmids in *Providencia rettgeri* and *Proteus mirabilis*

2.1 Abstract

A carbapenem resistant *Providencia rettgeri* (PR1) isolate was recovered from a wound infection in Missouri, USA. This isolate possessed an EDTA inhibitable carbapenemase that was unidentified using the Xpert CARBA-R assay. Our objective was to elucidate the molecular determinant of carbapenem resistance in this isolate. We then sought to test the transmissibility of *bla*_{IMP-27} loci in clinical *P. rettgeri* and *Proteus mirabilis* isolates. In October 2016 the novel ambler Class B carbapenemase *bla*_{IMP-27}, was reported in two different *Proteus mirabilis* (PM185 and PM187) isolates. Broth mating assays for transfer of carbapenemase activity were performed for the three clinical isolates with recipient sodium azide resistant *Escherichia coli* J53. Antibiotic susceptibility and phenotypic carbapenemase activity testing was performed on the clinical isolates, J53, and transconjugants using the Kirby-Bauer Disk diffusion method according to Clinical & Laboratory Standards Institute guidelines. Plasmid DNA from PM187, PR1, and their transconjugants were used as input for Nextera Illumina sequencing libraries and sequenced on a NextSeq platform. PR1 was resistant to both imipenem and meropenem. PM187 and PR1 could transfer resistance to *E. coli* via

plasmid conjugation (pPM187 and pPR1). pPM187 had a virB/virD4 type IV secretion system (T4SS) whereas pPR1 had traB/traD (T4SS). 2 of 3 *bla*_{IMP-27} bearing clinical isolates tested could conjugate resistance into *E. coli*. The resulting transconjugants became positive for phenotypic carbapenemase production but did not pass clinical resistance breakpoints. *bla*_{IMP-27} can be transmitted on different plasmid replicon types that rely on distinct classes of T4SS for horizontal transfer.

2.2 Introduction

In January 2016, we isolated a carbapenem resistant *Providencia rettgeri* (PR1) from a foot wound infection of a patient who visited an outside hospital affiliate of Barnes-Jewish Hospital (Missouri, United States). PR1 was positive for an EDTA-inhibited carbapenemase but no gene was identified by multiplex PCR. Whole genome sequencing (WGS) and antibiotic resistance gene (ARG) identification of the PR1 draft genome identified *bla*_{IMP-27}. *bla*_{IMP-27} was first reported in October 2016 from two *Proteus mirabilis* strains (PM185 and PM187) from the upper plains region of the United States (1). In December 2016, *bla*_{IMP-27} was identified on IncQ plasmids from a variety of swine associated Enterobacteriaceae in the United States (2). Given these recent reports, the greater Midwest region of the United States may be endemic for *bla*_{IMP-27}, and a potential source for wider geographic dissemination. Accordingly, we acquired PM185 and PM187 to understand, with PR1, the potential for lateral transfer of this resistance gene from *P. mirabilis* and *P. rettgeri* into *E. coli*, and the associated changes in antibiotic resistance (1).

2.3 Results

2.3.1 Southern blot confirmation of *bla*_{IMP-27} in transconjugants

Southern blot analysis indicates that PR1 has a single copy of *bla*_{IMP-27}, similar to the previous report on PM185 (1). In contrast, the PM187 had two copies of *bla*_{IMP-27} (1).

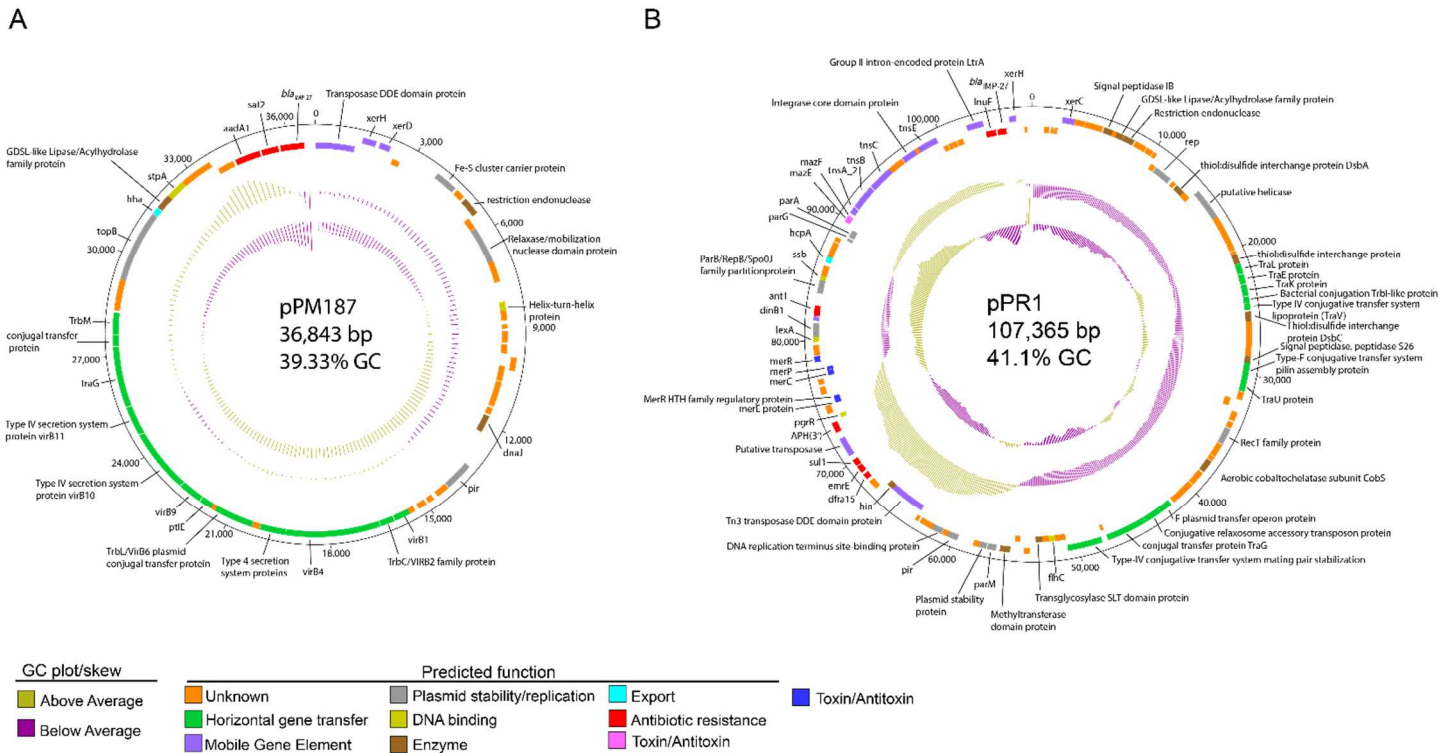


Figure 2.3.1. Depiction of plasmids harboring *bla*_{IMP-27} in PM187 and PR1
 (A) Annotated plasmid diagram from DNAPlotter of pPM187 (36,843 bp) displaying *bla*_{IMP-27} co-localized with a Class II integron gene cassette and type IV secretion system. The inner most ring shows GC plot, the second ring shows GC skew, the third ring represents open reading frames in the forward direction, and the fourth ring (adjacent to the nucleotide position counter) indicates open reading frames in the reverse direction. (B) Annotated plasmid diagram from DNAPlotter of pPR1 (107,365 bp) displaying *bla*_{IMP-27} co-localized with a Class II integron gene cassette, a *tra* operon, and additional resistance genes.

2.3.2 Conjugation effects on phenotypic resistance in *E. coli* J53

PM185 and PM187 were intermediate and susceptible to meropenem and imipenem, respectively (Table 1). Only PR1 was resistant to both carbapenems. PM185 was indeterminate for the carbapenem inactivation method but PM187 and PR1 were both phenotypically positive (Table 1). Transconjugants were obtained from conjugation

assays of PR1 and PM187 with the *E. coli* J53 recipient but not PM185. Although conjugation did not achieve clinical resistance guidelines, the zone size for meropenem decreased from 32 mm in J53 to 25 mm in J53: pPM187 and 27 mm in J53:pPR1 (Table 2.3.1) (27). The zone size for imipenem decreased a lesser amount, from 33 mm in J53 to 31 and 32 mm in J53: pPM187 and J53:pPR1, respectively. Both transconjugants were positive for phenotypic carbapenem production (Table 2.3.1).

2.3.3 Complete *bla*_{IMP-27} positive plasmids sequence

The plasmid from PM187, pPM187 (Genbank NOWA01000087.1), contains a putative virB/D4 IV secretion system operon, providing a potential mechanism for horizontal dissemination (Figure 2.3.2A). The *virB4* amino acid sequence had 100% identity over its entire length with a conjugal transfer protein (WP_012368868.1) from *P. mirabilis* HI4320 (11). pPM187 has an IncX8 backbone, a newly discovered IncX family member (12). Unlike pPM187, the assembled *bla*_{IMP-27} bearing plasmid, pPR1 (Genbank NOWC01000095.1) did not have a plasmid replicon identified. pPR1 also contained a putative type IV secretion system, though of the tra/trb type (Figure 2.3.2B). The *traN* amino acid sequence had 100% identity across its entire length to the *traN* (WP_023159916.1) of the *bla*_{NDM-1} bearing plasmid pPrY2001 from *P. rettgeri* 09ACRGNY2001 (13).

	Zone of Cleara	Interpret ation										
--	----------------------	--------------------	--	--	--	--	--	--	--	--	--	--

	nce (mm)																
Antibiotic	PM185		PM187		PR1		J53		J53:pPM 187		J53:pPR 1						
	Ampicillin	24	S	6	R	4	R	1	S	2	1	S	1	S			
Cefazolin	9	R	9	R	6	R	5	S	2	8	R	9	R				
Cefotetan	11	R	1	I	4	I	6	R	3	3	S	1	R	0	R		
Ceftriaxone	17	R	1	R	7	R	1	R	3	5	S	1	R	4	R		
Ceftazidime	23	S	2	I	0	I	2	S	3	0	S	1	R	5	R		
Cefepime	20	I	1	I	9	I	1	SDD	3	6	S	2	S	8	S	2	S
Meropenem	20	I	2	S	2	S	6	R	3	2	S	2	S	5	S	7	S
Imipenem	20	I	2	S	4	S	1	R	3	3	S	3	S	1	S	2	S

Pipercillin-Tazobactam	33	S	2 6	S	3 1	S	3 0	S	3 0	S	3 1	S
Ampicillin-Sulbactam	23	S	1 8	S	6 6	R	2 4	S	2 0	S	2 2	S
Ciprofloxacin	36	S	3 2	S	2 7	S	2 5	S	2 5	S	2 5	S
Levofloxacin	35	S	3 0	S	2 6	S	2 5	S	2 5	S	2 5	S
Gentamicin	23	S	1 5	S	1 6	S	2 5	S	2 5	S	2 6	S
Amikacin	22	S	2 4	S	2 4	S	2 5	S	2 5	S	2 6	S
Trimethoprim-sulfamethoxazole	30	S	2 3	S	6 6	R	3 5	S	2 5	S	6 6	R
Colistin	6	R	6	R	6	R	1 6	S	2 4	S	1 6	S
Aztreonam	38	S	3 5	S	3 5	S	3 6	S	3 5	S	3 5	S
Doxycycline	6	R	6	R	6	R	2 2	S	2 2	S	2 2	S

Table 2.3.1. Zone disk diffusion results for wildtype (*P. mirabilis* PM185, *P. mirabilis* PM187, *P. rettgeri* PR1, and *E. coli* J53) and transconjugant (*E. coli* J53:pPM187 and *E. coli* J53:pPR1) isolates in this study.

Minocycline			1			2		2		2
e	11	R	0	R	6	R	6	S	6	S
Tigecycline	18	I	2		2		2		2	
m			3	S	0	S	9	S	9	S
Carbapenem										
Inactivation		Indeterminate		Positive		Positive		Negative		Positive
Method				ve		ve		ve		ve

2.4 Discussion

In this study, we used conjugation experiments to determine that two *bla*_{IMP-27} positive clinical isolates, PM187 and PR1, could transfer carbapenemase production to *E. coli*. We used Illumina sequencing of the transconjugants and clinical isolates to assemble the *bla*_{IMP-27} bearing plasmids, pPM187 and pPR1.

E. coli transconjugants with these plasmids (pPR1 and pPM187) gain detectable carbapenemase activity, but this activity does not shift the transconjugants to a past clinical breakpoints for carbapenem resistance. It is possible that regulatory or translational optimization of the conjugated *bla*_{IMP-27} bearing plasmid in *E. coli* is required for clinical resistance (14). In addition to *bla*_{IMP-27} expression, it is also possible that porin mutations

or efflux activity in the clinical isolates could contribute to phenotypic carbapenem resistance (15).

A previous investigation found that while *bla*_{IMP-27} was plasmid-borne in swine-associated Enterobacteriaceae, the IncQ plasmids were not conjugatable. In contrast, the plasmids we have completely sequenced are capable of self-mobilization, likely due to a *virB/virD4* T4SS in pPM187 and a *traB/traA* T4SS in pPR1. The *virB4* and *traN* gene from these T4SS showed similarity to previously described systems from pathogenic *P. mirabilis* HI4320 and carbapenem resistant *P. rettgeri* 09ACRGNY2001 (11,13). A limitation of our approach was that Illumina whole genome sequencing could not unambiguously assemble the chromosome and all plasmids in PM187 and PR1. Further work is therefore warranted using long reads sequencing technology (e.g., from PacBio or Oxford Nanopore sequencing) on *bla*_{IMP-27} isolates to unequivocally determine chromosomal sequences and compare the nonconjugatable *bla*_{IMP-27} IncQ plasmids with pPM187 and pPR1. Although southern blot analysis indicates only a single *bla*_{IMP-27} loci exist in PM185 and PR1, this may further enable a comparison between the chromosomal and plasmid (pPM187) platforms of *bla*_{IMP-27} in PM187.

*bla*_{IMP-27} was unidentified using the FDA-cleared Xpert CARBA-R assay but the first report of *bla*_{IMP-27} used the ARM-D™ Multiplex PCR, which indicates some commercially available platforms can assay for *bla*_{IMP-27} (11). Therefore, further evaluations of commercial molecular diagnostic tests for *bla*_{IMP-27} is warranted.

2.5 Materials & Methods

2.5.1 Bacterial Isolates

The *Providencia rettgeri* isolate (PR1) was recovered from a chronic foot wound infection clinical culture. The isolate received for evaluation was a de-identified strain. As a result, the study team was not able to obtain patient consent. *Proteus mirabilis* strains (PM185 and PM187) were provided by Nancy Hanson at Creighton University (1). The sodium azide resistant *E. coli* J53 strain (ATCC number BAA-2730™) was used as a recipient for transconjugation experiments.

2.5.2 Broth Conjugation

Colonies of PM185, PM187, PR1, and wildtype *E. coli* J53 were separately suspended in Tryptic Soy Broth (TSB) (Sigma Aldrich, St. Louis, MO) and diluted to 0.05 OD₆₀₀. 100 µl of PM185, PM187, and PR1 were separately added to 100 µl *E. coli* J53 (for a 1:1 ratio) and diluted to 5 mL with TSB. Co-cultures were incubated at 37 °C without shaking for 24 hours. 50 µl of co-cultures were suspended onto MacConkey agar plates containing sodium azide (Thermo Fisher Scientific, Waltham, MA) (150 µg/ml) and ceftriaxone (5 µg/ml), spread with glass beads, and incubated for 18 hours at 37 °C. Individual transconjugant colonies were propagated overnight in TSB supplemented with 5 µg/ml ceftriaxone under shaking conditions (220 rpm).

2.5.3 Susceptibility Testing

Each clinical isolate, J53, J53:pPR1, and J53:pPM187 were cultured overnight as described previously. *E. coli* ATCC 25922 was used as a quality control. Susceptibility testing was performed using Kirby Bauer Disk Diffusion on Mueller Hinton Agar (Hardy Diagnostics) in accordance with CLSI Standards (3).

2.5.4 Southern blot

Total genomic DNA was extracted from PM185, PM187, PR1, J53, J53:pPM187, and J53:pPR1 using the Bacteremia kit (Qiagen). Southern blot protocol was used to separate the plasmid components from the chromosome and examine localization of *bla*_{IMP-27} using P-32 labeled primers (21).

2.5.5 Plasmid assembly and annotation

We used Illumina sequencing to specifically investigate *bla*_{IMP-27} bearing plasmids in PR1 and PM187. Plasmid DNA was obtained using a miniprep kit (Qiagen, Valencia, CA). Plasmid DNA for PR1 and PM187 was processed to remove Illumina adapters (trimmomatic) and contaminating DNA (deconseq). The paired reads were assembled into contigs with SPAdes v3.9.0 (4). Raw reads from the transconjugant minipreps were processed for quality in a similar manner. 100% of the transconjugant reads that aligned back to the clinical isolate plasmid assembly using Bowtie2 were assembled into contigs with SPAdes v3.9.0 (5) (4). Gaps were closed by PCR and Sanger-sequencing (Genewiz, South Plainfield, NJ) to yield finished plasmid assemblies. Open reading frames were annotated for coding sequence using prokka (6). Antibiotic resistance genes were additionally annotated with Resfams and the ResFinder web server (<https://cge.cbs.dtu.dk/services/ResFinder/>) (7, 8). pPM187 and pPR1 plasmid maps were made by viewing the gff3 files in DNAPlotter and manually annotated for putative open reading frame function(9). Select T4SS genes were submitted to blastp against the nonredundant protein sequence database on 12/10/17 (10).

2.6 Acknowledgments

The authors would like to thank Center for Genome Sciences & Systems Biology staff Jessica Hoisington-Lopez, Brian Koebbe, & Eric Martin for performing Illumina WGS and operating the High Throughput Computing Facility. The authors would also like to thank Nancy Hanson for generously providing PM185 and PM187. R.F.P presented a portion of this work as a poster at the 2017 American Society for Microbiology Microbe conference in New Orleans, LA. This work was supported in part by a grant to G.D. from the National Institute of General Medical Sciences (NIGMS: <http://www.nigms.nih.gov/>) of the NIH under award number R01 GM099538. R.F.P was supported by a NIGMS training grant through award T32 GM007067 (PI: James Skeath) and the Monsanto excellence fund graduate fellowship. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

2.7 References

1. Dixon N, Fowler RC, Yoshizumi A, Horiyama T, Ishii Y, Harrison L, Geyer CN, Moland ES, Thomson K, Hanson ND. IMP-27, a Unique Metallo-beta-Lactamase Identified in Geographically Distinct Isolates of *Proteus mirabilis*. *Antimicrob Agents Chemother.* 2016;60(10):6418-21. doi: 10.1128/AAC.02945-15. PubMed PMID: 27503648; PMCID: PMC5038328.
2. Mollenkopf DF, Stull JW, Mathys DA, Bowman AS, Feicht SM, Grooters SV, Daniels JB, Wittum TE. Carbapenemase-Producing Enterobacteriaceae Recovered from the Environment of a Swine Farrow-to-Finish Operation in the United States. *Antimicrob Agents Chemother.* 2017;61(2). doi: 10.1128/AAC.01298-16. PubMed PMID: 27919894; PMCID: PMC5278694.

3. Institute CaLS. Performance standards for antimicrobial susceptibility testing: Twenty-third Informational Supplement M100-S232013.
4. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 2012;19(5):455-77. doi: 10.1089/cmb.2012.0021. PubMed PMID: 22506599; PMCID: PMC3342519.
5. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012;9(4):357-9. doi: 10.1038/nmeth.1923. PubMed PMID: 22388286; PMCID: PMC3322381.
6. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics.* 2014;30(14):2068-9. doi: 10.1093/bioinformatics/btu153. PubMed PMID: 24642063.
7. Gibson MK, Forsberg KJ, Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.* 2015;9(1):207-16. doi: 10.1038/ismej.2014.106. PubMed PMID: 25003965; PMCID: PMC4274418.
8. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother.* 2012;67(11):2640-4. doi: 10.1093/jac/dks261. PubMed PMID: 22782487; PMCID: PMC3468078.
9. Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J. DNAPlotter: circular and linear interactive genome visualization. *Bioinformatics.* 2009;25(1):119-20. doi: 10.1093/bioinformatics/btn578. PubMed PMID: 18990721; PMCID: PMC2612626.

10. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421. doi: 10.1186/1471-2105-10-421. PubMed PMID: 20003500; PMCID: PMC2803857.
11. Pearson MM, Sebahia M, Churcher C, Quail MA, Seshasayee AS, Luscombe NM, Abdellah Z, Arrosmith C, Atkin B, Chillingworth T, Hauser H, Jagels K, Moule S, Mungall K, Norbertczak H, Rabinowitsch E, Walker D, Whithead S, Thomson NR, Rather PN, Parkhill J, Mobley HL. Complete genome sequence of uropathogenic *Proteus mirabilis*, a master of both adherence and motility. *J Bacteriol*. 2008;190(11):4027-37. doi: 10.1128/JB.01981-07. PubMed PMID: 18375554; PMCID: PMC2395036.
12. Guo Q, Su J, McElheny CL, Stoesser N, Doi Y, Wang M. IncX2 and IncX1-X2 Hybrid Plasmids Coexisting in a FosA6-Producing *Escherichia coli* Strain. *Antimicrob Agents Chemother*. 2017;61(7). doi: 10.1128/AAC.00536-17. PubMed PMID: 28438937; PMCID: PMC5487653.
13. Mataseje LF, Boyd DA, Lefebvre B, Bryce E, Embree J, Gravel D, Katz K, Kibsey P, Kuhn M, Langley J, Mitchell R, Roscoe D, Simor A, Taylor G, Thomas E, Turgeon N, Mulvey MR, Canadian Nosocomial Infection Surveillance P. Complete sequences of a novel bla_{NDM-1}-harbouring plasmid from *Providencia rettgeri* and an FII-type plasmid from *Klebsiella pneumoniae* identified in Canada. *J Antimicrob Chemother*. 2014;69(3):637-42. doi: 10.1093/jac/dkt445. PubMed PMID: 24275114.
14. Zeng X, Lin J. Beta-lactamase induction and cell wall metabolism in Gram-negative bacteria. *Front Microbiol*. 2013;4:128. doi: 10.3389/fmicb.2013.00128. PubMed PMID: 23734147; PMCID: PMC3660660.

15. Potter RF, D'Souza AW, Dantas G. The rapid spread of carbapenem-resistant Enterobacteriaceae. *Drug Resist Updat.* 2016;29:30-46. doi: 10.1016/j.drug.2016.09.002. PubMed PMID: 27912842; PMCID: PMC5140036.

Chapter 3: Population structure, antibiotic resistance, and uropathogenicity of *Klebsiella variicola*

3.1 Abstract

Klebsiella variicola is a member of the *Klebsiella* genus and often misidentified as *Klebsiella pneumoniae* or *Klebsiella quasipneumoniae*. The importance of *K. pneumoniae* human infections has been known; however, a dearth of relative knowledge exists for *K. variicola*. Despite its growing clinical importance, comprehensive analyses of *K. variicola* population structure and mechanistic investigations of virulence factors and antibiotic resistance genes have not yet been performed. To address this, we utilized *in silico*, *in vitro*, and *in vivo* methods to study a cohort of *K. variicola* isolates and genomes. We found that the *K. variicola* population structure has two distant lineages composed of two and 143 genomes, respectively. 10/145 of *K. variicola* genomes harbored carbapenem resistance genes and 6/145 contained complete virulence operons. While the β -lactam *bla*_{LEN} and quinolone *oqxAB* antibiotic resistance genes were generally conserved within our institutional cohort, unexpectedly 11 isolates were nonresistant to the β -lactam ampicillin and only one isolate was nonsusceptible to the quinolone ciprofloxacin. *K. variicola* isolates have variation in ability to cause urinary tract infections in a newly developed murine model, but importantly a strain had statistically significant higher bladder colony forming units compared to the model uropathogenic *K. pneumoniae* strain TOP52. Type 1 pilus and genomic identification of altered *fim* operon structure were associated with differences in bladder colony forming units for the tested strains. 9 newly reported types of pili genes were discovered in the *K. variicola* pan-genome, including the first identified P-pilus in *Klebsiella* spp. Infections caused by

antibiotic resistant bacterial pathogens is a growing public health threat. Understanding of pathogen relatedness and biology is imperative for tracking outbreaks and developing therapeutics. Here, we detail the phylogenetic structure of 145 *K. variicola* genomes from different continents. Our results have important clinical ramifications as high-risk antibiotic resistance genes are present in *K. variicola* genomes from a variety of geographies and as we demonstrate that *K. variicola* clinical isolates can establish higher bladder titers than *K. pneumoniae*. Differential presence of these pilus genes in *K. variicola* isolates may indicate adaptation for specific environmental niches. Therefore, due to the potential of multidrug resistance and pathogenic efficacy, identification of *K. variicola* and *K. pneumoniae* to a species level should be performed to optimally improve patient outcomes during infection. This work provides a foundation for our improved understanding of *K. variicola* biology and pathogenesis.

3.2 Introduction

Klebsiella variicola was initially believed to be a plant-associated, distant lineage of *Klebsiella pneumoniae*, however it has subsequently been recovered from human clinical specimens(1). Despite increasing knowledge on the distinctness of *K. variicola*, *K. pneumoniae* and *Klebsiella quasipneumoniae*, misidentification within the clinical microbiology lab commonly occurs (2, 3). This may have clinical implications, as one study demonstrated that *K. variicola*-infected patients have higher mortality than *K. pneumoniae*-infected patients (4). Furthermore, several virulence genes (VGs) including siderophores, allantoin utilization genes, and glycerate pathway genes have been reported in select *K. variicola* strains (5, 6). *K. variicola* has been shown to contain a

large pan-genome that is distinct from *K. quasipneumoniae* and *K. pneumoniae*, but the functional consequences of differential gene content has not been explored (2, 7).

In this study, we retrospectively analyzed a cohort of *Klebsiella* isolates collected from 2016-2017 at Washington University in St. Louis School of Medicine/Barnes-Jewish Hospital Clinical Microbiology Laboratory (WUSM) for possible *K. variicola* strains using matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS) and *yggE* PCR/restriction fragment length polymorphism (RFLP) assays. We performed Illumina whole-genome sequencing (WGS) to compare *K. variicola* from our institution with publicly available genomes in the first global evaluation of this species. We particularly focused on annotation of canonical *Klebsiella* VGs and ARGs, and then assessed their functional consequences using *in vitro* assays and *in vivo* murine infections. Our results demonstrate that population structure, antibiotic resistance, and uropathogenicity of *K. variicola* are generally similar to *K. pneumoniae*, but variability among *K. variicola* genomes has important clinical implications with varying strain efficacy in a murine model of urinary tract infection (UTI).

3.3 Results

3.3.1 Average nucleotide identity can delineate *Klebsiella variicola* from related species

We performed Illumina WGS on 113 isolates that are commonly misidentified as *K. pneumoniae* (*K. variicola* (n=56), *K. quasipneumoniae* (n=3), *K. pneumoniae* (n=53), and *Citrobacter freundii* (n=1)). They were identified by Bruker biotyper MALDI-TOF MS and *yggE* RFLP assays from a variety of adult infection sites. The isolates were

retrieved from the Barnes-Jewish hospital clinical microbiology laboratory (St. Louis, MO, USA) in 2016-2017. We used pyANI with the mummer method to calculate the pairwise average nucleotide identity (ANI_m) between the isolates in our cohort and retrieved publicly available *Klebsiella* genomes (n=90)(8, 9). The *C. freundii* was originally classified as *K. pneumoniae* from the VITEK MS MALDI-TOF MS v2.3.3 but was later determined to be *Citrobacter freundii* by Bruker Biotyper MALDI-TOF MS. The *yggE* PCR/RFLP was indeterminate for this isolate. Confirmatory *yggE* PCR/RFLP had 94.6% (53/56) concordance with MALDI-TOF for prediction of *K. variicola* within our cohort (Figure 3.3.1). While one genome was dropped from downstream analysis, the other 55 WUSM *K. variicola* genomes all had > 95% ANI_m with the reference genome of *K. variicola* At-22(5). *K. variicola* HKUPOLA (GCA_001278905.1) had > 95% ANI_m with *K. quasipneumoniae* ATCC 7000603 reference genome but not *K. variicola* At-22, indicating that it is likely a misannotated *K. quasipneumoniae* and not a *K. variicola*. The remainder of the NCBI *K. variicola* genomes clustered with *K. variicola* At-22 and the WUSM *K. variicola* cohort. 100% (41/41) of the *K. pneumoniae* genomes from NCBI that were suspected to be *K. variicola* due to BLAST similarity had > 95% ANI_m with *K.*

variicola At-22 but not *K. pneumoniae* HS11286 or *K. pneumoniae* CAV1042 (Figure 3.3.1).

Hierarchical clustering of the pairwise ANIm values replicated previous phylogenetic analysis showing that *K. pneumoniae* and *K. quasipneumoniae* are more closely related to each other than to *K. variicola* (Figure 3.3.1). Interestingly, the clustering pattern within *K. variicola* indicated that two isolates, KvMX2 (FLLH01.1) and YH43

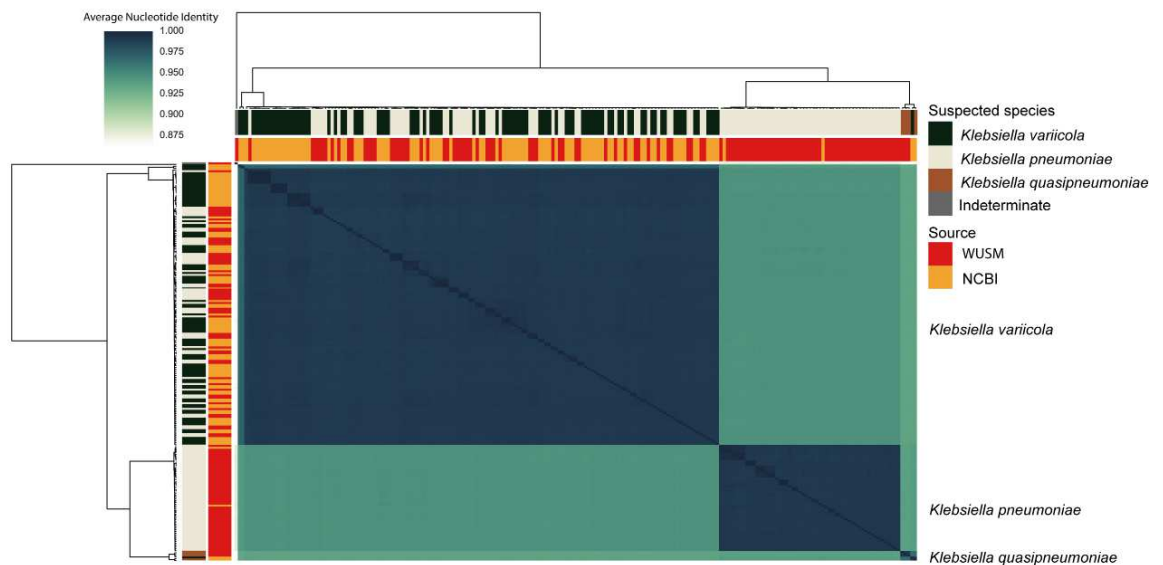


Figure 3.3.1 Pairwise Average Nucleotide Identity Clustermap of WUSM and NCBI *Klebsiella*

Hierarchical clustering and heatmap of pairwise ANIm values among all isolates. The source of isolates (WUSM or NCBI) and initial species delineation (*K. variicola*, *K. pneumoniae*, or *K. quasipneumoniae*) are shown as colored bars adjacent to the heatmap. The three major blocks are labeled by their final species determination.

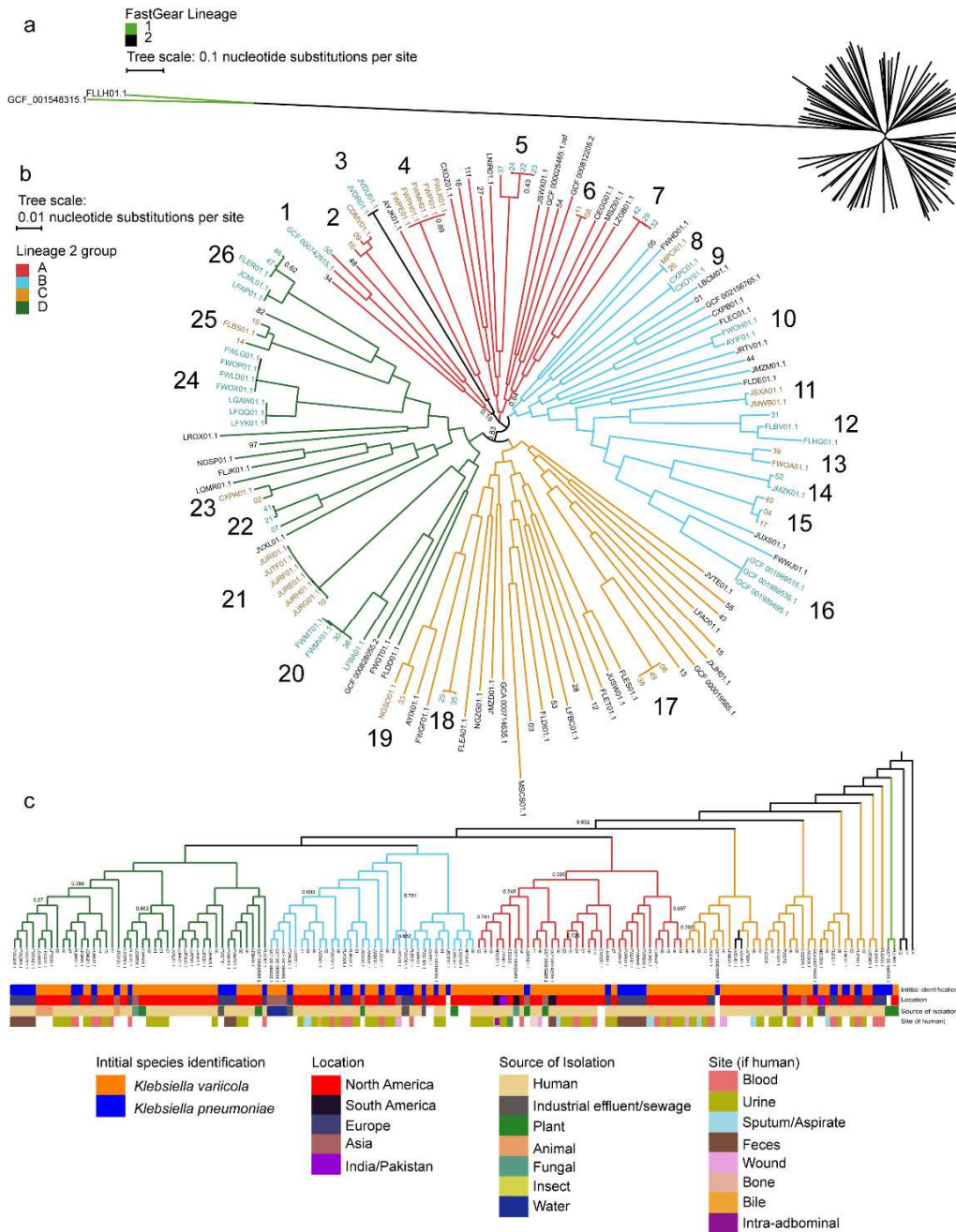
(GCF_001548315.1), are more closely related to one another than to the remainder (143/145) of the *K. variicola* genomes. Given that *K. quasipneumoniae* can be differentiated into two subspecies based on ANI with the BLAST method (ANIb), we used the JSpecies ANIb program to specifically compare KvMX2 and YH43 with *K. pneumoniae* ATCC BAA-1705, *K. quasipneumoniae* ATCC 7000603, and 3 other *K.*

variicola genomes(10). KvMX2 and Yh43 have 98.02% ANIb with one another but an average of 96.67% ,96.65%, 96.68% ANIb with WUSM_KV_53, WUSM_KV_15, and *K. variicola* At-22, respectively. Consistent with our pyANI ANIm result, none of the *K. variicola* strains had >95% ANIb with *K. pneumoniae* ATCC BAA-1705 or *K. quasipneumoniae* ATCC 7000603. These data suggest that MALDI-TOF MS or *yggE* PCR/RFLP may be effective means to differentiate *K. variicola* from *K. pneumoniae* in the absence of WGS.

3.3.2 *Klebsiella variicola* is composed of two distantly related lineages

Core-genome alignment of the 1,262 genes at 90% identity shared by strains in all *Klebsiella* species and a *Kluyvera georgiana* outgroup show that the *K. variicola* isolates are in a cluster with *K. pneumoniae*, *K. quasipneumoniae*, and the newly described *K. quasivariicola*(11). Core genome alignment of the 3,430 core-genes at 95% nucleotide identity for the entire gene length by all 145 *K. variicola* genomes indicate that KvMX2 and Yh43 are distantly related to the other 143 genomes (Figure 3.3.2a). These other genomes form a star-like phylogeny showing deep-branching clusters radiating from the center of the tree. FastGear, which uses hierBAPS to identify lineages and then searches for recombination between lineages, supported the differentiation of KvMX2 and Yh43 into a separate lineage from the other genomes and identified 6 instances of recombination between these two lineages(12, 13).

Phylogenomic network analysis and quantification of recombination from parSNP showed minimal recombination within the 143 *K. variicola* lineage 2 genomes, with approximately 1.62% of the *K. variicola* genome believed to be recombinant (14). The Nearest Neighbor network of the 3,496 genes shared by the lineage 2 genomes and a



recombination-free phylogenetic tree of the 143 genomes from parSNP showed many deep-branching clades with a star-like phylogeny. This tree topology was similar with and without recombination, which suggests that *K. variicola* lineages emerged early from a single

Figure 3.3.2. Population structure of *K. variicola* genomes

(a) Approximate-maximum-likelihood tree of the total 145 *K. variicola* genomes and annotation of FastGear lineage identification. (b) Recombination free parSNP tree of the closely related lineage 2 genomes with quantitative clustering from ClusterPicker added as alternating teal and brown labels adjacent to Cluster number (1-26). Bootstrap support values below 80% are depicted as node labels. Monophyletic groups of these clusters were colored if they were similar in (c) the dendrogram showing the evolutionary context of the cluster when compared to *K. pneumoniae* (KP), *K. quasipneumoniae* (KQ), and *K. aerogenes* (KA). Relevant metadata shows for initial identification, geographic location, source of isolation, and body site are adjacent to the assembly names. Bootstrap support values below 80% are depicted as node labels.

common ancestor into equally distant clades across different environments. Quantitative clustering of the 143 genomes in the second lineage with ClusterPicker showed that 56.6% (81/143) genomes fall into 26 clusters, with 57.7% (15/26) of the clusters containing more than 2 genomes (Fig. 2b)(15). Only 46.2% (12/26) of clusters contain isolates from both WUSM_KV and NCBI. The largest clusters, 24 and 21, each contain 7 genomes. Cluster 21 contained WUSM_KV_10 and 6 genomes from an analysis of patient isolates at an intensive care unit in Seattle, Washington (USA). Although they were in the same cluster WUSM_KV_10 differed from these isolates at 1,882 sites across the 4,867 genes shared at 95% identity.

To better understand the context of the 4 groups in lineage 2, we aligned the 2,932 genes shared among the 145 *K. variicola* genomes, *Klebsiella* (Formerly *Enterobacter*) *aerogenes* KCTC 2190, *K. quasipneumoniae* ATCC 700603, and *K. pneumoniae* ATCC BAA-1705 at $\geq 90\%$ identity to create a dendrogram (Figure 3.3.2c). This method preserved the conservation of the lineage 2 groups but showed a different order. The only discrepancy observed is that in the lineage 2 phylogenetic tree, cluster 3 appeared to be in the A group, however, both 521_SSON and 524_SBOY are more similar to C group genomes in the dendrogram. This incongruency is consistent with cluster 3 radiating away from cluster 4 near the center of the phylogenetic tree (Figure 3.3.2b). Addition of metadata onto the dendrogram showed that the *K. variicola* cohort spans most geographic locations, with the notable exception of Africa and Oceania (Figure 3.3.2c). The *K. variicola* genomes showed a remarkable level of source diversity, with representative isolates from animals (n=4), fungi (n=2), plants (n=7), water (n=3), and industrial waste (n=6). However, as a testament to the pathogenic potential of *K.*

variicola, 79.5% (114/145) genomes came from sites associated with humans. Of the human-associated sites, 40.4% (46/114) came from urine and 19.2% (22/114) came from blood (Figure 3.3.2c). We did not observe any apparent association with geography, habitat, or infection site for any of the *K. variicola* clades. 67/145 isolates had a sequence type (ST) identified using the *K. pneumoniae* multilocus sequence type scheme. Consistent with the distance between lineages, 44 different STs were identified. ST1562 and ST641 had the highest number of isolates (n=4). In summary, these data demonstrate that *K. variicola* has diverse population structure and can be found in a variety of environmental and host niches.

3.3.3 Acquired ARGs and VGs are not restricted to any *K. variicola* cluster

We applied ResFinder to determine the burden of acquired ARGs amongst the *K. variicola* strains (Figure 3.3.3a) (16). β -lactamase genes were the most abundant ARG in the *K. variicola* cohort (n = 26). As expected, *bla*_{LEN} was almost universally conserved, as 837_KPNE was the only isolate without one identified. 10 different *bla*_{LEN} alleles were found. *bla*_{LEN-16} was most common (51/145), followed by *bla*_{LEN-24} (40/145) and *bla*_{LEN-2} (31/145). Carbapenemases were rare but *bla*_{KPC-2} (4/145), *bla*_{KPC-6} (1/145), *bla*_{NDM-1} (1/145), *bla*_{NDM-9} (3/145), and *bla*_{OXA-48} (1/145) were each identified across a total of 10/145 strains. *bla*_{CTX}, *bla*_{SHV}, *bla*_{TEM}, and non-carbapenemases *bla*_{OXA} genes were also identified, but we did not detect any Class C β -lactamase genes or non-*bla*_{NDM} Class B β -lactamase genes. Aminoglycoside ARGs (n=10), including members of the *aac*, *aad*, *aph*, and *str* families, comprised the second most abundant class. ARGs against folate synthesis inhibitors (n=8), quinolones (n=7), amphenicols (n=4), tetracyclines (n=2), macrolides/lincosamides/streptogramins (n=2), and fosfomycin

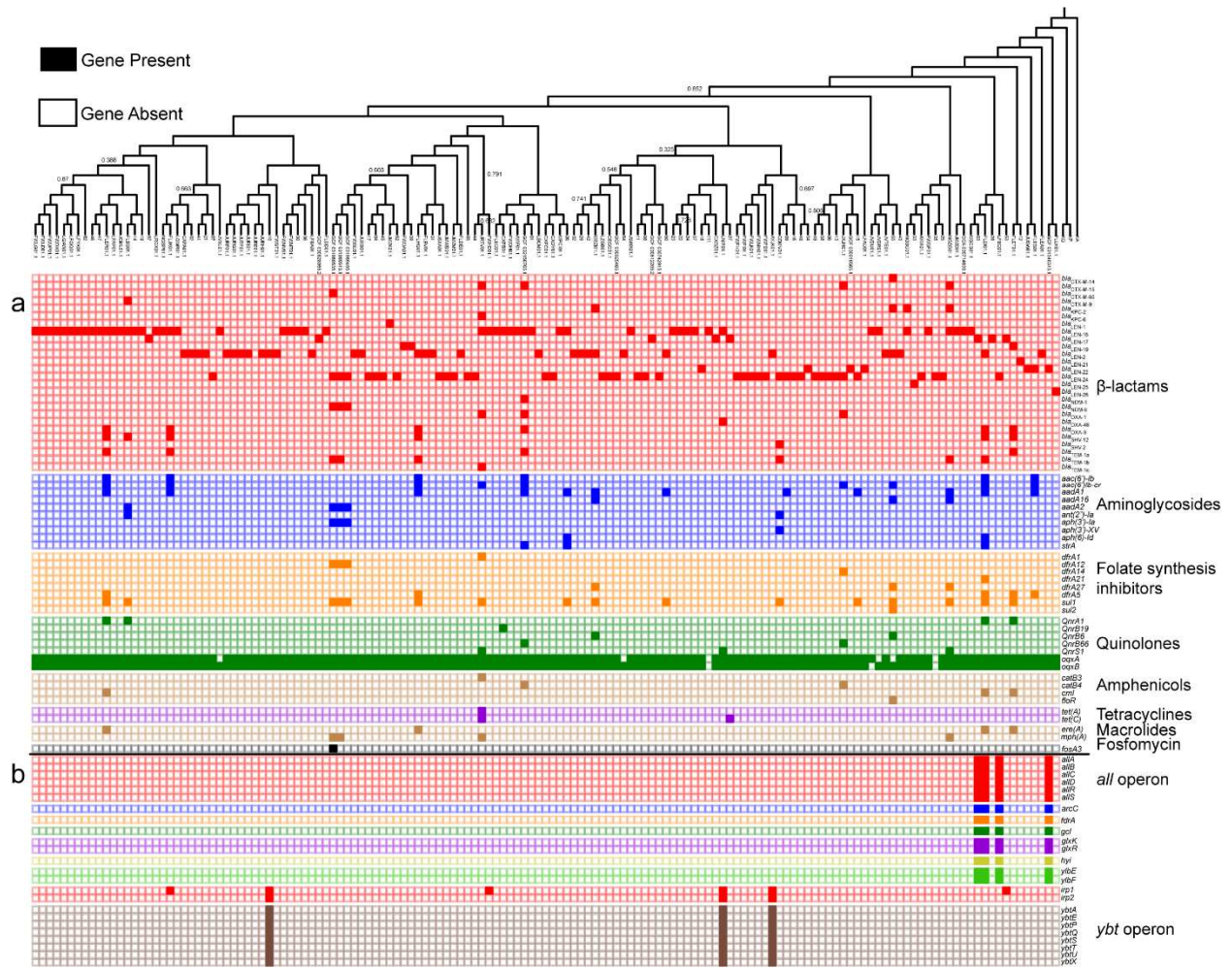


Figure 3.3.3 Distribution of acquired antibiotic resistance and virulence genes in the *K. variicola* cohort

Presence/Absence matrix of ARGs (a), virulence genes (b), and plasmid replicons (c) ordered for all *K. variicola* genomes against the dendrogram from Figure 3.3.2c.

(n=1) were also found (Fig. 3a). In addition to the near-total conservation of *bla*_{LEN}, the quinolone efflux pump components *oqxAB* were found in almost all isolates (139/145). Across the 145 genomes, the median and mode number of ARGs were both 3. 6.89% (10/145) genomes harbored ≥ 10 ARGs, including WUSM_KV_55 from our cohort.

We used the *K. pneumoniae* BIGSdb database

(bigsd.b.pasteur.fr/klebsiella/klebsiella.html) and BLASTN to identify canonical *Klebsiella*

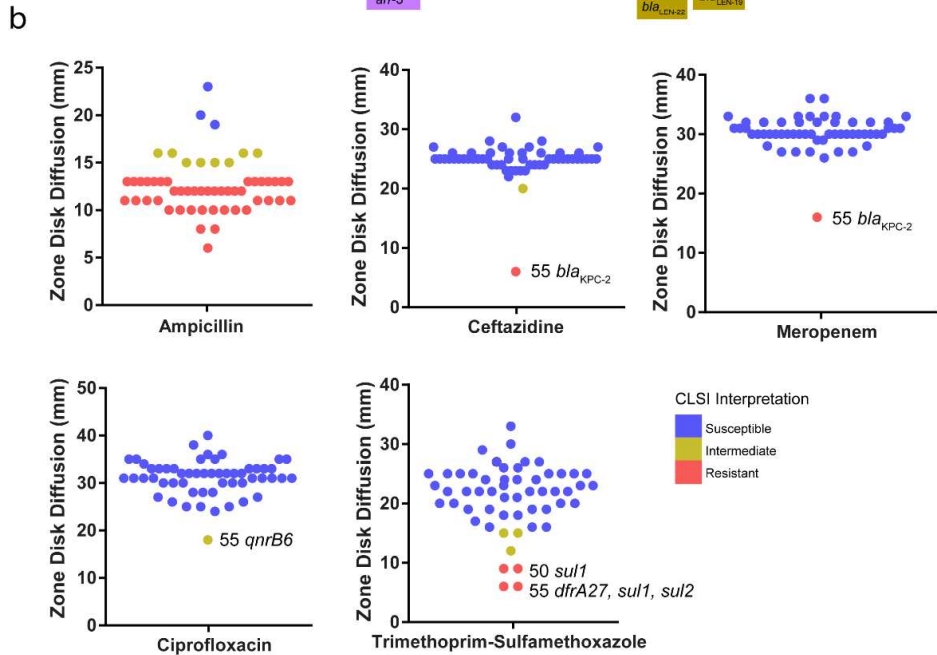
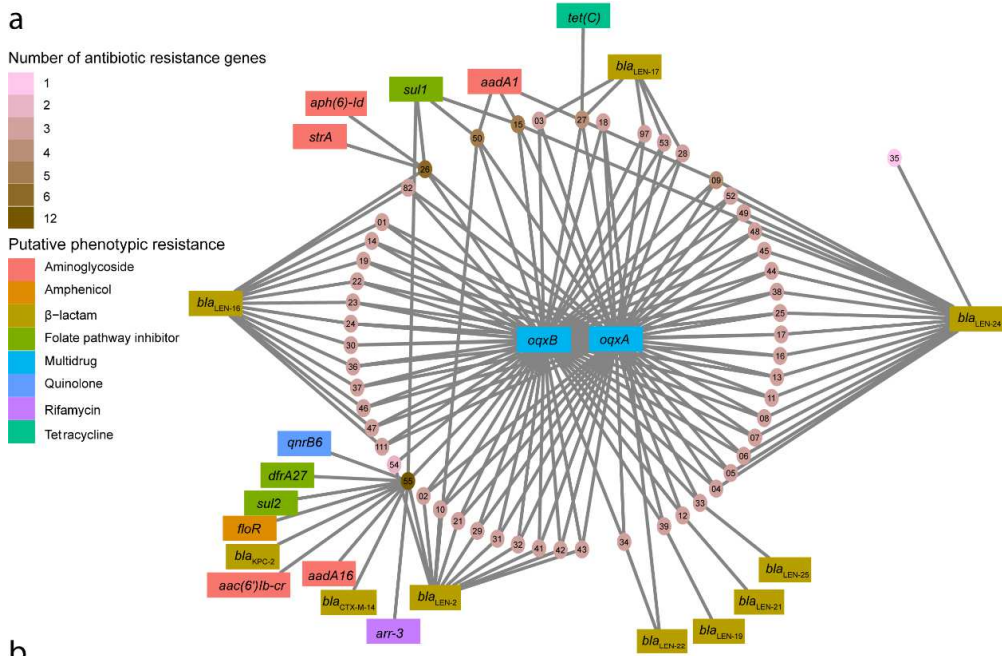
VGs in the *K. variicola* strains (Figure 3.3.3b). In contrast to ARGs, previously characterized *Klebsiella* VGs were found only sporadically in the *K. variicola* cohort. Interestingly, the *all* allantoin utilization operon, *arc*, *fdxA*, *gcl*, *glxKR*, *hyi*, and *ybbWY* genes were found in the distantly related YH43 genome as well as the closely related BIDMC90, k385, and WUSM_KV_03 genomes. *Irp12* and the *ybt* operon were found together in the three isolates 50878013, MGH 20, and WUSM_KV_10. *Irp1* was found on 3 additional instances but with no other VGs. Among 8 isolates containing the full *all* or *ybt* operon, six had only 3 ARGs; however, 50878013 contains the *ybt* operon, *irp12*, and has 5 ARGs including the *bla*_{OXA-48} carbapenemases, while k385 had 17 ARGs but no carbapenemases.

3.3.4 WUSM *K. variicola* cohort are susceptible to most antibiotics

We constructed a network diagram of ARGs and isolates to identify connectivity within the *K. variicola* strains from our cohort (Figure 3.3.4a). WUSM_KV_55 had twice as many ARGs (n=12) as the next closest isolate, WUSM_KV_26 (n=6). Most notably, WUSM_KV_55 contained the carbapenemase gene *bla*_{KPC-2}. In addition to the core β -lactamase *bla*_{LEN-2}, this isolate also contained a *bla*_{CTX-M-14} gene. Redundancy was again observed for the ARGs against aminoglycosides and sulfonamides, as WUSM_KV_55 contained *aac(6')Ib-cr*, *aadA16* and *sul1*, *sul2*. Within our cohort, this isolate was the only isolate found to harbor additional quinolone (*qnrB6*), rifampin (*arr-3*), and amphenicol (*floR*) ARGs. Interestingly, it possesses *oqxB* but not *oqxA*. Conversely, WUSM_KV_35 harbored the lowest number of acquired ARGs, as it lacked *oqxAB* but carried *bla*_{LEN-24}.

We used Kirby-Bauer disk diffusion to quantify phenotypic resistance of the WUSM *K. variicola* strains to several clinically relevant antibiotics (Figure 3.3.4b). *Klebsiella* species are generally considered intrinsically resistant to ampicillin due to a conserved β -lactamase gene. In our cohort, 3/55 isolates were unexpectedly susceptible to ampicillin while the rest were resistant. Despite phenotypic sensitivity to ampicillin, the genomes for WUSM_KV_25, WUSM_KV_34, and WUSM_KV_82 encode *bla*_{LEN-24}, *bla*_{LEN22}, and *bla*_{LEN-16}, respectively. These *bla*_{LEN} alleles were also found in isolates intermediate and resistant to ampicillin. As expected, WUSM_KV_55 was the only isolate resistant to both meropenem and ceftazidime, presumably due to carriage of *bla*_{KPC-2}. Additionally, it was the only isolate intermediate to ciprofloxacin. Four isolates

were resistant to trimethoprim-sulfamethoxazole, but only WUSM_KV_50 and



WUSM_KV_55 had identified ARGs that would explain this phenotype.

Review of 2017 composite antibiogram from a microbiology laboratory serving 5 hospitals in the St. Louis region (Missouri, USA), based on first isolate per patient per year, revealed that, in general, *K. pneumoniae* (n = 1522) had decreased susceptibility to all reported antimicrobials compared to *K. variicola* (n=144), except for meropenem (99% susceptibility for both species). Most notably, *K. pneumoniae* exhibited decreased susceptibility, as compared to *K. variicola*, with ampicillin-sulbactam (63 vs 93% susceptible), nitrofurantoin (66 vs 86% susceptible), and trimethoprim-sulfamethoxazole (80 vs 90% susceptible).

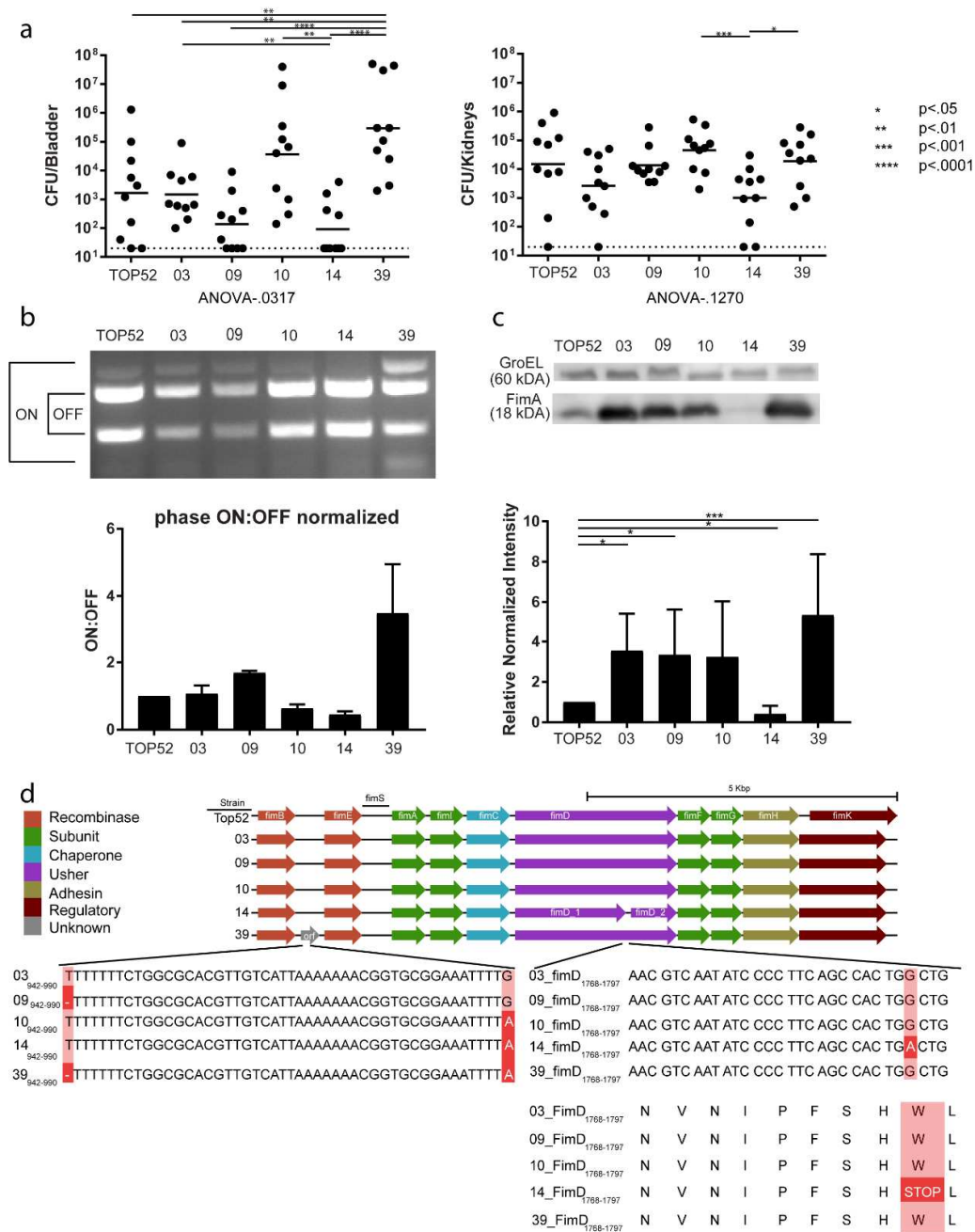
3.3.5 Changes in *fim* operon are associated with uropathogenicity in a murine UTI model

Given that 70% (39/56) of *K. variicola* strains from our cohort were isolated from the human urinary tract, we wanted to assess uropathogenicity in a diverse subset of these isolates. We transurethrally inoculated C3H/HeN mice with 10^7 CFU/ml of 5 individual *K. variicola* strains, or the model uropathogenic *K. pneumoniae* TOP52 strain, for comparison (Figure 3.3.5a) (3, 17, 18). Similar to previously published infections with *K. pneumoniae* TOP52, the *K. variicola* strains exhibited large variations in bacterial

Figure 3.3.4 WUSM *K. variicola* have a low burden of ARGs and are generally susceptible to antibiotics

(a) Network diagram depicting each WUSM_KV isolate and ARG as nodes. ARGs are colored in accordance with predicted phenotypic resistance from ResFinder, and WUSM_KV genomes are colored by the burden of ARGs. (b) Scatterplots depicting Kirby-Bauer disk diffusion size (mm) from phenotypic susceptibility testing. Each plot represents an isolate, and the plots are colored according to CLSI interpretation. Those with atypical resistance are listed by name with putative ARGs.

colony-forming units (CFUs) recovered from the bladder at 24 hours post infection (hpi). When compared to TOP52, WUSM_KV_39 was the only isolate with a significantly increased bladder burden ($P=0.0094$). Bacterial loads of WUSM_KV_10 and WUSM_KV_39 were both significantly higher than WUSM_KV_09 and WUSM_KV_14 (Figure 3.3.5a). Despite this variability among bladder CFU results, the results of kidney titers 24 hpi were not significantly different among strains by ANOVA ($P=0.1270$). As observed in the bladder, though, WUSM_KV_10 and WUSM_KV_39 achieved significantly higher kidney CFU compared to WUSM_KV_14.



Given the variation in bladder burden, we wanted to assess if differences in uropathogenicity could be related to expression of type 1 pili, a key virulence factor for UTI encoded by the *fim* operon(18, 19). In *K.*

pneumoniae and *Escherichia coli*,

expression of type 1 pili is controlled by a region of invertible DNA

Figure 3.3.5 Changes in *fim* operon are associated with outcomes in mouse UTI model

(a) CFU/bladder and CFU/kidney of *K. pneumoniae* TOP52 and WUSM_KV isolates 24 hour post transurethral bladder inoculation of C3H/HeN mice. Short bars represent geometric means of each group and dotted lines represent limits of detection. (b) *fimS* phase assay and quantification with respective bands indicating the “ON” and “OFF” position labeled. (c) Immunoblot for FimA and GroEL, with quantification shown below. (d) EasyFig illustration of genes in the *fim* operon and JALview of the nucleotides and amino acids for the *fimB/fimE* intergenic region and *fimD* gene.

(*fimS* site)(19, 20). Orientation of the *fimS* site in the “ON” position enables production of type 1 pili and increased urovirulence. Under identical growth conditions, WUSM_KV_39 had a higher population with the *fimS* promoter region in the “ON” orientation compared to the other strains tested (Figure 3.3.5b). Furthermore, consistent with its success in the bladder, WUSM_KV_39 was found to produce the greatest amount of FimA (the main structural component of type 1 pili), as measured by immunoblot (Figure 3.3.5c). WUSM_KV_03, WUSM_KV_09, and WUSM_KV_39 all produced significantly more FimA than *K. pneumoniae* TOP52. Interestingly, WUSM_KV_14 did not produce appreciable levels of FimA by this assay (Figure 3.3.5c).

As we discovered significant variability in type 1 piliation, we specifically investigated changes in *fim* operon sequence between these isolates by viewing the prokka coding sequence annotation in EasyFig and JALview (Figure 3.3.5d)(21, 22). We found that WUSM_KV_14 had a predicted truncated FimD usher sequence. A guanine-to-adenine single nucleotide polymorphism (SNP) in the *fimD* gene changed a predicted tryptophan residue into a premature stop codon, likely explaining the observed lack of production of type 1 pili. Additionally, in WUSM_KV_39, prokka annotated a hypothetical protein in the intergenic region between *fimB* and *fimE* and included a gap replacing a thymine and a guanine-to-adenine SNP. The altered *fimB/fimE* intergenic region in WUSM_KV_39 may play a role in its increased expression of type 1 pili. Together, these data demonstrate that variation exists amongst *K. variicola* genomes that may account for differential urinary tract niche proclivity among isolates.

3.3.6 *K. variicola* encodes both conserved and novel usher genes

The *fim* operon is one of the best characterized chaperone-usher pathways (CUP); given the observed importance of the *fim* operon in *K. variicola* uropathogenicity, we searched the pan-genome of our *K. variicola* cohort to identify the complete repertoire of CUP operons(23). 17 unique usher sequences at 95% identity were identified across the 55 WUSM *K. variicola* genomes, and an amino acid sequence alignment showed

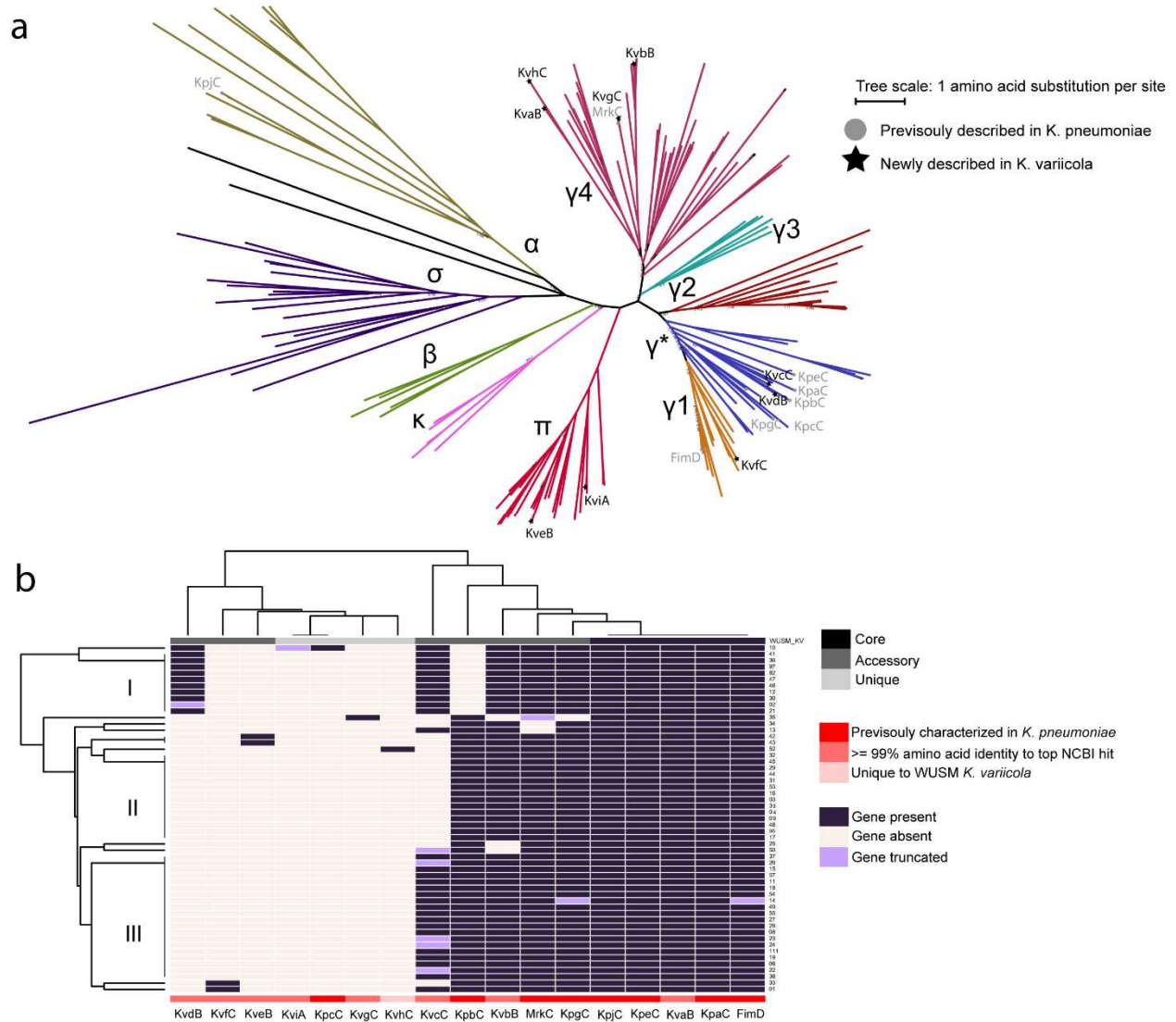


Figure 3.3.6 *K. variicola* encodes both conserved and novel usher genes

(a) Approximate-maximum-likelihood tree of the usher amino acid sequences described by Nuccio & Baumber and representatives of the 17 usher sequences identified in the WUSM_KV pan-genome. (b) Hierarchical clustering of the presence/absence matrix of each and annotation of relevant features related to each usher.

that they were distributed in 5 Nuccio & Baumber clades (Figure 3.3.6a)(24). From this analysis we discovered 9 new usher genes previously undescribed in *Klebsiella*, which we name *kva-kvi*. *KviA* and *KveB* usher sequences were found to cluster within the pi (π) clade, making them the first description of a P pilus apparatus in *Klebsiella*. The recently named γ^* subclade contained the greatest amount (7/17) of *K. variicola* usher

sequences; 5 of these 7 were previously reported in *K. pneumoniae*, while KvcC and KvdB are first reported here.

FimD, and the usher sequences for KpaC, KvaB, KpeC, and KpjC were present in all 55 WUSM *K. variicola* isolates (Figure 3.3.6b). KvgC, KvhC, KviA, and KpcC were each found in only one isolate. KpgC, MrkC, kvbC, KpbC, KvcC, KveB, KvfC, and KvdB can be considered accessory usher sequences in this cohort, as they were absent in certain strains. The most notable pattern evident from the hierarchical clustering of the presence/absence for all usher genes in our *K. variicola* cohort is that isolates WUSM_KV_10 through WUSM_KV_21 all carry the KvdB sequence but not KpbC.

Eight of the 9 newly described usher sequences had highest BLASTP hits $\geq 99\%$ identity across the entire length of the gene against the non-redundant protein sequences database in April 2018 and all of them were previously annotated as being found in *Enterobacteriaceae*, *Klebsiella*, or *K. variicola*. All of the usher genes except *kvi* were in operons that included a chaperone, at least one subunit, and a putative adhesin. KvhC, the usher protein with the lowest BLASTP identify value, had 76% identity to several genes from *Enterobacter* species. The contig with the *kvh* operon also contained several genes that had possible roles in prophage integration and transposase activity. Our results indicate that *K. variicola* strains harbor a diverse set of usher genes, which may augment *K. variicola* fitness across a variety of environmental niches, and these operons may be acquired from other *Enterobacteriaceae*.

3.4 Discussion

A previous phylogenomic study used split-network analysis to demonstrate that the *K. variicola* phylogroup (formerly KPIII) is distinct from *K. pneumoniae* (KPI) and *K. quasipneumoniae* (KPII) (25). As an orthogonal method, we used ANI software, the gold standard for *in silico* species delineation, to recreate this differentiation of phylogroups as separate species (8). Historically, differentiation between *K. pneumoniae* and *K. variicola* has been difficult, as evidenced by misannotation of *K. variicola* as *K. pneumoniae* in public genome sequence databases. These misannotated *K. variicola* strains came from a variety of geographic regions and were not exclusive to any cluster. Within our sequenced cohort, differentiation of *K. variicola* from *K. pneumoniae* and *K. quasipneumoniae* using MALDI-TOF MS and *yggE* PCR/RFLP was supported by ANI. This indicates that *yggE* PCR/RFLP (3) would be a feasible alternative for clinical labs across the globe lacking access to MALDI-TOF MS or WGS. Additionally, hierarchical clustering of the ANI values and core-genome phylogeny demonstrated that 2 *K. variicola* genomes were distinctly separate from the other 143 in our cohort. ANI values between these genomes with the other *K. variicola* genomes were ~96%, similar to what was observed for *K. quasipneumoniae*. The differences in ANI values contributed to the delineation of *K. quasipneumoniae* into two subspecies, *K. quasipneumoniae* subsp. *quasipneumoniae* and *K. quasipneumoniae* subsp. *similipneumoniae* (26). However, further phenotypic comparisons between FLLH01.1 and GCF_001548315 with other *K. variicola* isolates is required to unequivocally qualify these as separate subspecies. Further phenotypic comparisons, including the sole carbon source utilization used for differentiation of the *K. quasipneumoniae* subspecies,

between KvMX2/Yh43 and other *K. variicola* isolates is required to unequivocally qualify these as separate subspecies(26).

Numerous studies have shown that *K. pneumoniae* has a deep-branching phylogenetic structure with minimal recombination occurring within *K. pneumoniae* strains and between *K. pneumoniae* and *K. variicola*/*K. quasipneumoniae* (25, 27). Importantly, though, large-scale recombination events may be clinically relevant, as evidenced by research on the origin of the frequently carbapenem-resistant ST258 lineage (28, 29). Our results demonstrate that like *K. pneumoniae*, *K. variicola* shows minimal recombination within its genome, and its population structure is composed of numerous clades in a star-like phylogeny. A star-like population structure with deep-branching relationship between isolates (n=29 and n=28) was also found in two previously published *K. variicola* phylogenetic trees(2, 30).

Similar to our work, a previous investigation did not identify any geographic distinction when genomes from within the United States were compared to those from outside of the United States (2). The 6 genomes in cluster 23 with WUSM_KV_10 were from ICU patient samples in Seattle, Washington, which provides the first evidence of clonal groups responsible for *K. variicola* infections in some settings (31). Although they were closely related when compared against all *K. variicola* genomes, there was still 1,882 SNPs between WUSM_KV_10 and the other 6 genomes. Interestingly, clusters were not restricted to human infections, as Cluster 24 contains 3 genomes from bovine mastitis (NL49, NL58, NL58) and hospital isolates (VRCO0246, VRCO00242, VRCO00244, and VRCO00243) (<https://www.ncbi.nlm.nih.gov/bioproject/361595>)(32).

As expected for *K. variicola*, *bla*_{LEN} β -lactamases were the most conserved ARGs. A previous report unexpectedly found a *K. variicola* isolate that harbored the *bla*_{OKP} gene commonly found in *K. quasipneumoniae*; however, we did not identify such instances within our cohort (2). Although chromosomally encoded in *K. pneumoniae*, *fosA* was identified in only 1/145 of the *K. variicola* genomes (33, 34). Additionally, as previously found in *K. pneumoniae* clinical isolate cohorts we found *oqxAB* efflux pump genes widespread across *K. variicola* genomes (35-37). Although these genes may be ubiquitous in *K. variicola*, 0/55 of isolates we tested had resistance to ciprofloxacin; the single example with intermediate susceptibility carried a *qnrB6* gene. This is not atypical for *Enterobacteriaceae* possessing *oqxAB*, as one study found 100% prevalence of *oxqAB* in *K. pneumoniae* but no quinolone resistance (36). It is possible that for *K. variicola*, similar to *K. pneumoniae*, high expression of *oqxAB* is essential for phenotypic resistance to quinolones (35). In *K. pneumoniae*, expansion of clonal groups are associated with carbapenemase carriage (i.e. ST258 and *bla*_{KPC}) however we did not observe any associations between carbapenemase genes and *K. variicola* clusters. Indeed only 1.81% (1/55) of *K. variicola* within our institutional cohort had a carbapenemase gene and the regional resistance rate for meropenem between *K. pneumoniae* and *K. variicola* in 2017 was similar. *bla*_{NDM} positive *K. variicola* have been identified in clinical and environmental samples but *bla*_{KPC} positive genomes came exclusively from clinical sources. KPN1481 (*bla*_{NDM-1}) was annotated as a urine derived isolate but GJ1, GJ2, and GJ3 (all *bla*_{NDM-9}) were found in the Gwangju tributary in South Korea (2, 38). In contrast, WUSM_KV_55 (*bla*_{KPC-2}) was isolated from bronchoalveolar lavage fluid, KP007 (*bla*_{KPC-2}) from intraabdominal site, and 223/14

(*bla*_{KPC-6}) from a laparotomy wound(39, 40). IncF plasmids, the most abundant replicon identified in the *K. variicola* cohort are known carriers of antibiotic resistance genes, including *bla*_{CTX-M} and *bla*_{OXA} β -lactamases(41). Consistent with their widespread identification in *K. variicola*, IncF plasmids are frequently found in *K. pneumoniae* and *E. coli*(42, 43).

K. pneumoniae is the second leading cause of urinary tract infections (44). Given previous misclassification of *K. variicola* as *K. pneumoniae* and the high frequency at which *K. variicola* was isolated from the urinary tract, we were interested in comparing the uropathogenicity of our *K. variicola* isolates to the well-studied model *K. pneumoniae* TOP52 isolate (3, 17, 18). We identified strain-dependent virulence capacity, with UTI infections from WUSM_KV_39 yielding statistically significant higher bladder CFU than *K. pneumoniae* TOP52. Quantification of metrics used to study uropathogenicity in *E. coli* and *K. pneumoniae* show increased *fimS* in the “ON” orientation and increased FimA production by WUSM_KV_39; these findings provide a plausible explanation for why WUSM_KV_39 performed better than *K. pneumoniae* TOP52 and all WUSM_KV isolates excluding WUSM_KV_10 (45). While we do not yet understand the role of the putative protein identified between recombinases *fimB* and *fimE* in WUSM_KV_39, one could postulate that this difference may affect fimbrial expression. Additionally, the poorest performer in the urinary tract, WUSM_KV_14, encodes a mutation resulting in a truncated *fimD* usher sequence which likely explains its lack of FimA production. As with other bacterial pathogens, it is likely that specific virulence factors are required for *K. variicola* competency in distinct body niches (46, 47). Further work is therefore warranted to test if yersinibactin and allantoin utilization

promote lung and liver infections, respectively, in *K. variicola* as they do in *K. pneumoniae* (48-51).

K. variicola encodes usher genes previously identified in *K. pneumoniae* and 9 novel ushers (52). Interestingly, KveB and KviA are the first report of π usher proteins in *Klebsiella*. The best studied π operon, *pap* in *E. coli*, is a major contributor to pyelonephritis as the PapG adhesin can bind Gal- α (1–4)-Gal exposed on human kidney cells (53). Other usher genes have been shown to be essential for biofilm formation, plant cell adhesion, and murine gut colonization, further demonstrating their role in niche differentiation (52). Clustering of the presence/absence of these ushers showed the absence of KpbC but presence of KvdB in 11 of the WUSM_KV genomes, a phenomenon similar to that observed for UshC and YraJ in *E. coli*(54). All 4 of these usher types were found in the γ^* clade, suggesting an exclusionary form of functional redundancy between usher genes (54). Usher genes and CUP operons are frequently exchanged horizontally between *Enterobacteriaceae* genera (54). Indeed, we have found that the KvhC usher protein has only 76% amino acid identity to any existing proteins in the non-redundant protein sequence database and that the *kvh* operon is situated next to multiple prophage and transposase associated genes.

In this investigation, we used phenotypic and genomic analyses to better understand the diversity of *K. variicola* genomes, both from our institution and across the globe (using publicly available NCBI genomes). Then we assessed the functional consequences of ARGs and VGs towards antibiotic resistance and uropathogenicity. One limitation of our study is that our mouse infections and phenotypic analyses are performed with non-isogenic strains. If existing genetic modification systems in *K.*

pneumoniae are shown to be useful for gene knockouts in *K. variicola*, further work can be performed to mechanistically validate our findings. An additional limitation is that ~30 genomes of *K. variicola* have been uploaded to NCBI since we initiated our comparison. These may further elucidate differences in population structure, although even with almost 300 genomes, one study indicates that *K. pneumoniae* diversity remains under sampled (25).

Our work represents the first large-scale genomic analysis of *K. variicola* across multiple institutions and the first use of a murine model to study *K. variicola* pathogenesis. We unequivocally show that whole-genome comparisons can separate *K. variicola* from *K. pneumoniae* and offer convenient alternative methods for laboratories without access to WGS to differentiate these species. Importantly, we demonstrate that high-risk ARGs and VGs are present in *K. variicola* genomes from a variety of geographies. This may have clinical ramifications, as we demonstrate that some *K. variicola* clinical isolates can be superior uropathogens compared to *K. pneumoniae*. Similar to *E. coli* and *K. pneumoniae*, the diversity of CUP operons in these isolates could complement additional acquired virulence genes and enable infection of specific niches. Therefore, it is imperative that *K. variicola* and *K. pneumoniae* continue to be differentiated in the clinical laboratory, so that we may apply data on differential gene repertoire, clinical behavior, and niche specificity to the goal of ultimately improving patient outcomes.

3.5 Materials & Methods

3.5.1 Clinical *Klebsiella* Collection

113 clinical *Klebsiella* spp. isolates recovered in the Barnes-Jewish Hospital Microbiology laboratory (St. Louis, MO) from 2016-2017 were evaluated in this study. Of these, 56 were consecutively collected isolates identified by Bruker Biotyper MALDI-TOF MS as *K. variicola* (research-use only database v6). This identification was confirmed using a PCR/restriction fragment length polymorphism (RFLP) assay targeting the *yggE* gene (F: 5'-TGTTACTTAAATCGCCCTTACGGG-3'; R: 5'-CAGCGATCTGCAAAACGTCTACT-3'; restriction enzyme: BciVI) that was designed to distinguish *K. variicola* from *K. pneumoniae*. 94.6% (53/56) confirmed as *K. variicola* using the *yggE* PCR-RFLP assay.

The remaining 58 isolates were randomly selected from a banked collection of *K. pneumoniae* strains historically recovered from clinical specimens (29 from urine, 25 from blood, and 1 each abdominal wound, tracheal aspirate, bronchial washing, and bile). Each of these isolates underwent Bruker MALDI-TOF MS and *yggE* PCR/RFLP to confirm their identification. Five percent (5%; 3/58) confirmed as *K. variicola* using MALDI-TOF MS and the *yggE* PCR-RFLP assay.

3.5.2 Illumina Whole Genome Sequencing and publicly available *Klebsiella* genomes

Pure frozen stocks of the presumptive 113 *Klebsiella* isolates were plated on blood agar to isolate single colonies. ~10 colonies were suspended using a sterile cotton swab into water, and total genomic DNA was extracted using the Bacteremia Kit (Qiagen). 0.5 ng of DNA was used as input for sequencing libraries using the Nextera kit (Illumina) (55). Libraries were pooled and sequenced on an Illumina NextSeq 2500 High Output system

to obtain ~2.5 million 2×150 bp reads. Demultiplexed reads had Illumina adapters removed with trimmomatic v.36 and decontaminated with DeconSeq v0.4.3 (56, 57). Draft genomes were assembled with spades v3.11.0, and the scaffolds.fasta files were used as input for QUAST v 4.5 to measure the efficacy of assembly (58, 59). All contigs ≥ 500 bp in length were annotated for open reading frames with prokka v1.12 (60). The genomes have all been deposited to NCBI under BioProject [PRJNA473122](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA473122).

To increase the number of genomes for downstream analysis, 50 *K. variicola* genomes were obtained from NCBI genomes (<https://www.ncbi.nlm.nih.gov/genome/>) in September 2017. Additionally, as it is possible that previously sequenced *K. variicola* may be incorrectly described as *K. pneumoniae*, we submitted the complete genome of the *K. variicola* reference strain At-22 to NCBI BLASTN against the non-redundant nucleotide collection and the whole-genome shotgun sequence databases using default settings in September 2017. Using this method, we obtained 41 genomes of *K. pneumoniae* with the minimum observed query length of 38% at 99% identity. Given that the cohort of genomes analyzed in our study includes isolates initially misannotated, we refer to them as either the NCBI genome or assembly (<https://www.ncbi.nlm.nih.gov/assembly>) accession key. Sequenced and acquired isolates were analyzed using a variety of computational programs. *In silico* sequence typing was performed using mlst v2.11 (<https://github.com/tseemann/mlst>) and the BIGSdb database (bigsdatabases.org/klebsiella/klebsiella.html).

3.5.3 Antimicrobial Susceptibility Testing

K. variicola isolates underwent antimicrobial susceptibility testing per laboratory standard operating procedures using Kirby-Bauer disk diffusion on Mueller Hinton Agar

(BD BBL™ Mueller Hinton II Agar), in accordance with Clinical and Laboratory Standards Institute (CLSI) standards. Disk diffusion results were interpreted using CLSI *Enterobacteriaceae* disk diffusion breakpoints (CLSI. *Performance Standards for Antimicrobial Susceptibility Testing*. 27th ed. CLSI Supplement M100. Wayne, PA: Clinical and Laboratory Standards Institute; 2017). Briefly, 4-5 colonies from pure isolates were used to create a 0.5 McFarland suspension of the organism in sterile saline. A sterile, non-toxic cotton swab was dipped into the bacterial suspension, and a lawn of the organism was plated to Mueller-Hinton agar. Antimicrobial Kirby-Bauer disks were applied, and the plate was incubated at 35°C in room air for 16-24 h. The diameters of the zones of growth inhibition surrounding each antimicrobial disk were recorded in mm.

3.5.4 Mouse Urinary Tract Infections

Bacterial strains from our *K. variicola* cohort and *K. pneumoniae* TOP52 were used to inoculate 7- to 8-week-old female C3H/HeN mice (Envigo) by transurethral catheterization as previously described (17, 18, 61). The *K. variicola* strains were selected to encompass a range of genetically distinct isolates. WUSM_KV_03 and WUSM_KV_10 were specifically chosen as they contain the *all* and *ybt* operons, respectively. Static 20-mL cultures were started from freezer stocks, grown in Luria-Bertani (LB) broth at 37°C for 16 h, centrifuged for 5 min at 8,000 × *g*, and the resultant pellet was resuspended in phosphate-buffered saline (PBS) and diluted to approximately 4 × 10⁸ CFU/ml. Fifty mL of this suspension was used to infect each mouse with an inoculum of 2 × 10⁷ CFU/ml. Inocula were verified by serial dilution and plating. At 24 hpi, bladders and kidneys were aseptically harvested, homogenized in

sterile PBS via Bullet Blender (Next Advance) for 5 min, serially diluted and plated on LB agar. All animal procedures were approved by the Institutional Animal Care and Use Committee at Washington University School of Medicine.

3.5.5 Phase Assays

To determine the orientation of the *fimS* phase switch in *Klebsiella*, a phase assay was adapted as previously described (19). An 817 bp fragment including *fimS* was PCR amplified using *Taq* polymerase (Invitrogen) and the primers 5'-GGGACAGATACGCGTTTGAT-3' and 5'-GGCCTAACTGAACGGTTTGA-3' and then digested with *HinfI* (New England Biolabs). Digestion products were separated by electrophoresis on a 1% agarose gel. A phase-ON switch yields products of 605 and 212 bp, and a phase-OFF switch yields products of 496 and 321 bp.

3.5.6 FimA and GroEL Immunoblots

Acid-treated, whole-cell immunoblotting was performed as previously described using 1:2,000 rabbit anti-type 1 pilus and 1:500,000 rabbit anti-GroEL (Sigma-Aldrich) primary antibodies(62, 63). Amersham ECL horseradish peroxidase-linked donkey anti-rabbit IgG (GE Healthcare) secondary antibody (1:2,000) was applied, followed by application of Clarity enhanced chemiluminescence (ECL) substrate (Bio-Rad Laboratories). The membrane was developed and imaged using a ChemiDoc MP Imaging System (Bio-Rad Laboratories). Relative band intensities were quantified using Fiji (<https://fiji.sc/>) (64).

3.5.7 Statistics

CFU/bladder and CFU/kidney for both experimental replicates were used as input for ordinary one-way ANOVA to judge significance. Pairwise comparisons of CFU/bladder and CFU/kidney values were performed by using the nonparametric Mann-Whitney U test. Similarly, normalized quantifications of relative FimA amounts (FimA/GroEL) and *fimS* in “ON” position (*fimS* “ON”/*fimS* “OFF”) were compared using the Mann-Whitney U test. All P values <0.05 were considered significant, and all calculations were performed in GraphPad Prism v7.04.

3.6 Acknowledgments

We thank members of the Dantas lab for insightful discussions of the results and conclusions. This work is supported in part by awards to G.D. through the Edward Mallinckrodt, Jr. Foundation (Scholar Award), and from the National Institute of General Medical Sciences, the National Institute of Allergy and Infectious Diseases, and the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health (NIH) under award numbers R01GM099538, R01AI123394, and R01HD092414, respectively. Experiments performed by JT and DR used funding from the NIH (award K08-AI127714) and the Children’s Discovery Institute of Washington University and St. Louis Children’s Hospital. The authors would like to thank Center for Genome Sciences & Systems Biology staff Brian Koebbe and Eric Martin for operation of the High-Throughput Computing Facility. The authors additionally thank David Hunstad for constructive feedback during manuscript authoring. The authors additionally thank Center for Genome Sciences & Systems Biology staff Jessica Hoisington-Lopez and MariaLynn Jaeger for performing the Illumina sequencing and demultiplexing. RFP was supported by a NIGMS training grant through award T32

GM007067 (PI: James Skeath) and the Monsanto Excellence Fund graduate fellowship.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication

3.8 References

1. Rosenblueth M, Martinez L, Silva J, Martinez-Romero E. *Klebsiella variicola*, a novel species with clinical and plant-associated isolates. *Syst Appl Microbiol*. 2004;27(1):27-35. doi: 10.1078/0723-2020-00261. PubMed PMID: 15053318.
2. Long SW, Linson SE, Ojeda Saavedra M, Cantu C, Davis JJ, Brettin T, Olsen RJ. Whole-Genome Sequencing of Human Clinical *Klebsiella pneumoniae* Isolates Reveals Misidentification and Misunderstandings of *Klebsiella pneumoniae*, *Klebsiella variicola*, and *Klebsiella quasipneumoniae*. *mSphere*. 2017;2(4). doi: 10.1128/mSphereDirect.00290-17. PubMed PMID: 28776045; PMCID: PMC5541162.
3. Berry GJ, Loeffelholz MJ, Williams-Bouyer N. An Investigation into Laboratory Misidentification of a Bloodstream *Klebsiella variicola* Infection. *J Clin Microbiol*. 2015;53(8):2793-4. doi: 10.1128/JCM.00841-15. PubMed PMID: 26063851; PMCID: PMC4508421.
4. Maatallah M, Vading M, Kabir MH, Bakhrouf A, Kalin M, Naucner P, Brisse S, Giske CG. *Klebsiella variicola* is a frequent cause of bloodstream infection in the stockholm area, and associated with higher mortality compared to *K. pneumoniae*. *PLoS One*. 2014;9(11):e113539. doi: 10.1371/journal.pone.0113539. PubMed PMID: 25426853; PMCID: PMC4245126.

5. Andrade BG, de Veiga Ramos N, Marin MF, Fonseca EL, Vicente AC. The genome of a clinical *Klebsiella variicola* strain reveals virulence-associated traits and a pI9-like plasmid. *FEMS Microbiol Lett.* 2014;360(1):13-6. doi: 10.1111/1574-6968.12583. PubMed PMID: 25135672.
6. Martinez-Romero E, Rodriguez-Medina N, Beltran-Rojel M, Toribio-Jimenez J, Garza-Ramos U. *Klebsiella variicola* and *Klebsiella quasipneumoniae* with capacity to adapt to clinical and plant settings. *Salud Publica Mex.* 2018;60(1):29-40. doi: 10.21149/8156. PubMed PMID: 29689654.
7. Martin RM, Bachman MA. Colonization, Infection, and the Accessory Genome of *Klebsiella pneumoniae*. *Front Cell Infect Microbiol.* 2018;8:4. doi: 10.3389/fcimb.2018.00004. PubMed PMID: 29404282; PMCID: PMC5786545.
8. Richter M, Rossello-Mora R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A.* 2009;106(45):19126-31. doi: 10.1073/pnas.0906412106. PubMed PMID: 19855009; PMCID: PMC2776425.
9. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. Versatile and open software for comparing large genomes. *Genome Biol.* 2004;5(2):R12. doi: 10.1186/gb-2004-5-2-r12. PubMed PMID: 14759262; PMCID: PMC395750.
10. Richter M, Rossello-Mora R, Oliver Glockner F, Peplies J. JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. *Bioinformatics.* 2016;32(6):929-31. doi: 10.1093/bioinformatics/btv681. PubMed PMID: 26576653.

11. Long SW, Linson SE, Ojeda Saavedra M, Cantu C, Davis JJ, Brettin T, Olsen RJ. Whole-Genome Sequencing of a Human Clinical Isolate of the Novel Species *Klebsiella quasivariicola* sp. nov. *Genome Announc.* 2017;5(42). doi: 10.1128/genomeA.01057-17. PubMed PMID: 29051239; PMCID: PMC5646392.
12. Mostowy R, Croucher NJ, Andam CP, Corander J, Hanage WP, Marttinen P. Efficient Inference of Recent and Ancestral Recombination within Bacterial Populations. *Mol Biol Evol.* 2017;34(5):1167-82. doi: 10.1093/molbev/msx066. PubMed PMID: 28199698; PMCID: PMC5400400.
13. Cheng L, Connor TR, Siren J, Aanensen DM, Corander J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol Biol Evol.* 2013;30(5):1224-8. doi: 10.1093/molbev/mst028. PubMed PMID: 23408797; PMCID: PMC3670731.
14. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol.* 2014;15(11):524. doi: 10.1186/PREACCEPT-2573980311437212. PubMed PMID: 25410596; PMCID: PMC4262987.
15. Rose R, Lamers SL, Dollar JJ, Grabowski MK, Hodcroft EB, Ragonnet-Cronin M, Wertheim JO, Redd AD, German D, Laeyendecker O. Identifying Transmission Clusters with Cluster Picker and HIV-TRACE. *AIDS Res Hum Retroviruses.* 2017;33(3):211-8. doi: 10.1089/AID.2016.0205. PubMed PMID: 27824249; PMCID: PMC5333565.
16. Kleinheinz KA, Joensen KG, Larsen MV. Applying the ResFinder and VirulenceFinder web-services for easy identification of acquired antibiotic resistance and *E. coli* virulence genes in bacteriophage and prophage nucleotide sequences.

Bacteriophage. 2014;4(1):e27943. doi: 10.4161/bact.27943. PubMed PMID: 24575358; PMCID: PMC3926868.

17. Johnson JG, Spurbeck RR, Sandhu SK, Matson JS. Genome Sequence of *Klebsiella pneumoniae* Urinary Tract Isolate Top52. *Genome Announc.* 2014;2(4). doi: 10.1128/genomeA.00668-14. PubMed PMID: 24994806; PMCID: PMC4082006.

18. Rosen DA, Pinkner JS, Jones JM, Walker JN, Clegg S, Hultgren SJ. Utilization of an intracellular bacterial community pathway in *Klebsiella pneumoniae* urinary tract infection and the effects of FimK on type 1 pilus expression. *Infect Immun.* 2008;76(7):3337-45. doi: 10.1128/IAI.00090-08. PubMed PMID: 18411285; PMCID: PMC2446714.

19. Struve C, Bojer M, Krogfelt KA. Characterization of *Klebsiella pneumoniae* type 1 fimbriae by detection of phase variation during colonization and infection and impact on virulence. *Infect Immun.* 2008;76(9):4055-65. doi: 10.1128/IAI.00494-08. PubMed PMID: 18559432; PMCID: PMC2519443.

20. Abraham JM, Freitag CS, Clements JR, Eisenstein BI. An invertible element of DNA controls phase variation of type 1 fimbriae of *Escherichia coli*. *Proc Natl Acad Sci U S A.* 1985;82(17):5724-7. PubMed PMID: 2863818; PMCID: PMC390624.

21. Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer. *Bioinformatics.* 2011;27(7):1009-10. doi: 10.1093/bioinformatics/btr039. PubMed PMID: 21278367; PMCID: PMC3065679.

22. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. *Bioinformatics.*

2009;25(9):1189-91. doi: 10.1093/bioinformatics/btp033. PubMed PMID: 19151095; PMCID: PMC2672624.

23. Busch A, Waksman G. Chaperone-usher pathways: diversity and pilus assembly mechanism. *Philos Trans R Soc Lond B Biol Sci.* 2012;367(1592):1112-22. doi: 10.1098/rstb.2011.0206. PubMed PMID: 22411982; PMCID: PMC3297437.

24. Nuccio SP, Baumler AJ. Evolution of the chaperone/usher assembly pathway: fimbrial classification goes Greek. *Microbiol Mol Biol Rev.* 2007;71(4):551-75. doi: 10.1128/MMBR.00014-07. PubMed PMID: 18063717; PMCID: PMC2168650.

25. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, Jenney A, Connor TR, Hsu LY, Severin J, Brisse S, Cao H, Wilksch J, Gorrie C, Schultz MB, Edwards DJ, Nguyen KV, Nguyen TV, Dao TT, Mensink M, Minh VL, Nhu NT, Schultz C, Kuntaman K, Newton PN, Moore CE, Strugnell RA, Thomson NR. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. *Proc Natl Acad Sci U S A.* 2015;112(27):E3574-81. doi: 10.1073/pnas.1501049112. PubMed PMID: 26100894; PMCID: PMC4500264.

26. Brisse S, Passet V, Grimont PA. Description of *Klebsiella quasipneumoniae* sp. nov., isolated from human infections, with two subspecies, *Klebsiella quasipneumoniae* subsp. *quasipneumoniae* subsp. nov. and *Klebsiella quasipneumoniae* subsp. *similipneumoniae* subsp. nov., and demonstration that *Klebsiella singaporensis* is a junior heterotypic synonym of *Klebsiella variicola*. *Int J Syst Evol Microbiol.* 2014;64(Pt 9):3146-52. doi: 10.1099/ijss.0.062737-0. PubMed PMID: 24958762.

27. Moradigaravand D, Martin V, Peacock SJ, Parkhill J. Evolution and Epidemiology of Multidrug-Resistant *Klebsiella pneumoniae* in the United Kingdom and Ireland. *MBio*. 2017;8(1). doi: 10.1128/mBio.01976-16. PubMed PMID: 28223459; PMCID: PMC5358916.
28. Wyres KL, Gorrie C, Edwards DJ, Wertheim HF, Hsu LY, Van Kinh N, Zadoks R, Baker S, Holt KE. Extensive Capsule Locus Variation and Large-Scale Genomic Recombination within the *Klebsiella pneumoniae* Clonal Group 258. *Genome Biol Evol*. 2015;7(5):1267-79. doi: 10.1093/gbe/evv062. PubMed PMID: 25861820; PMCID: PMC4453057.
29. Chen L, Mathema B, Pitout JD, DeLeo FR, Kreiswirth BN. Epidemic *Klebsiella pneumoniae* ST258 is a hybrid strain. *MBio*. 2014;5(3):e01355-14. doi: 10.1128/mBio.01355-14. PubMed PMID: 24961694; PMCID: PMC4073492.
30. Gorrie CL, Mirceta M, Wick RR, Edwards DJ, Thomson NR, Strugnell RA, Pratt NF, Garlick JS, Watson KM, Pilcher DV, McGloughlin SA, Spelman DW, Jenney AWJ, Holt KE. Gastrointestinal Carriage Is a Major Reservoir of *Klebsiella pneumoniae* Infection in Intensive Care Patients. *Clin Infect Dis*. 2017;65(2):208-15. doi: 10.1093/cid/cix270. PubMed PMID: 28369261; PMCID: PMC5850561.
31. Roach DJ, Burton JN, Lee C, Stackhouse B, Butler-Wu SM, Cookson BT, Shendure J, Salipante SJ. A Year of Infection in the Intensive Care Unit: Prospective Whole Genome Sequencing of Bacterial Clinical Isolates Reveals Cryptic Transmissions and Novel Microbiota. *PLoS Genet*. 2015;11(7):e1005413. doi: 10.1371/journal.pgen.1005413. PubMed PMID: 26230489; PMCID: PMC4521703.

32. Davidson FW, Whitney HG, Tahlan K. Genome Sequences of *Klebsiella variicola* Isolates from Dairy Animals with Bovine Mastitis from Newfoundland, Canada. *Genome Announc.* 2015;3(5). doi: 10.1128/genomeA.00938-15. PubMed PMID: 26358587; PMCID: PMC4566169.
33. Guo Q, Tomich AD, McElheny CL, Cooper VS, Stoesser N, Wang M, Sluis-Cremer N, Doi Y. Glutathione-S-transferase FosA6 of *Klebsiella pneumoniae* origin conferring fosfomycin resistance in ESBL-producing *Escherichia coli*. *J Antimicrob Chemother.* 2016;71(9):2460-5. doi: 10.1093/jac/dkw177. PubMed PMID: 27261267; PMCID: PMC4992852.
34. Ito R, Mustapha MM, Tomich AD, Callaghan JD, McElheny CL, Mettus RT, Shanks RMQ, Sluis-Cremer N, Doi Y. Widespread Fosfomycin Resistance in Gram-Negative Bacteria Attributable to the Chromosomal *fosA* Gene. *MBio.* 2017;8(4). doi: 10.1128/mBio.00749-17. PubMed PMID: 28851843; PMCID: PMC5574708.
35. Rodriguez-Martinez JM, Diaz de Alba P, Briales A, Machuca J, Lossa M, Fernandez-Cuenca F, Rodriguez Bano J, Martinez-Martinez L, Pascual A. Contribution of OqxAB efflux pumps to quinolone resistance in extended-spectrum-beta-lactamase-producing *Klebsiella pneumoniae*. *J Antimicrob Chemother.* 2013;68(1):68-73. doi: 10.1093/jac/dks377. PubMed PMID: 23011289.
36. Perez F, Rudin SD, Marshall SH, Coakley P, Chen L, Kreiswirth BN, Rather PN, Hujer AM, Toltzis P, van Duin D, Paterson DL, Bonomo RA. OqxAB, a quinolone and olaquinox efflux pump, is widely distributed among multidrug-resistant *Klebsiella pneumoniae* isolates of human origin. *Antimicrob Agents Chemother.* 2013;57(9):4602-3. doi: 10.1128/AAC.00725-13. PubMed PMID: 23817374; PMCID: PMC3754307.

37. Yuan J, Xu X, Guo Q, Zhao X, Ye X, Guo Y, Wang M. Prevalence of the *oqxAB* gene complex in *Klebsiella pneumoniae* and *Escherichia coli* clinical isolates. *J Antimicrob Chemother.* 2012;67(7):1655-9. doi: 10.1093/jac/dks086. PubMed PMID: 22438434.
38. Di DY, Jang J, Unno T, Hur HG. Emergence of *Klebsiella variicola* positive for NDM-9, a variant of New Delhi metallo-beta-lactamase, in an urban river in South Korea. *J Antimicrob Chemother.* 2017;72(4):1063-7. doi: 10.1093/jac/dkw547. PubMed PMID: 28087584.
39. Cienfuegos-Gallet AV, Chen L, Kreiswirth BN, Jimenez JN. Colistin Resistance in Carbapenem-Resistant *Klebsiella pneumoniae* Mediated by Chromosomal Integration of Plasmid DNA. *Antimicrob Agents Chemother.* 2017;61(8). doi: 10.1128/AAC.00404-17. PubMed PMID: 28507118; PMCID: PMC5527652.
40. Ahmad N, Chong TM, Hashim R, Shukor S, Yin WF, Chan KG. Draft Genome of Multidrug-Resistant *Klebsiella pneumoniae* 223/14 Carrying KPC-6, Isolated from a General Hospital in Malaysia. *J Genomics.* 2015;3:97-8. doi: 10.7150/jgen.13910. PubMed PMID: 26816553; PMCID: PMC4716803.
41. Carattoli A. Resistance plasmid families in Enterobacteriaceae. *Antimicrob Agents Chemother.* 2009;53(6):2227-38. doi: 10.1128/AAC.01707-08. PubMed PMID: 19307361; PMCID: PMC2687249.
42. Dolejska M, Villa L, Dobiasova H, Fortini D, Feudi C, Carattoli A. Plasmid content of a clinically relevant *Klebsiella pneumoniae* clone from the Czech Republic producing CTX-M-15 and QnrB1. *Antimicrob Agents Chemother.* 2013;57(2):1073-6. doi: 10.1128/AAC.01886-12. PubMed PMID: 23229477; PMCID: PMC3553734.

43. Shin J, Choi MJ, Ko KS. Replicon sequence typing of IncF plasmids and the genetic environments of blaCTX-M-15 indicate multiple acquisitions of blaCTX-M-15 in *Escherichia coli* and *Klebsiella pneumoniae* isolates from South Korea. *J Antimicrob Chemother.* 2012;67(8):1853-7. doi: 10.1093/jac/dks143. PubMed PMID: 22566590.
44. Flores-Mireles AL, Walker JN, Caparon M, Hultgren SJ. Urinary tract infections: epidemiology, mechanisms of infection and treatment options. *Nat Rev Microbiol.* 2015;13(5):269-84. doi: 10.1038/nrmicro3432. PubMed PMID: 25853778; PMCID: PMC4457377.
45. Schwan WR, Ding H. Temporal Regulation of fim Genes in Uropathogenic *Escherichia coli* during Infection of the Murine Urinary Tract. *J Pathog.* 2017;2017:8694356. doi: 10.1155/2017/8694356. PubMed PMID: 29445547; PMCID: PMC5763102.
46. Chmiela M, Miszczyk E, Rudnicka K. Structural modifications of *Helicobacter pylori* lipopolysaccharide: an idea for how to live in peace. *World J Gastroenterol.* 2014;20(29):9882-97. doi: 10.3748/wjg.v20.i29.9882. PubMed PMID: 25110419; PMCID: PMC4123370.
47. Hill C. Virulence or niche factors: what's in a name? *J Bacteriol.* 2012;194(21):5725-7. doi: 10.1128/JB.00980-12. PubMed PMID: 22821969; PMCID: PMC3486107.
48. Lawlor MS, O'Connor C, Miller VL. Yersiniabactin is a virulence factor for *Klebsiella pneumoniae* during pulmonary infection. *Infect Immun.* 2007;75(3):1463-72. doi: 10.1128/IAI.00372-06. PubMed PMID: 17220312; PMCID: PMC1828572.

49. Bachman MA, Oyler JE, Burns SH, Caza M, Lepine F, Dozois CM, Weiser JN. *Klebsiella pneumoniae* yersiniabactin promotes respiratory tract infection through evasion of lipocalin 2. *Infect Immun*. 2011;79(8):3309-16. doi: 10.1128/IAI.05114-11. PubMed PMID: 21576334; PMCID: PMC3147564.
50. Chou HC, Lee CZ, Ma LC, Fang CT, Chang SC, Wang JT. Isolation of a chromosomal region of *Klebsiella pneumoniae* associated with allantoin metabolism and liver infection. *Infect Immun*. 2004;72(7):3783-92. doi: 10.1128/IAI.72.7.3783-3792.2004. PubMed PMID: 15213119; PMCID: PMC427404.
51. Compain F, Babosan A, Brisse S, Genel N, Audo J, Ailloud F, Kassis-Chikhani N, Arlet G, Decre D. Multiplex PCR for detection of seven virulence factors and K1/K2 capsular serotypes of *Klebsiella pneumoniae*. *J Clin Microbiol*. 2014;52(12):4377-80. doi: 10.1128/JCM.02316-14. PubMed PMID: 25275000; PMCID: PMC4313302.
52. Khater F, Balestrino D, Charbonnel N, Dufayard JF, Brisse S, Forestier C. In silico analysis of usher encoding genes in *Klebsiella pneumoniae* and characterization of their role in adhesion and colonization. *PLoS One*. 2015;10(3):e0116215. doi: 10.1371/journal.pone.0116215. PubMed PMID: 25751658; PMCID: PMC4353729.
53. Verger D, Bullitt E, Hultgren SJ, Waksman G. Crystal structure of the P pilus rod subunit PapA. *PLoS Pathog*. 2007;3(5):e73. doi: 10.1371/journal.ppat.0030073. PubMed PMID: 17511517; PMCID: PMC1868955.
54. Stubenrauch CJ, Dougan G, Lithgow T, Heinz E. Constraints on lateral gene transfer in promoting fimbrial usher protein diversity and function. *Open Biol*. 2017;7(11). doi: 10.1098/rsob.170144. PubMed PMID: 29142104; PMCID: PMC5717340.

55. Baym M, Kryazhimskiy S, Lieberman TD, Chung H, Desai MM, Kishony R. Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS One*. 2015;10(5):e0128036. doi: 10.1371/journal.pone.0128036. PubMed PMID: 26000737; PMCID: PMC4441430.
56. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-20. doi: 10.1093/bioinformatics/btu170. PubMed PMID: 24695404; PMCID: PMC4103590.
57. Schmieder R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS One*. 2011;6(3):e17288. doi: 10.1371/journal.pone.0017288. PubMed PMID: 21408061; PMCID: PMC3052304.
58. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012;19(5):455-77. doi: 10.1089/cmb.2012.0021. PubMed PMID: 22506599; PMCID: PMC3342519.
59. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29(8):1072-5. doi: 10.1093/bioinformatics/btt086. PubMed PMID: 23422339; PMCID: PMC3624806.
60. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30(14):2068-9. doi: 10.1093/bioinformatics/btu153. PubMed PMID: 24642063.
61. Mulvey MA, Lopez-Boado YS, Wilson CL, Roth R, Parks WC, Heuser J, Hultgren SJ. Induction and evasion of host defenses by type 1-piliated uropathogenic *Escherichia coli*. *Science*. 1998;282(5393):1494-7. PubMed PMID: 9822381.

62. Garofalo CK, Hooton TM, Martin SM, Stamm WE, Palermo JJ, Gordon JI, Hultgren SJ. *Escherichia coli* from urine of female patients with urinary tract infections is competent for intracellular bacterial community formation. *Infect Immun*. 2007;75(1):52-60. doi: 10.1128/IAI.01123-06. PubMed PMID: 17074856; PMCID: PMC1828379.
63. Pinkner JS, Remaut H, Buelens F, Miller E, Aberg V, Pemberton N, Hedenstrom M, Larsson A, Seed P, Waksman G, Hultgren SJ, Almqvist F. Rationally designed small compounds inhibit pilus biogenesis in uropathogenic bacteria. *Proc Natl Acad Sci U S A*. 2006;103(47):17897-902. doi: 10.1073/pnas.0606795103. PubMed PMID: 17098869; PMCID: PMC1693844.
64. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B, Tinevez JY, White DJ, Hartenstein V, Eliceiri K, Tomancak P, Cardona A. Fiji: an open-source platform for biological-image analysis. *Nat Methods*. 2012;9(7):676-82. doi: 10.1038/nmeth.2019. PubMed PMID: 22743772; PMCID: PMC3855844.

Chapter 4: Spatiotemporal dynamics of multidrug resistant bacteria on intensive care unit surfaces

4.1 Abstract

Bacterial pathogens that infect hospital patients also contaminate hospital surfaces. These surface contaminants impact hospital infection control and epidemiology, but spatial, temporal, and phylogenetic relationships of these diverse bacteria are still under exploration. We investigate spatiotemporal and phylogenetic relationships of multidrug resistant bacteria on intensive care unit surfaces from two hospitals in the United States and Pakistan

collected over a year. These bacteria include common nosocomial pathogens, rare opportunistic pathogens, and novel taxa. Most of our common nosocomial isolates are dominated by single lineages composed of different clones, are phenotypically multidrug resistant, and have high resistance gene burdens. Many resistance genes are shared by multiple species and are flanked by mobilization elements. With permutation testing we identify *Acinetobacter baumannii* and *Enterococcus faecium* co-association, and our *in vitro* experiments find supporting synergistic biofilm interactions. Our results highlight drug resistant nosocomial pathogen burdens in hospital built-environments, provide evidence for spatiotemporal dependent transmission, and demonstrate a potential mechanism for dual-species bacterial surface persistence.

4.2 Introduction

Global treatment of bacterial infections is increasingly compromised by evolution and transmission of multidrug resistant organisms (MDROs) and their antibiotic

resistance genes (ARGs) between multiple habitats(1). Infections caused by MDROs are associated with increased mortality risk compared to infections by matched species susceptible isolates(2-4). Through international travel, clonal expansion, and promiscuous mobile genetic elements, MDROs and the ARGs they harbor have rapidly swept across the globe(1, 5-11). Resistant infections cause over 23,000 annual deaths in the United States of America (USA) and cost the economy over 55 billion dollars(12). The annual global death toll from MDROs is at least 700,000 people(13). Improved surveillance and understanding of MDRO and ARG transmission are key factors in reducing these death tolls(1).

Hospitalized patients are more vulnerable to bacterial infections than the general population(14), and healthcare associated infections (HAIs) acutely threaten patient safety worldwide(15, 16). The 'ESKAPE' pathogens, named by the Infectious Disease Society of America, are common causes of HAIs and the most common MDROs(17). These include the gram-positive microorganisms *Enterococcus* spp. and *Staphylococcus aureus*, and the gram-negative microorganisms *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterobacter* spp.(17). These ESKAPE pathogens can be acquired while hospitalized, but some patients may be colonized or infected prior to hospital admission(18). Patients harboring these putative pathogens can transmit these bacteria to healthcare workers, other patients, medical equipment, and hospital surfaces(18), but the contribution of this contamination route compared to other routes is unknown. The presence of these microorganisms on surfaces in healthcare settings is a local and global public health concern(19). Some putatively pathogenic strains of bacteria persist for months on hospital surfaces, and they may even survive

surface decontamination efforts, partly aided by biofilm formation(20-23). Though studies clearly demonstrate that bacterial pathogens exist on hospital surfaces, key knowledge gaps exist regarding the levels, types, and dynamics of contamination in hospitals from different geographies(14, 18). Specifically, there is a lack of information on the spatial, temporal, and phylogenetic relationships between different bacterial taxa on surfaces from countries endemic for a high burden of ARGs. This information gap is especially true for physical colocalization and horizontal gene transfer between clinically relevant ESKAPE pathogens and benign environmental bacteria.

Monitoring high contact surfaces for clinically relevant pathogenic bacteria and understanding the dynamics of their persistence and spread is one approach to thwart MDRO transmission and protect vulnerable hospitalized patients(24). Additionally, such surveillance provides an opportunity to identify and characterize potential emerging pathogens before they are recognized in clinical infections(12, 25).

To address the question of MDRO spatiotemporal dynamics and persistence on healthcare surfaces we conducted a year-long longitudinal study at a tertiary care hospital in Pakistan (PAK-H) where endemic ARG burden is high(26-28). A previous investigation found differing resistance mechanisms to last-resort carbapenem antibiotics in genetically similar *Enterobacteriaceae* strains and plasmids isolated from hospitals in Pakistan and the USA(29). Accordingly, we included a matched tertiary care hospital in the USA (USA-H) as a comparison group. For our collections and subsequent analysis, we took an Eulerian approach by selecting and measuring fixed hospital surfaces over time to understand bacterial contamination dynamics. This approach allows us to leverage collection time information and surface spatial information to draw epidemiological

insights. In both hospitals, we sampled 4 intensive care unit (ICU) rooms with 5 surfaces in each room (Figure 4.3.1). We collected surface swabs every other week for 3 months, and again at 6 months, and at 1 year, for a total of 180 samples per hospital. We identified high burdens of known MDROs on PAK-H ICU surfaces including ESKAPE pathogens and novel taxa(30). This investigation is the first from Pakistan to show such widespread contamination with multidrug resistant, extensively drug resistant, and pan-drug resistant bacteria. We found evidence that bacteria are non-randomly distributed on hospital surfaces with respect to both space and time, and we used this information to narrow possible contamination routes. We found cross-contamination of MDRO clones both across different surfaces within rooms, as well as between rooms at the same sampling time-points. From our results, it is likely that bacteria are seeded to hospital surfaces from diverse human and/or environmental reservoirs in a time dependent manner. These seedings result in waves of contamination that are often, but not always restricted to a single collection time. We show high numbers of ARGs are shared between common nosocomial pathogens and rarer bacterial species, including several novel taxa which are close phylogenetic relatives to nosocomial pathogens. Co-association analysis of *A. baumannii* and *E. faecium* led us to identify synergistic biofilm formation between these two ESKAPE pathogens. This discovery points to a possible explanation of bacterial persistence on hospital surfaces. Longitudinal persistence of these high impact pathogenic species alongside highly resistant bacteria classically identified as "environmental" paints a concerning picture of hospital surface contamination. These results lay groundwork for future surveillance efforts and infection control interventions to reduce healthcare associated bacterial surface contamination.

4.3 Results

4.3.1 PAK-H ICU surfaces had high bacterial burden

We recovered 1163 bacterial isolates from hospital surfaces in PAK-H and predicted their species identities by MALDI-TOF MS. We chose a subset of 289 unique isolates for phenotypic and genomic analysis, using the criterion of a single isolate per unique MALDI-TOF MS identified species per culture condition per surface per time-point. These 289 bacteria represent 31 species and 10 families (Figure 4.3.2a). 25.9% (75/289) of isolates recovered from PAK-H were identified as *A. baumannii*. 16.2% (47/289) were the gram-positive pathogen *E. faecium*, and 11.8% (34/289) were *K. pneumoniae*. Interestingly, similar numbers of the soil-associated opportunistic pathogen *Pseudomonas stutzeri* were recovered (28/289, 9.7%) as the common nosocomial pathogen *P. aeruginosa* (27/289, 9.3%). In addition to these expected nosocomial organisms, we identified a variety of other clinically relevant species such as

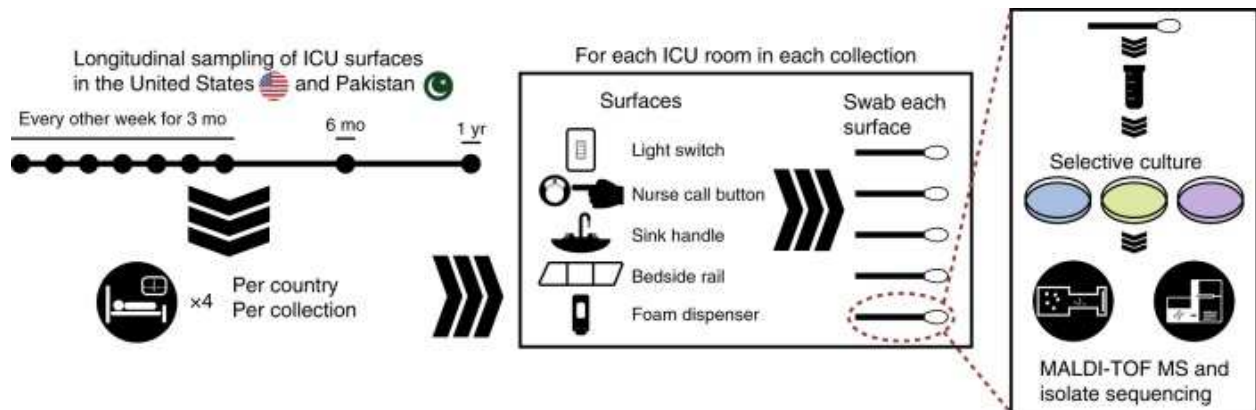


Figure 4.3.1 Bacterial isolate taxonomic identification and location.

Samples were collected from surfaces longitudinally over the course of 1 year from PAK-H ICU and USA-H ICU. Four rooms from each ICU were chosen for sampling and five surfaces within each room were surveyed for every collection time. Bacteria were cultured from the collection swabs, identified by MALDI-TOF MS, and then whole-genome sequenced.

Stenotrophomonas maltophilia, *Shewanella putrefaciens*, and *Providencia rettgeri*. These results starkly contrast with USA-H, where we only recovered 6 unique isolates which MALDI-TOF MS identified as *A. baumannii* (4/6) and *E. coli* (2/6) (Figure 4.3.2a). The majority of PAK-H (156/180, 86.7%) surface collections yielded bacteria (Figure 4.3.2b), but only a few (6/180, 3.3%) USA-H surface collections yielded isolates using the same culture conditions.

4.3.2 Sequence based bacterial identification outperformed MALDI-TOF MS

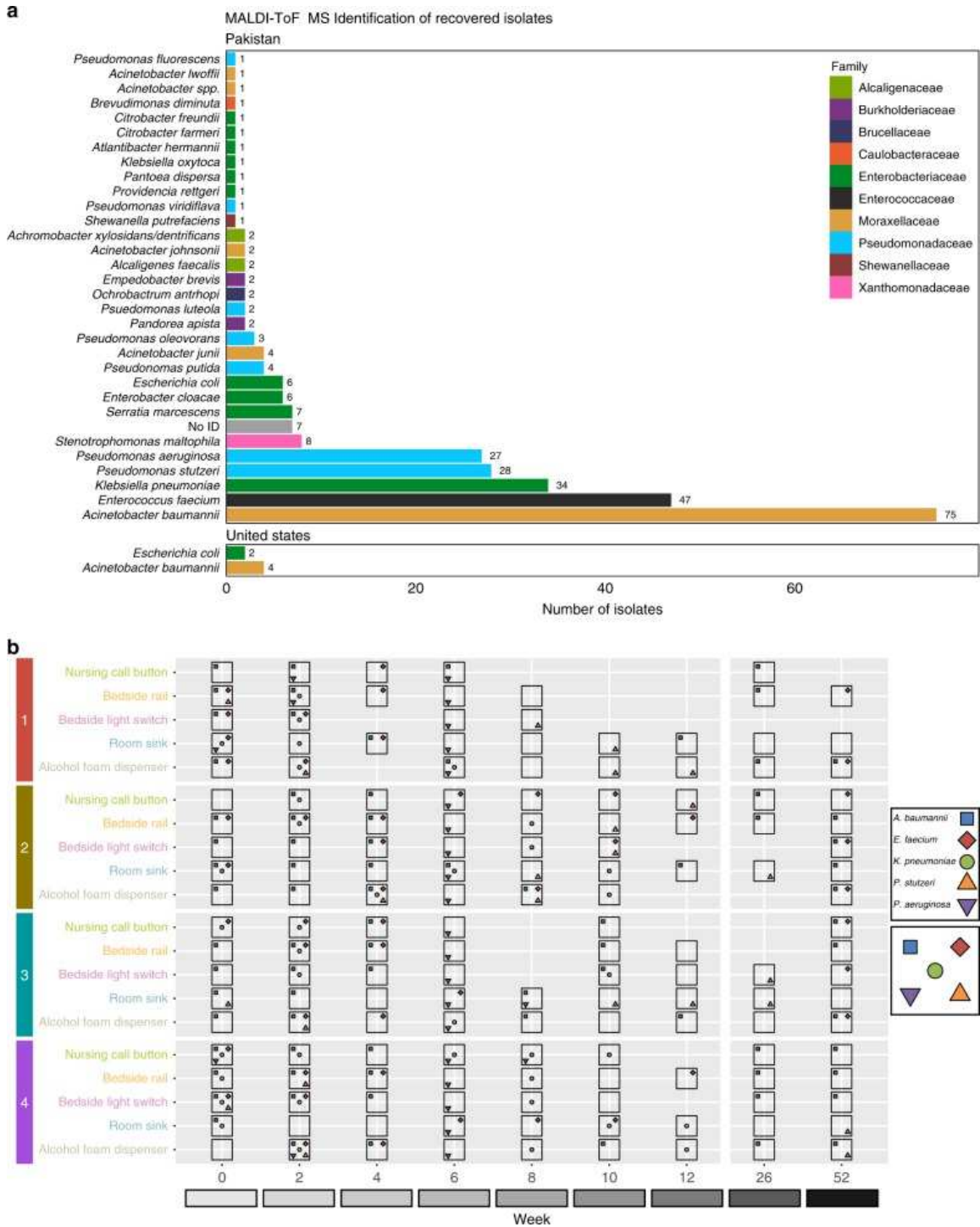


Figure 4.3.2 MALDI-TOF Identification and distribution

a MALDI-TOF MS identifications of bacterial isolates recovered from surfaces at PAK-H (above) and USA-H (below), colored by family. **b** Overview of PAK-H bacterial surface collections. Each horizontal gray panel represents a PAK-H room. Large, open black boxes are around any surface where one or more bacteria were collected. Blue squares are *A. baumannii*, red diamonds are *E. faecium*, green circles are *K. pneumoniae*, orange triangles are *P. stutzeri*, and purple triangles are *P. aeruginosa*.

We performed draft Illumina whole genome sequencing (WGS) on the 289 isolates to improve taxonomic resolution, quantify transmission dynamics for abundantly recovered organisms, and analyze ARG content. Initially, we constructed a Hadamard matrix, which represents the product of the average nucleotide identity (ANI) and percent of the genome aligned, between every pairwise combination of the 289 genomes sequenced from PAK-H surfaces. Hierarchical clustering of Hadamard values confirms 74/75 isolates identified by MALDI-TOF MS as *A. baumannii*, 47/47 as *E. faecium*, 33/34 as *K. pneumoniae*, 27/27 as *P. aeruginosa*, and 24/28 as *P. stutzeri*. These isolates cluster into the first 5 blocks. Analysis of the clustering pattern in the *K. pneumoniae* group found one isolate distant from the rest of the cohort; separate ANI analysis demonstrated this isolate is *Klebsiella quasipneumoniae*. Similarly, 3 isolates annotated as *P. stutzeri* are *Pseudomonas xanthomarina*. The isolate identified as *A. baumannii* that did not cluster with the rest of the cohort was *Acinetobacter soli*. In total, we found 27 cases where initial MALDI-TOF MS identifications differed from subsequent WGS dependent identifications. Additionally, both (2/2) isolates initially identified as *Empedobacter brevis* are *Empedobacter falsenii*. 2/3 of genomically confirmed *Atlantibacter subterranea* were unidentified by MALDI-TOF MS but 1/3 was identified as the closely related *Atlantibacter hermanii*.

We found 12 instances where genomes did not have $\geq 95\%$ ANI with the identified MALDI-TOF MS hit or the most closely related genomes as determined by 16S rRNA gene sequence in the EzBioCloud database, indicating that these are putative novel genomospecies. A separate investigation found that 2/7 of the isolates unidentified by MALDI-TOF MS are a new genus of multidrug resistant *Enterobacteriaceae*, termed *Superficieibacter electus*(30). The previously unreported genomospecies come from the

Caulobacteriaceae, *Xanthomonadaceae*, and *Enterobacteriaceae* families, and 5 of the proposed new genomospecies are *Pseudomonadaceae*. Importantly, these unreported genomospecies are found on the same healthcare surfaces as common human

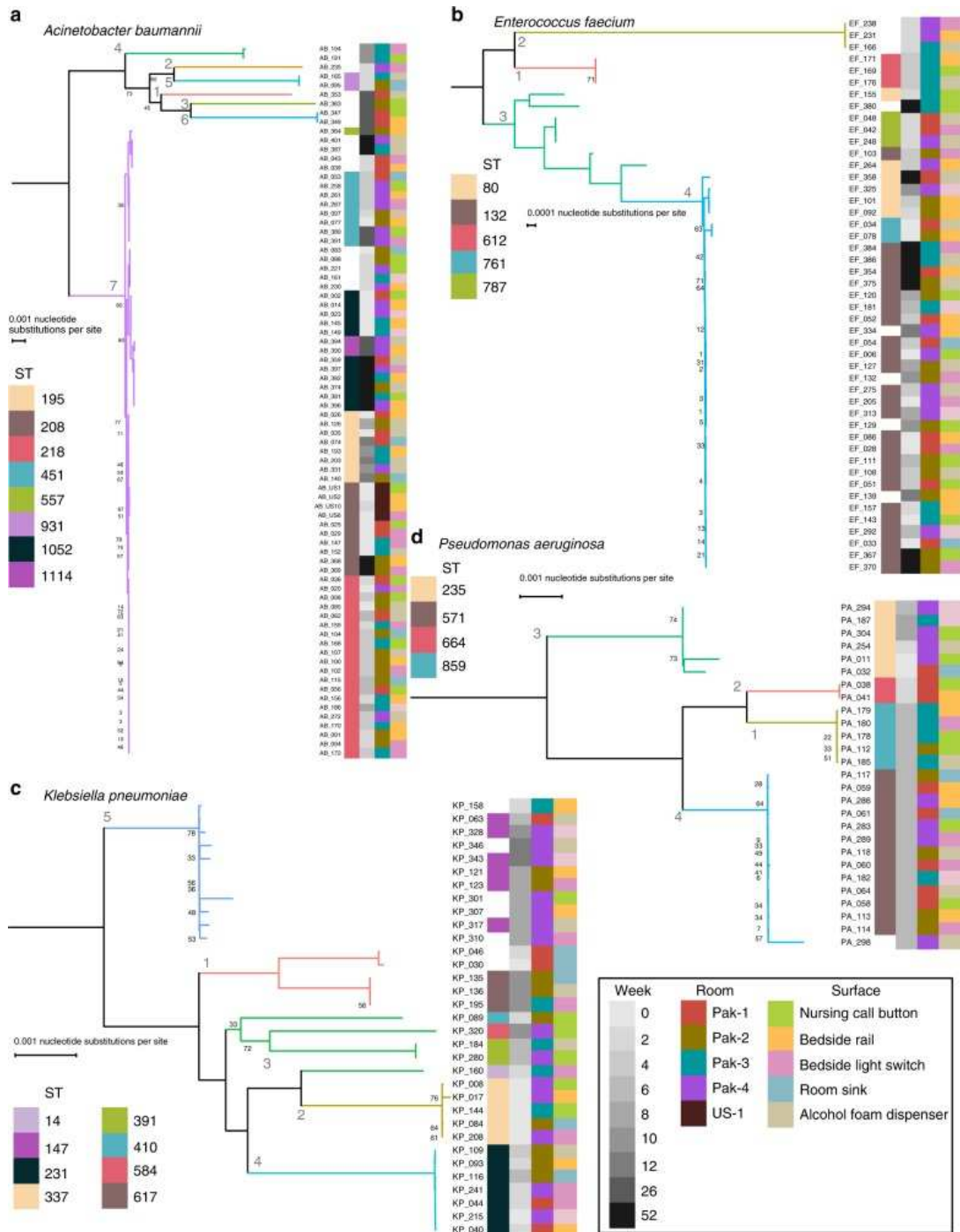


Figure 4.3.3 Phylogenetic trees of high abundance species from core genome alignments.

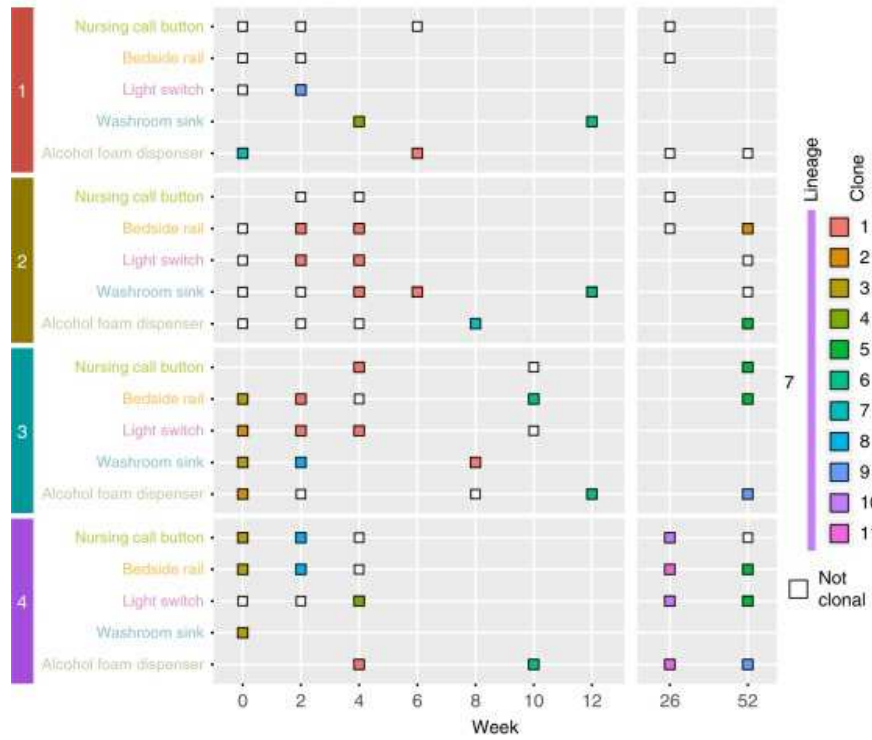
Maximum likelihood phylogenetic trees from core genome alignments of *A. baumannii* (a), *E. faecium* (b), *K. pneumoniae* (c), and *P. aeruginosa* (d). Tree branches are colored by hierBAPS lineage and these lineages are colored in subsequent figures. Sequence type, week, room, and surface are annotated as colored bars next to the isolate number. Week is given as grayscale with darker values corresponding to later weeks. The US room that yielded isolates is annotated dark brown

pathogens. Our results indicate WGS offers improved resolution for species delineation

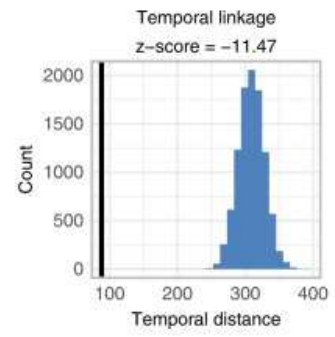
compared to conventional clinical diagnostic tools, for both common human pathogens and rarer species.

4.3.3 Single lineages dominated *A. baumannii* and *E. faecium* populations

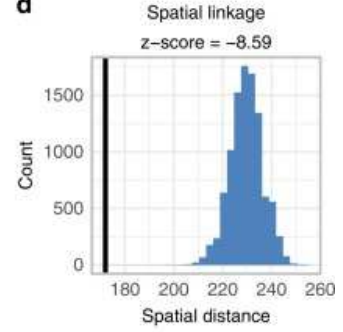
a *Acinetobacter baumannii* clonality results



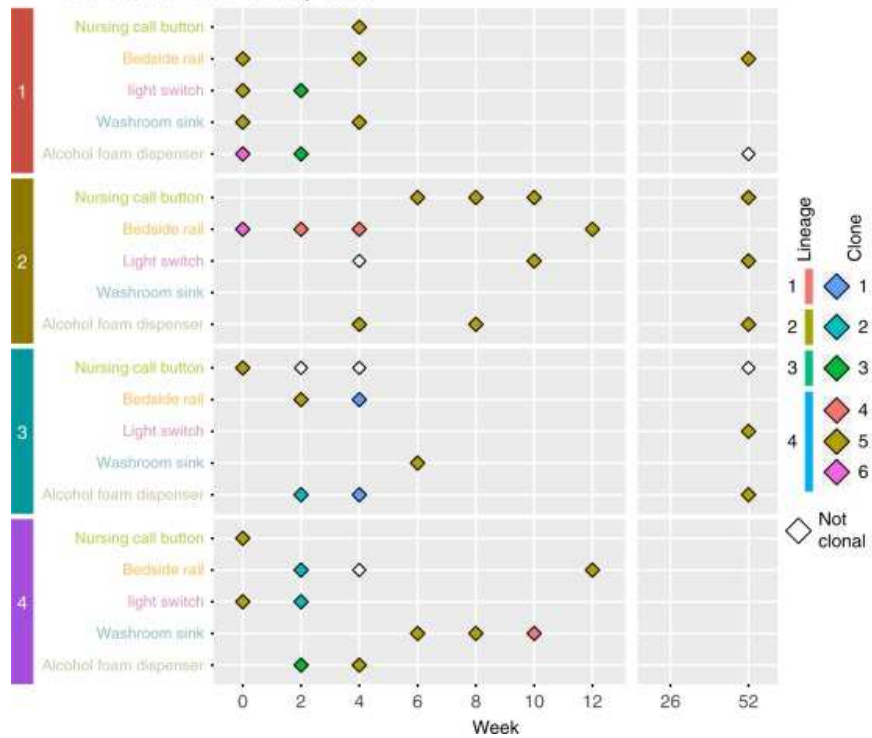
c



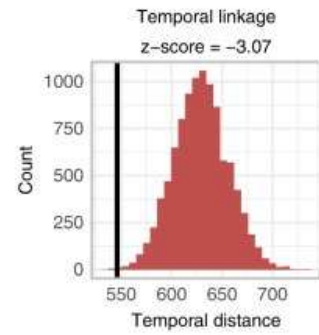
d



b *Enterococcus faecium* clonality results



e



f

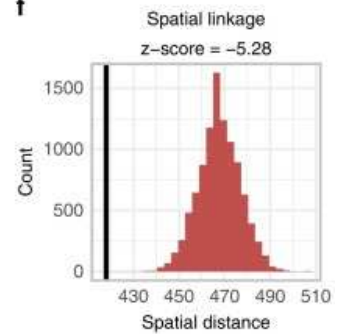


Figure 4.3.4 Relationship of core genome SNP groups to spatial and temporal distance.

a Clonality results for *A. baumannii*. Squares represent *A. baumannii* collected from surfaces. Colors represent clonal subgroup membership. Each colored set is a clonal subgroup with fewer than five SNPs different between all members of the group. Unfilled squares did not have fewer than five SNPs different with any other isolates. Lineage from BAP (identified in Fig. 3 by branch color) is indicated in the legend on the left. **b** Clonality results for *E. faecium*. Diamonds represent *E. faecium* collected from surfaces. Colors represent clonal subgroup membership. Each colored set is a clonal subgroup with fewer than five SNPs different between all members of the group. Unfilled diamonds did not have fewer than five SNPs different with any other isolates. Lineage from BAP is indicated in the legend on the left. For **c**, **d**, temporal distances are calculated as +1 for every 2-week span separating isolate collections. Spatial distances are given as +0 if isolates were collected from the same surface and room, +1 if they were collected from the same room, but different surfaces, and +2 if they were collected from different rooms. **c** Temporal linkage for *A. baumannii* clones. The expected temporal distance distribution is shown in blue and the observed temporal distribution is shown as a solid black line. **d** Spatial linkage for *A. baumannii* clones. The expected spatial distance distribution is shown in blue and the observed spatial distribution is shown as a solid black line. **e** Temporal linkage for *E. faecium* clones. The expected temporal distance distribution is shown in red and the observed temporal distribution is shown as a solid black line. **f** Spatial linkage for *E. faecium* clones. The expected spatial distance distribution is shown in red and the observed spatial distribution is shown as a solid black line.

As our taxonomic analysis demonstrated *A. baumannii*, *E. faecium*, *K. pneumoniae*, and *P. aeruginosa* were the most abundant putative pathogens collected at PAK-H, we next endeavored to determine population structure for isolates in these species. For each species, we annotated protein coding sequences with Prokka, constructed core-genome maximum-likelihood phylogenetic trees with Roary and RAxML, then identified lineages with fastGEAR/BAPS(31-34). Our results demonstrate that for *A. baumannii* and *E. faecium* but not *K. pneumoniae* or *P. aeruginosa*, a single lineage represented >70% of all isolates collected over 12 months. For all four species, time of collection, but not room or surface had the greatest concordance with phylogenetic position.

88.4% (69/78) of the *A. baumannii* isolates were from lineage 7 (Figure 4.3.3a), which was composed of several untypable isolates, and 7 sequence types (STs). Interestingly, the 4 USA-H genomes in ST208 clustered adjacent to one another and next to the 7 ST208 genomes from PAK-H. 72.3% (34/47) of the *E. faecium* isolates come from BAPS lineage 4. All lineage 2 and lineage 1 *E. faecium* isolates came from the 2nd and 4th week, respectively. *K. pneumoniae* contained 5 BAPS lineages with ST617, ST337, ST231, and ST147 relating to lineages 1, 2, 4, and 5, respectively. All the lineage 2 *K. pneumoniae* came from week 4 of our collections. *P. aeruginosa* had the greatest concordance between lineages and sequence types, as ST859, ST664, ST235, and ST571 corresponded to lineages 1, 2, 3, and 4, respectively. 74% (20/27) of the *P. aeruginosa* isolates came from week 8 of our collection, including all lineage 4 and lineage 1 isolates. Our analysis of population structure for recovered *A. baumannii*, *E. faecium*, *K. pneumoniae*, and *P. aeruginosa* indicates that specific lineages of closely related isolates dominated PAK-H surfaces. We next wanted to investigate if clonal groups of highly related isolates existed within lineages we identified for these pathogens.

4.3.4 Spatiotemporal distance identifies relevant epidemiologic groups

To identify epidemiologically meaningful groupings, we leveraged space and time information from our collections. For *A. baumannii* and for *E. faecium*, we iterated through every unique variant distance cutoff from the lowest distance between any two isolates until the lowest distance between any two isolates not in the same lineage (Figure 4.3.4a-e). We used these cutoffs to filter the isolate pairwise links edge list. For each cutoff, we found perfectly reciprocal groups with maximal graph coverage and recorded the number of cliques and the number of isolates per clique (Figure 4.3.4). Here we define cliques as

complete subgraphs within the network where each node in the clique is connected to each other node in the clique. Both *A. baumannii* and *E. faecium* showed a similar pattern where number of cliques rises sharply initially and then peaks. During this peak, there is a gradual increase in the number of isolates per clique, with cliques staying relatively balanced. After peaking, the number of cliques rapidly declines as formerly independent cliques merge. This merging interestingly results in one major clique with several other minor cliques. We then determined how much each clique grouping's spatial and temporal distances deviated from a null model generated with 10000 permutations for that clique grouping (Figure 4.3.4). If isolates spread randomly on surfaces, we would expect z-scores close to 0 for the spatial and temporal data. We projected the lowest z-score cutoffs onto the pairwise variant distances histogram (Figure 4.3.4). The greatest deviation from the null model for significant temporal (Figure 4.3.4) and spatial linkage (Figure 4.3.4) coincided with cutoffs that yielded the highest number of cliques. In this case, we found nine cliques for *A. baumannii* with both the time-minimizing distance and space-minimizing distance cutoff. For *E. faecium*, we found ten cliques for the time-minimizing cutoff and 8 cliques for the space-minimizing cutoff. The cutoff values in that range best fit the radiation of isolates on these surfaces. After cutoff values increase beyond the clique-maximizing value, within-clique spatial and temporal distance observations rapidly increase to match and even exceed null estimations, indicating that the epidemiologically relevant variant cutoff was likely passed.

For *A. baumannii* and for *E. faecium*, cliques are mostly restricted to single collection times, but some cliques, like clique 8 for *A. baumannii*, deviate from this trend and are instead broadly spread over surfaces in both time and space. Though most

cliques are restricted by time, cliques that are spread in time show room restricted patterning. This distribution of isolates could be explained by a reservoir of multiple clones with continual seeding to surfaces. In this scenario, most seeding events would not result in long-term surface, persistence, but a few clones could pass this strong filter to successfully survive for multiple weeks within rooms in a space dependent fashion.

4.3.5 PAK-H isolates have high genotypic and phenotypic resistance

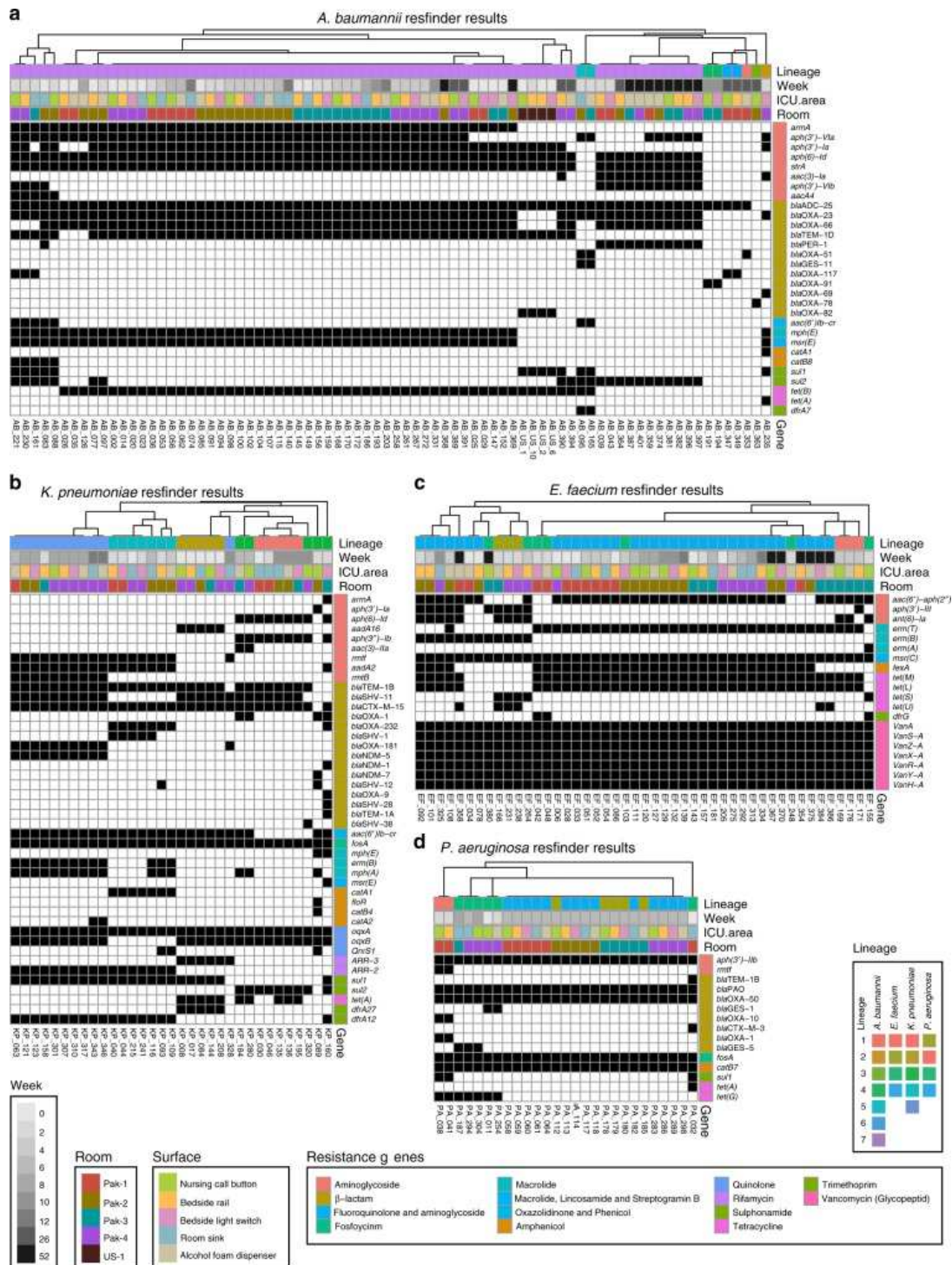


Figure 4.3.5 Genotypic antibiotic resistance in major species

Resfinder results for *A. baumannii* (a), *K. pneumoniae* (b), *E. faecium* (c), *P. aeruginosa* (d).

Resistance genes are grouped by antibiotic class on the y-axis and individual isolates are hierarchically clustered by their resistance genes on the x-axis. Black squares indicate the presence of a specific resistance gene in an isolate. Colored annotations are added next to the resistance genes for resistance gene class and above the charts for hierBAPS lineage (identified in Fig. 3 by branch color), week, surface, and room

We used ResFinder to identify ARGs in draft genomes of our sequenced *A. baumannii*, *E. faecium*, *K. pneumoniae*, and *P. aeruginosa* isolates(35). Additionally, we determined if these isolates were phenotypically resistant, intermediate, or susceptible using Kirby-Bauer disk diffusion assays in accordance with Clinical and Laboratory Standards Institute (CLSI) guidelines(36). For all species, we found hierarchical clustering of isolates based on ARG presence or phenotypic susceptibility indicated lineage was the major predictor of resistance-based clustering patterns. Specific lineages can dominate clinical infections and tight correlation of lineage with resistance may relate to this phenomenon(37). This linkage between lineage and antimicrobial resistance may also allow for rapid, sequence based rather than gene-based susceptibility predictions(38).

A. baumannii isolates harbored 30 unique ARGs against 9 different classes of antimicrobials (Figure 4.3.5a). 40% (12/30) of these ARGs were β -lactamases and 26.7% (8/30) were expected to confer phenotypic resistance against aminoglycosides (Figure 4.3.5a). 100% (65/65) of lineage 7 PAK-H isolates harbored *bla*_{OXA-23} and 95.4% (62/65) also had *bla*_{OXA-66}, while none (0/4) of the USA-H isolates had either of these carbapenemases. Interestingly, USA-H isolates clustered close together with most other lineage 7 PAK-H samples rather than as a separate group (Figure 4.3.5a). 92.3% (72/78) of the bacteria were resistant to 3 or more classes of antimicrobials including two carbapenems. 4.05% (3/74) of the PAK-H *A. baumannii* isolates were resistant to all 14 antimicrobials tested. Minocycline was most efficacious against PAK-H strains, with 92.3% (72/78) non-resistant.

E. faecium isolates had 20 unique resistance genes against 7 classes of antimicrobials (Figure 4.3.5a). Only *erm(A)* was unique to a single isolate. Components

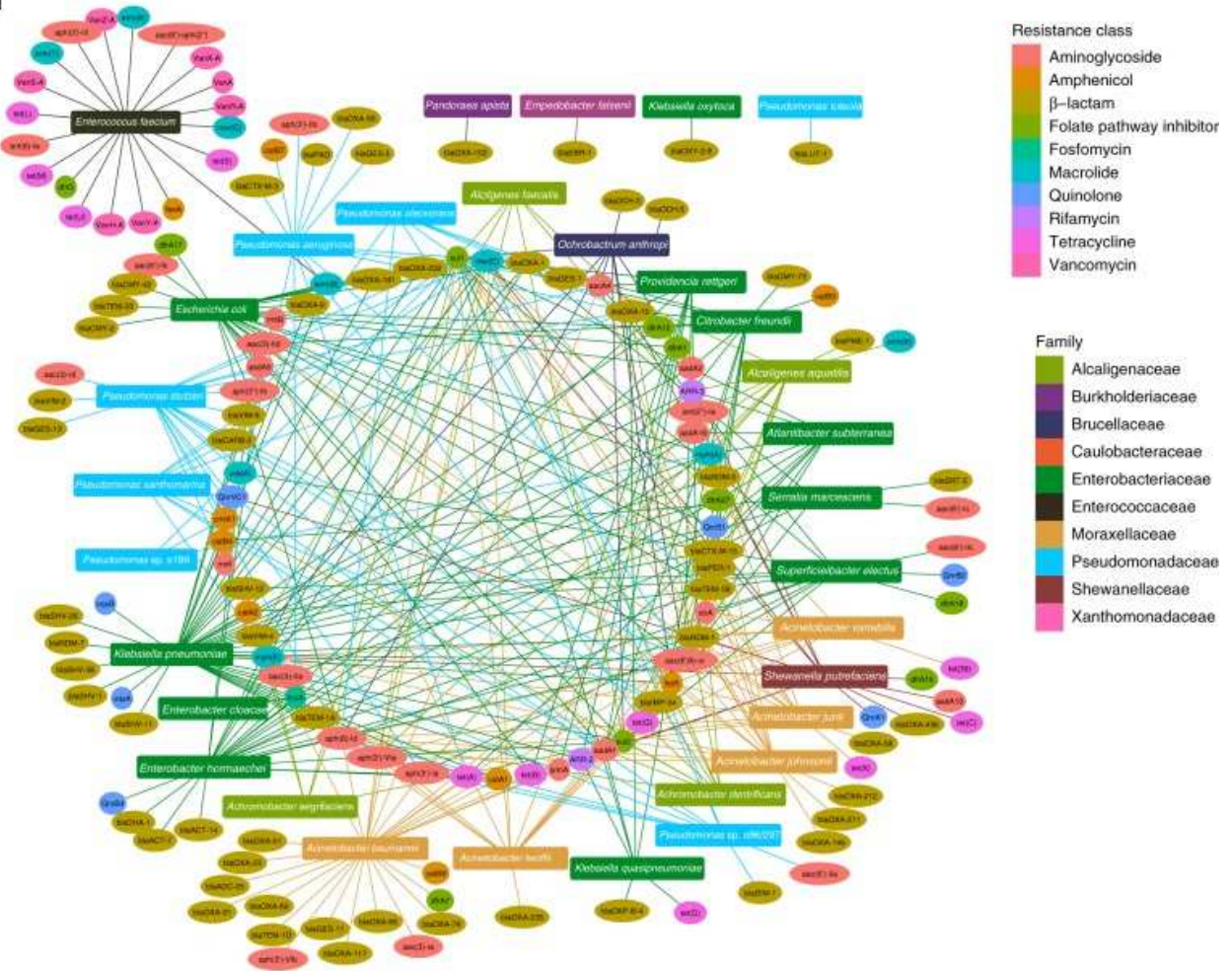
of the *vanA* operon and the macrolide ARG *msr(C)* were common to all isolates. As expected for *vanA* containing *E. faecium*, all isolates were resistant to vancomycin. 42.1% (24/57) were additionally resistant to chloramphenicol and doxycycline. All isolates were susceptible to daptomycin.

The *K. pneumoniae* isolates harbored 44 unique resistance genes and of these, 25.0% (11/44) were unique to single isolates (Figure 4.3.5c). 3 *bla*_{NDM} (*bla*_{NDM-1}, *bla*_{NDM-5}, and *bla*_{NDM-7}) and 2 *bla*_{OXA} (*bla*_{OXA-181} and *bla*_{OXA-232}) carbapenemase genes were identified. *bla*_{NDM-5} was found in *K. pneumoniae* on 10 surfaces and in all 4 PAK-H ICU rooms. 39.4% (13/33) of *K. pneumoniae* isolates were resistant to meropenem and imipenem. 100% of lineage 1 (5/5) and lineage 2 (5/5) isolates and 60% (3/5) of lineage 3 isolates were susceptible to these two antibiotics. All (33/33) isolates harbored the fosfomycin ARG *fosA* and an efflux pump component *oqxA*, however all lineage 4 isolates lacked the second component, *oqxB*.

P. aeruginosa isolates harbored 15 unique resistance genes against 6 classes of antimicrobials. All isolates had *aph(3')-Ib*, *bla*_{PAO}, *bla*_{OXA-50}, *fosA*, and *catB7* (Figure 4.3.5). 50% (3/6) of lineage 3 genomes had the carbapenemase *bla*_{GES-5}. All lineage 4 *P. aeruginosa* isolates and 3/5 lineage 1 isolates were pan-susceptible to antibiotics. In contrast, all (8/8) lineage 2 and 3 isolates were resistant to meropenem, ciprofloxacin, and gentamicin. Our results demonstrate that the major abundant HAI pathogens contain a high ARG burden and exhibit profound levels of multidrug resistance. Infections from these bacteria could have limited treatment options due to high phenotypic multidrug resistance.

4.3.6 ARGs against almost all antimicrobials are shared between species

a



b

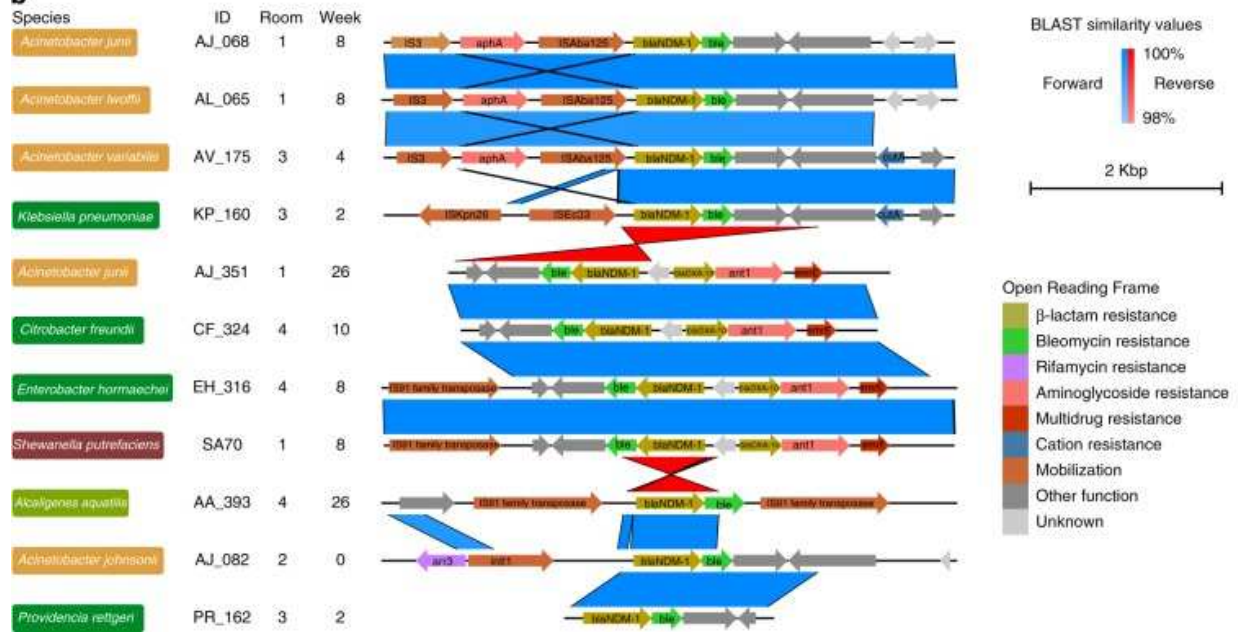


Figure 4.3.6 Shared antibiotic resistance genes across diverse taxonomic groups

a Species and resistance gene network diagram. Species are represented as rectangles colored by family. Resistance genes are represented by ovals colored by resistance gene class. Lines colored by species family are drawn from each species to all the resistance genes annotated by Resfinder in that species isolates. **b** Annotated *bla*_{NDM-1} contigs in 11 isolates. Protein annotations colored by putative function are shown as arrows for each isolate's *bla*_{NDM-1} contig. BLAST similarity values greater than 98% between contigs are shown in blue if they are oriented in the forward direction and red if they are oriented in the reverse direction. Species names are shown on the left in rectangular boxes colored by family and isolate ID, room, and week are also labeled

Given the extensive diversity and burden of high-risk ARGs found in *A. baumannii*, *E. faecium*, *K. pneumoniae*, and *P. aeruginosa*, we analyzed potential lateral transfer of ARGs between all collected species. To accomplish this, we concatenated identified acquired ARGs within each species and created a network diagram connecting each taxa with its ARGs (Figure 4.3.6a). The high connectivity of this network highlights the extensive promiscuity of ARGs we observe in these data. Strikingly, 57 ARGs were found in 2 or more species. These genes were expected to confer resistance against all classes of antibiotics, excluding vancomycin. *E. faecium* contained the macrolide resistance gene *erm(B)*, which was also shared with *E. coli*. Given that *E. faecium* is the sole gram-positive species in this collection, it unsurprisingly had the most species specific ARGs (n=17). *Sul1* was the most promiscuous ARG within our cohort, as it was identified in 22 different species, including those in *Acinetobacter*, *Achromobacter*, *Alcaligenes*, *Atlanibacter*, *Citrobacter*, *Escherichia*, *Enterobacter*, *Klebsiella*, *Ochrobactrum*, *Pseudomonas*, *Providencia*, *Shewanella*, and *Superficieibacter*. β -lactam ARGs were the most abundant class in our cohort, with a total of 57 identified from all 4 Ambler classes. Alarmingly, 40.3% (23/57) of these genes have putative carbapenemase activity. *bla*_{GES-5} is the only Ambler Class A carbapenemase. 34.7% (8/23) of genes we identified are Ambler Class B Metallo- β -lactamases, from the *bla*_{VIM}, *bla*_{IMP}, *bla*_{EBR}, *bla*_{DIM} and *bla*_{NDM} families. The

remaining 60.8% (14/23) were *bla*_{OXA} variants including the *bla*_{OXA-48-like} family members *bla*_{OXA-181} and *bla*_{OXA-232}. *bla*_{NDM-1} showed the greatest diversity of host species, as it was identified 11 times in 10 different species from *Alcaligenaceae*, *Enterobacteriaceae*, *Moraxellaceae*, and *Shewanellaceae*.

*bla*_{NDM} is a globally proliferated family of carbapenem resistance genes endemic to India and Pakistan (39). To better understand the local genetic context of *bla*_{NDM-1}, we performed long-read sequencing with the Oxford NanoPore MinION platform on all *bla*_{NDM-1} positive isolates (Figure 4.3.6b). *bla*_{NDM-1} in all genetic contexts was adjacent to *ble*, a bleomycin resistance gene. The *bla*_{NDM-1} loci region was nearly identical between *A. junii* AJ_068/*A. lwoffii* AL_065/*A. variabilis* AV_175 and *A. junii* AJ_351/*C. freundii* CF_324, *E. hormaechei* EH_316, and *S. putrefaciens* SA70. *A. junii* AJ_351, *C. freundii* CF_324, *E. hormaechei* EH_316, and *S. putrefaciens* SA70 additionally contained *bla*_{OXA-10} and *ant1*. *A. junii* AJ_068, *A. lwoffii* AL_065, and *A. variabilis* AV_175 had a different aminoglycoside resistance gene, *aph*. *A. johnsonii* AJ_082 contained the only rifamycin resistance gene, *arr3*. *A. junii* AJ_351, *C. freundii* CF_324, *E. hormaechei* EH_316, and *S. putrefaciens* SA70 also contained the *emrE* multidrug resistance transporter. On 72.7% (8/11) of the loci, *bla*_{NDM-1} was co-localized with a transposase associated gene. Our analysis of ARG content across species identified high interconnectivity between most gram-negative species and determined *bla*_{NDM-1} is situated in similar genetic contexts across diverse taxonomic groups, suggesting extensive horizontal ARG transfer.

4.3.7 *A. baumannii* and *E. faecium* have synergistic biofilm interactions

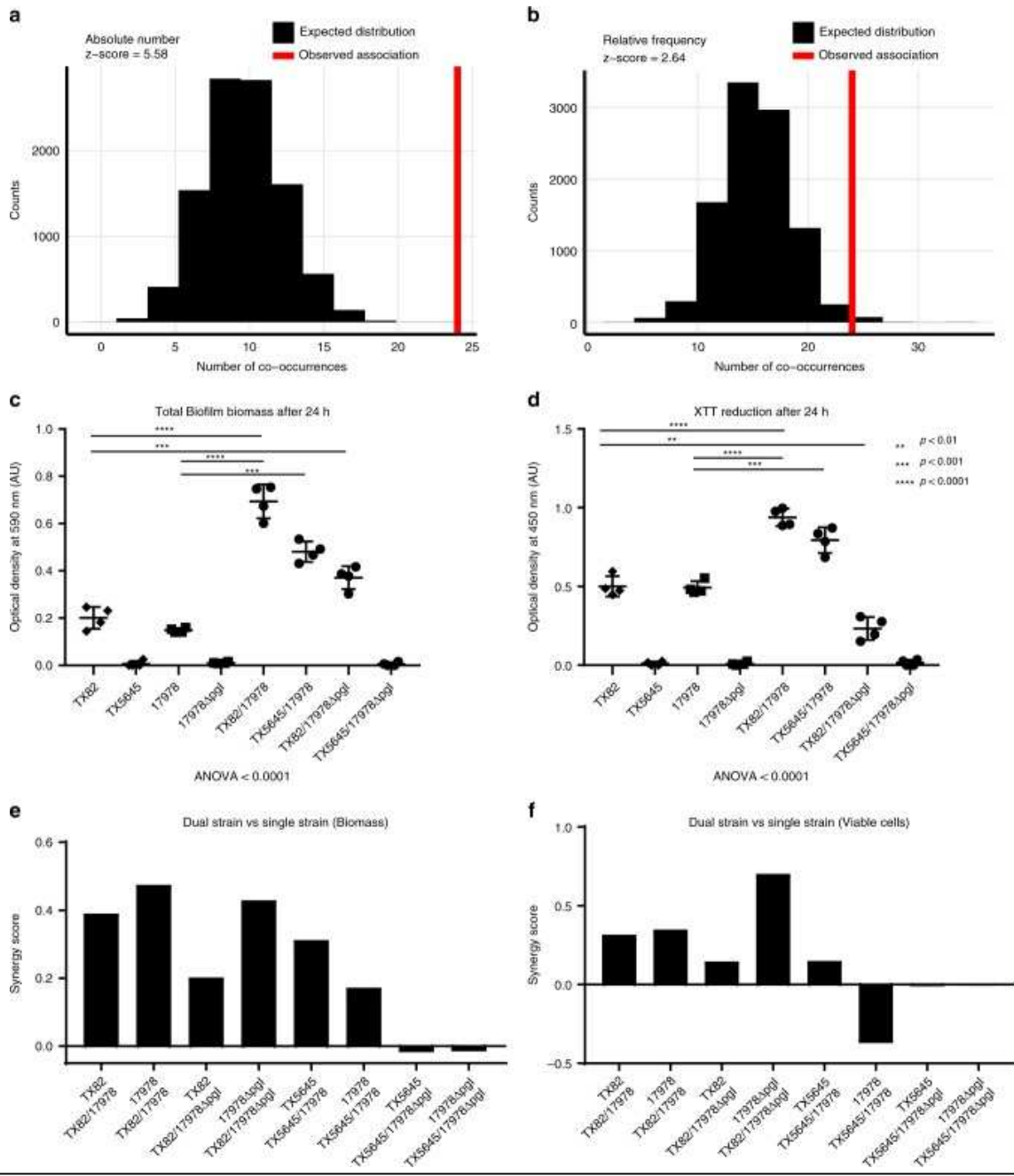


Figure 4.3.7 Synergistic biofilm interactions for *A. baumannii* and *E. faecium* predicted by surface collections

Permutation test of co-association between *A. baumannii* and *E. faecium* on surfaces conducted using species **a** absolute counts and **b** relative species frequencies. The expected distribution of the number of co-occurrences is shown in red and the observed number of co-occurrences in the dataset is shown as a vertical blue line. Total crystal violet stained **c** biofilm biomass and **d** XTT reduction for *A. baumannii* and *E. faecium* model biofilm strains grown in single and in co-culture (*P*-values were generated using unpaired, nonparametric Mann–Whitney statistical tests are indicated using the following mapping: **<math>p < 0.01</math>, ***<math>p < 0.001</math>, ****<math>p < 0.0001</math>). *y*-Axis for both plots is optical density at 590 nm and 450 nm, respectively, and error bars are 1 standard deviation. Synergy scores of dual vs single strain cultures for **e** biofilm biomass and **f** viable cells.

Bacteria harboring diverse ARGs may be recalcitrant to treatment regimens and could continually transmit from patients onto ICU surfaces, likely forming sessile biofilms to survive the dry conditions(40, 41). Indeed, biofilms composed of MDROs have been previously demonstrated to contaminate 93% (41/44) of hospital surfaces surveyed(23). To assay potential microbe-microbe interactions that may explain long-term surface persistence, we examined co-occurrences between abundant species in first three collection months using permutation testing. To remove potential bias from overrepresentation of certain taxa, we performed this analysis with both total counts and relative frequency. Both metrics demonstrated *A. baumannii* and *E. faecium* co-occurred on surfaces more often than predicted by chance ($P < 0.00001$ for *A. baumannii* and $P = 0.0083$ for *E. faecium* [permutation test]) (Figure 4.3.7ab).

We then obtained isogenic strains of *E. faecium* (TX82/TX5645) and *A. baumannii* (ATCC-17978, 17978 Δ pgl) capable of or deficient in biofilm formation, respectively(42, 43). Using every pairwise combination between the different species, we found co-culture of *E. faecium* TX82 with *A. baumannii* ATCC-17978 or *A. baumannii* 17978 Δ pgl, and *E. faecium* TX5645 with *A. baumannii* ATCC-17978 resulted in statistically significant increases ($P < 0.0001$ [Mann-Whitney U test]) in biofilm biomass relative to either of the parent strains (Figure 4.3.7b). This effect did not occur when both species were incapable of forming biofilms individually (Figure 4.3.7e).

As dead cells may be included in total analysis of biofilm biomass, we next specifically quantified the population of total viable cells between each pairwise interaction. Like results for total biofilm biomass, the number of viable cells increased significantly in *E. faecium* TX82/*A. baumannii* ATCC-17978 and *E. faecium* TX82/*A.*

baumannii 17978 Δ pgl compared to either parent strain ($P < 0.0001$ [Mann-Whitney U test]) (Figure 4.3.7d). However, in contrast to the increase in biofilm biomass observed for *E. faecium* TX5645/*A. baumannii* ATCC-17978 relative to both parent strains, we found a decrease in viable cells compared to *A. baumannii* ATCC-17978 (Figure 4.3.7f). Quantification of biofilm biomass synergy values between each strain combination shows all interactions except those between *E. faecium* TX5645 and *A. baumannii* 17978 Δ pgl are synergistic. For viable cells, interactions between *E. faecium* TX5645 and *A. baumannii* 17978 Δ pgl and *A. baumannii* ATCC-17978 versus *E. faecium* TX5645/*A. baumannii* 17978 Δ pgl are synergistic. These data suggest interspecies interactions between organisms identified on PAK-H ICU surfaces may enable increased survival due to synergistic growth inside biofilms. Importantly, relative efficacy of those interspecies biofilms depends strongly on individual strain capabilities.

4.4 Discussion

HAIs are a substantial patient health threat and economic burden(44). While pathogenic bacteria that often cause HAIs can be transferred via invasive medical procedures or directly between patients or healthcare providers, inanimate surfaces and shared equipment are also an important reservoir for bacterial transmission(14, 40). Here we report an in-depth, year-long investigation of bacterial colonization of hospital surfaces in two ICUs in Pakistan (PAK-H) and the USA (USA-H). We found substantially more contamination by MDROs on PAK-H surfaces compared to USA-H surfaces using identical differential and selective culture conditions.

In addition commonly recognized HAI causing bacteria, we found many potentially opportunistic pathogens and novel genomospecies from commonly pathogenic genera

(*Pseudomonas*, *Stenotrophomonas*, *Brevundimonas*). The first novel genomospecies from this collection to be fully characterized, *S. electus*, is a new genus of *Enterobacteriaceae* that harbored extended spectrum β -lactamases and was multidrug resistant(30). A previous taxonomic investigation determined that species from another novel genus of *Enterobacteriaceae*, *Pseudocitrobacter faecalis* and *Pseudocitrobacter anthropi*, harbored *bla*_{NDM-1} carbapenemases, and were identified in fecal samples from patients at hospitals in Pakistan(45). Currently no clinical evidence indicates these 3 species are human pathogens, but it is concerning that they exist proximal to known pathogens, encode clinically-relevant ARGs, and are phenotypically resistant to multiple drugs. Furthermore, increasing implementation of WGS in clinical laboratories is enabling identification of emerging pathogens which were previously misidentified by traditional methods, such as the first report of a bloodstream infection by *Kosakonia radicincitans*(46). Our results provide additional utility for the implementation of WGS for bacterial delineation from clinically-relevant environments. Further comparative analysis and molecular and phenotypic evidence for pathogenesis is required to demonstrate that this level of identification is clinically relevant or actionable.

A. baumannii, *E. faecium*, *K. pneumoniae*, and *P. aeruginosa*, the 4 most abundant bacteria in our cohort, are also common pathogens and common HAI agents. Interestingly, through core-genome phylogenetic analysis we found that our *A. baumannii* and *E. faecium* isolates are dominated by single lineages, but *K. pneumoniae* and *P. aeruginosa* have nearly equal numbers of isolates from multiple lineages. Previous reports of *K. pneumoniae* and *E. cloacae* isolates from a US hospital system and Italy showed they were similarly composed of diverse sequence types(47, 48). Timepoint of

sample collection was the variable that showed greatest concordance with phylogenetic lineage. Lineage 7, the main group of *A. baumannii* isolates, was composed of several sequence types, including ST218, ST208, and ST195. These STs correspond to major strains collected of *bla*_{OXA-23} bearing *A. baumannii* in Indonesia; additionally, *bla*_{OXA-23} positive ST195 isolates were responsible for an outbreak of infections in North China(49, 50). The 4 ST208 USA-H *A. baumannii* isolates were genomically similar to the PAK-H isolates, although the PAK-H isolates harbored *bla*_{OXA-23} whereas the USA-H isolates have *bla*_{OXA-81}. This parallels a previous investigation which found near identical genomes and plasmids from carbapenem resistant *Enterobacteriaceae* in the US and Pakistan, but US isolates exclusively contained *bla*_{KPC} while *bla*_{NDM} was only found in isolates from Pakistan(29). The most abundant *E. faecium* sequence type, ST132, was primarily contained in lineage 4. Isolates from this ST have been reported as both etiological urinary tract infection agents and as commensal animal bacteria(51, 52).

Though *A. baumannii* and *E. faecium* were dominated by a single lineage, we had evidence that the 6 clones of *E. faecium* came from all 4 lineages, but that 3/6 of the clones were from the dominant lineage 4 group. In contrast, all identified *A. baumannii* clones were in the dominant lineage 7. Given that clone 5 of *E. faecium* was found on 8/9 timepoints during our collections, including the 0th and 52nd week, it is possible that PAK-H surfaces are being colonized by a common seeding source or that these isolates represent the predominant clone circulating in the PAK-H region. Source investigation of carbapenemase producing organisms in a US hospital system determined that plasmids mobilizing the ARGs originated from building plumbing(53). As further evidence of this, we found that *A. baumannii* and *E. faecium* clones are more likely to co-localize in space

and time than if they were randomly distributed. This may have important clinical ramifications, as one analysis determined that although only 8.7% of ICU bacteria sequenced are from a clonal lineage, they were associated with clinical infection in 62% of occurrences(37). Therefore, eradication of the common contaminating source could drastically reduce spread of these clones and thereby reduce potential of spread to hospital patients. If bacteria are transmitting between surfaces, spatial and temporal linkage of these surfaces could mean effective decontamination of surfaces will have a combinatorial effect.

In our variant analysis, we identified that most cliques (complete subgraphs within the network where each node in the clique is connected to each other node in the clique) were time restricted, but a few cliques persisted across multiple collections. These persistent cliques subsequently showed room restriction. Several contamination routes could explain these results. For example, seeding bacteria may originate from patients occupying hospital rooms(54). Bacteria coming from different patients are likely genetically distinct in variant analysis; even within a patient, multiple lineages of the same species could co-exist(55). Seeding events from patient to surface would represent a bottleneck event and persistence on surfaces would represent another bottleneck. Bacteria passing the first bottleneck would be detected within a single collection time, and bacteria passing the second would be found in multiple collection times. Bacterial clones on many surfaces would have higher chances to spread to other surfaces in the same room or different rooms. Similar contamination patterns could also be observed due to water contamination in the hospital(53). PAK-H uses tap water with Virkon S disinfectant tablets (Lanxess) to clean hospital surfaces. If tap water has high bacteria burden or if not

enough tablets are used, the disinfectant protocol could contaminate rather than decontaminate surfaces. This tap water environmental source could contain a polymicrobial community, thus acting as reservoir for multiple bacterial lineages(56). With tap water, the first significant bottleneck would be getting from the water system to surfaces, but subsequent steps would be in line with the patient contamination scenario. In support of these potential contamination routes, the bacteria we observe in this study are a mixture of human fecal bacteria and water environmental bacteria(57, 58). This analysis demonstrates how a surface focused sampling and analysis approach can generate epidemiologically meaningful insights for future investigation. In our case, the hospital water system and ICU room patients can both be tested as potential reservoirs for observed ICU surface bacterial contaminants, and a longitudinal sampling scheme similar to the one used in our study would enable estimation of transmission dynamics between these putative contamination sources and sinks.

The *A. baumannii*, *E. faecium*, *K. pneumoniae*, and *P. aeruginosa* isolates we recovered from PAK-H surfaces had high ARG burdens and were often phenotypically resistant to multiple classes of antibiotics commonly used as treatment against them. This is particularly troublesome for local patient safety at PAK-H given that a retrospective cohort analysis found significant increases in 30-day mortality after infection when comparing patients infected by multidrug resistant versus susceptible organisms(59). Particularly problematic are the 3 *A. baumannii* isolates we recovered that were resistant to all antibiotics tested with CLSI interpretive criteria, similar to 20 pandrug-resistant isolates recovered from countries bordering the Mediterranean Sea(60). While we were unable to determine directionality of transfer, linkage analysis between acquired ARGs

and species harboring them show numerous instances of identical ARGs in different species. This is best exemplified by *bla*_{NDM-1} presence in 10 different species. Using long-read nanopore sequencing, we found *bla*_{NDM-1} situated in a variety of genetic contexts, even between the two *A. junii* isolates that contained it. Similar to previous reports of *bla*_{NDM-1} in isolates recovered from Pakistan patient stool samples, the mobilization element ISAb₁₂₅ was co-localized with *bla*_{NDM-1} in 4/11 of our isolates(61). Additionally, 4/11 isolates also contained *bla*_{NDM-1} close to *bla*_{OXA-10}, similar to numerous *bla*_{NDM-1} harboring Enterobacteriaceae isolates from hospitalized patients(61).

Bacteria surviving in the built environment likely exist in sessile biofilms, which can make them difficult to eradicate(56). Numerous reports have determined dual or multi-species biofilms have distinct characteristics to enhance survival and pathogenicity(62-64). Direct sampling of ICU samples showed polymicrobial biofilms are widespread(23). Biofilm formation is an important component for pathogenesis of *Enterococcus* and *Acinetobacter*(65, 66). In both organisms, biofilm formation often requires extracellular attaching proteins including LH92_11085 and OmpA in *A. baumannii* or the Emp pilus in *E. faecium*(67-69). Variation has been observed among the ability of *A. baumannii* clinical isolates to form biofilms, but several strains are capable of growing on urinary catheter surfaces(70). In *E. faecium*, adaption to a biofilm is associated with changes in the transcriptional program(71). 16S rRNA gene sequencing of high-touch surfaces at large public hospitals in Brazil identified both *E. faecium* and *A. baumannii* co-localized to the same surface(72). Despite this observation and the role of individual genes in biofilm formation for both species, there is a dearth of relative knowledge on specific interactions between these two species that may occur in the built environment. Our analysis of co-

occurrence between organisms indicates *A. baumannii* and *E. faecium* isolates were cultured together more frequently than expected by chance. Additionally, we found co-culture of model *E. faecium* and *A. baumannii* biofilm-forming and biofilm-deficient strains resulted in changes in total biofilm biomass and total viable cells dependent on the biofilm formation capacity of input strains. These results are consistent with a previous report on changes between *Enterococcus faecalis* and *P. aeruginosa* biofilms, where synergistic interactions between the exopolysaccharide produced by *P. aeruginosa* is responsible for spatial segregation of the two species in biofilms(73). It is therefore possible that conserved interspecies interactions between *Enterococcus* spp. and gram-negative nonfermenting bacteria may explain prolonged surface survival.

One limitation of our study is some bacterial species may be more robust than others in surviving on surfaces and in the sampling protocol. For example, bacteria could exist transiently between sampling times in concert on surfaces. However, the number of rare species we collected helps to allay this concern. We also did not concurrently characterize isolates recovered from clinical specimens. Therefore, we are unable to determine if lineages found on surfaces correlate with lineages associated with clinical infection in the hospital and in addition, we cannot corroborate linkage of lineages (e.g. *P. aeruginosa* in week 4) or clones (e.g. *E. faecium* clone 5) with time to determine if outbreaks occurred. Detailed analysis of clinical isolates may additionally inform associations of identified *A. baumannii*, *E. faecium*, *K. pneumoniae*, or *P. aeruginosa* lineages with specific infection niches and elucidate novel virulence factors or identify contaminated medical equipment. Additionally, samples of patient/healthcare workers

skin, stool, or oral microbiota, or of the room plumbing system could be used to further track transmission of the recovered MDROs to a specific source.

Our work represents a thorough longitudinal analysis of hospital surface contamination in Pakistan. We unequivocally demonstrate that MDRO burden is higher on PAK-H surfaces than on analogous USA-H surfaces. Using WGS we found that while the recognized human pathogens *A. baumannii*, *E. faecium*, *K. pneumoniae*, and *P. aeruginosa* are the most abundant organisms, a variety of potentially pathogenic taxa and novel genomospecies were also recovered. Analysis of lineages in the 4 most abundant species and clones in *A. baumannii* and *E. faecium* provide evidence of a common point source of contamination. Particularly alarming is our determination that these isolates harbor a high burden of ARGs, are often phenotypically multidrug resistant, and that identical ARGs are housed on a variety of genetic platforms in multiple species. Synergistic growth of *E. faecium* and *A. baumannii* in dual species biofilms may explain statistically significant co-occurrence on PAK-H surfaces. Complex ecology revealed by our hospital sampling highlights that common human pathogens and rare species frequently colocalize and share clinically relevant genes. Rapid dissemination of bacterial pathogens and plasmid borne ARGs stress importance of surveilling bacterial isolates in high-risk areas to protect vulnerable hospitalized patients around the globe.

4.5 Materials and Methods

4.5.1 Sample collection and culturing

Intensive care unit rooms were sampled every other week for three months and then at six months and one year after the initial sampling. At each time point, five surfaces were

sampled in each patient room (if available in that room): the nursing call button (sampled the call button that is attached to the right of the bedside rail, swabbing as much of the surface as possible), the bedside rail (swabbing approximately 6 inches of the rail, swabbing the side that is closest to the room door), the main room light switch (swabbing the entire switch and switch plate), the sink handles (swabbing the handles on the sink inside the patient room, swabbing both handles, front and back), the alcohol hand foam dispenser (swabbing the one closest to the patient room, swabbing the high touch area of the dispenser). If a bedpan, commode or toilet was present in the patient room, this was also sampled, including the seat and handle. The Eswab collection and transport system (Copan, Murietta, CA) was used to collect all specimens; swabs were moistened prior to sample collection. Two swabs were held together for specimen collection. Specimens collected in Pakistan were shipped to the US site for workup and analysis.

One Eswab specimen was vortexed and 90 μ L of eluate was inoculated to each of the following culture medium: Sheep's blood agar (Hardy Diagnostics), MacConkey agar (Hardy Diagnostics), VRE chromID (bioMerieux), Spectra MRSA (Remel), HardyCHROM ESBL (Hardy), Pseudo agar (Hardy), and MacConkey agar with cefotaxime (Hardy). Plates were incubated at 35 °C in an air incubator and incubated up to 48 hours prior to discard if no growth. Up to 4 colonies of each colony morphotype (as appropriate for the agar type) were subcultured and identified using MALDI-TOF MS with the VITEK MS system(74-78). A second Eswab specimen was used for *Clostridium difficile* culture with a heat-shock broth enrichment method as previously described(79). All isolates recovered were stored at -80 °C in TSB with glycerol.

4.5.2 Antibiotic susceptibility testing

Antimicrobial susceptibility testing was performed using Kirby Bauer disk diffusion, interpreted according to CLSI standards(36).

4.5.3 Illumina Whole Genome Sequencing

Unique colony morphotypes from the initial swab plates were streaked for isolation on blood agar. After a culture was deemed pure by visual determination, ~10 colonies were suspended in deionized water with a sterile cotton swab. Total genomic DNA was extracted from the suspension using the bacteremia kit (Qiagen, Germantown, MD, USA). DNA was quantified with the Quant-iT PicoGreen dsDNA assay (Thermo Fisher Scientific, Waltham, MA, USA) .5 ng/ul of DNA was used as input for Illumina sequencing libraries with the nextera kit (Illumina, San Diego, CA, USA)(80). The libraries were pooled and sequenced on a NextSeq HighOutput platform (Illumina) to obtain 2x150 bp reads. The reads were demultiplexed by barcode, had adapters removed with Trimmomatic v.36, and contaminating sequences with Deconseq v.4.3(81, 82). Processed reads were assembled into draft genomes using the de-novo assembler SPAdes v3.11.0(83). The scaffolds.fasta files were used for all downstream analysis. Assembly statistics on the assemblies was quantified using QUAST v4.5(84). Prokka v1.12 was ran on the scaffolds file to identify open reading frames > 500 bp in length(31).

For the 11 isolates chosen to be sequenced with Nanopore technology, Genomic DNA was extracted using the Genomic-Tip 500/G (Qiagen) and genomic DNA buffer set (Qiagen) per manufactures instructions. The DNA was converted into a sequencing library on with the Rapid Barcoding Kit (Nanopore, Cambridge, MA, USA) per manufactures instructions and sequenced on the MinION platform. The output fastq files were used in a hybrid assembly with SPAdes v3.11.0 and processed Illumina reads.

These assemblies are uploaded to NCBI under BioProject: [PRJNA497126](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA497126).

4.5.4 Taxonomic assignment

All isolates were initially identified using the VITEK MS MALDI-TOF MS v2.3.3. Following draft genome assembly, the species determination for all isolates were then investigated using an *in silico* approach. MASH was performed against all of the isolate genomes(85). Isolates that had 100% concordance between the MALDI-TOF MS assignment and the top 10 MASH hits were determined to be the species assigned by MALDI-TOF MS. Isolates that had discrepant analysis were then manually investigated further, by using RNAmmer v1.2 to identify the 16S rRNA sequence, submission of that sequence to the EZ BioCloud taxonomic database, and finally ANI analysis with the mummer method between the isolate in question and the appropriate type genome (if available) using the JSpecies webserver (<http://jspecies.ribohost.com/jspeciesws>)(86-88). Species were determined if the genome in question had > 95% ANIm with the type genome (if available), or > 99% 16S rRNA identity (if type genome is not available)(89, 90). Isolates that did not pass either of these thresholds are therefore considered to be novel genomospecies. Finally, all the isolates sequenced in this study were used to construct a Hadamard matrix, representing the product of the average nucleotide identity and percent genome aligned, with the ANIm method from pyANI (<https://github.com/widdowquinn/pyani>). The matrix was visualized using the python package Seaborn (<http://seaborn.pydata.org>) and annotated for initial MALDI-TOF MS identification, and *in silico* assignment if discrepancies were identified.

4.5.5 Core genome alignment

The gff files produced from Prokka for *A. baumannii*, *E. faecium*, *K. pneumoniae*, and *P. aeruginosa* were used to construct a core-genome alignment with Roary v3.8.0 and PRANK v1.0(32, 91). fastGEAR was ran on the respective core_genome_alignment.aln output of Roary to identify instances of recombination within these species(34). The recombinant regions were removed using custom python scripts. The recombination purged core genome alignment was used to generate a maximum likelihood tree with RAxML v8.2.11(33). The output newick file was visualized in iTOL. *In silico* multilocus sequence typing (MLST) was performed with the MLST program. The sequence type information, week of collection, room of collection, and surface was viewed as a color strip in iTOL(92). Lineages identified by hierBAPS during fastGEAR were also marked on the trees(93).

4.5.6 Clonality analysis

Pairwise SNP counts between all isolates in the recombinant corrected core genome alignment were calculated. All paired distances greater than 5 SNPs were excluded from further clonality analysis. Pairwise groupings with 5 or fewer SNPs were imported to Gephi as an unweighted pairwise links table. Gephi's built in modularity analysis was used to isolate perfectly reciprocal groupings. R was used to visualize these groupings.

Pairwise SNP distances were calculated as the number of SNPs between two isolates divided by the total number of positions in the core genome alignment.

4.5.7 Calculate temporal and spatial distances for variant cliques

Spatial and temporal analysis for variant distance for variant cliques used the same distances as core genome SNP linkage analysis. Spatiotemporal linkage analysis was

conducted for isolates in the first 3 months of collection. For each cutoff value, observed distances were calculated by adding together the spatial or temporal distances within clique and expected distributions were calculated by conducting 10,000 permutations of the spatial and temporal distances using the *sample* function in R v3.53. Thus, permutations kept clique structure, but shuffled distance information.

4.5.8 ARG identification

Acquired ARGs against aminoglycosides, amphenicols, β -lactam, folate pathway inhibitors, fosfomycin, macrolides/lincosamides/streptogramins, quinolones, rifamycin, tetracycline, vancomycin were annotated using the ResFinder BLAST identification program(35). For the abundant species, the presence/absence matrix of ARGs was visualized in pheatmap (R). Associated metadata was displayed as a color strip to represent bacterial isolate demographics and expected resistance to antibiotics. To identify connectivity between the recovered species from the Pakistan ICU, we constructed a Source/Target/Edge formatted file, where each source represented a novel or curated genomospecies, a target was the unique ARG, and Edge weight was determined to be the number of times that ARG was identified within that species. The file was visualized in Cytoscape v3.4.0(95).

4.5.9 *bla*_{NDM-1} loci annotation and comparison

A ~6-2kB series of nucleotides flanking the *bla*_{NDM-1} loci in all positive strains was manually retrieved from SPAdes output of MinION & Illumina hybrid assembly. The nucleotides were re-annotated with prokka. The .gff file was used as input for Roary, to identify identical genes within the loci pan-genome. The .gbk files from prokka were

viewed for open reading frames and BLAST similarity in EasyFig(96). The sequences were ordered by their relationship from the newick tree created from the presence/absence matrix of genes. All loci in the pan-genome were submitted to BLASTX against the refseq_proteins in October 2017 to identify a putative function(97). The pairwise BLAST similarity was visualized on the EasyFig v2.2.2 construction by BLASTN similarity between the fasta files.

4.5.10 *A. baumannii* and *E. faecium* co-association permutation testing

Testing for significant association of *A. baumannii* and *E. faecium* was conducted using MALDI-TOF MS identifications from the first 3 months of PAK-H collections. The number and type of unique bacteria on each surface was tabulated. The number of surfaces with both *A. baumannii* and *E. faecium* was recorded as the observed frequency of co-occurrence. Absolute number and relative frequency expected distributions for *A. baumannii* and *E. faecium* co-occurrence were calculated using permutation tests with 10,000 random subsamples. For absolute number, the exact number of each bacterial species we collected was randomly distributed to a blank surface space with the restriction that each surface could not have more than one of the same species and that each surface had to get the same number of bacteria that was originally collected from the surface. This resulted a new permuted collection space with the same overall number of each bacterial species, but with randomized placement for each bacterium. The co-occurrence of *A. baumannii* and *E. faecium* for this permuted collection space was then recorded for the expected distribution. For relative frequency, the number of each species collected was used to calculate the frequency of that bacterial species in the collections. During permutation species were randomly chosen, weighted by their frequency in the

collections. R was used to visualize the *A. baumannii* and *E. faecium* co-association expected distributions and observed values.

4.5.11 Biofilm assays

Frozen cultures of *A. baumannii* ATCC-17978 (17978), *A. baumannii* ATCC-17978 Δ pgl, *E. faecium* TX82, and *E. faecium* TX5645 were streaked onto tryptic soy agar (Difco, Detroit, MI, USA) and grown overnight at 37 C. Isolated colonies were suspended in tryptic soy broth (Difco, Detroit, MI, USA) supplemented with .5% glucose (MP Biomedicals, Santa Ana, CA, USA) to promote the growth of *E. faecium* biofilm and quantified for OD600 using a 1:10 dilution. In concordance with previous investigations using respective strains, the *A. baumannii* isolates were normalized to .05 OD600 and the *E. faecium* were normalized to .10 OD600. For functional assays.

To grow biofilms, 200 μ l of each single strain or 100 μ l of *A. baumannii* and 100 μ l of *E. faecium* dual species biofilms were added to tissue culture treated 96 well polystyrene microtiter plates (Sigma Aldrich, St. Louis, MO, USA) in triplicate. We additionally plated cell-free controls to ensure that no contamination occurred and to subtract out background absorbance reading. After pipetting, the plates were gently pipetted up and down to ensure that the strains mixed thoroughly. The plates were covered with breath ez membrane (Diversified Biotech, Dedham, MA, USA) and grown on the benchtop at approximately 22 Celsius for 16 hours.

Following a growth period, the biofilm plates had planktonic cells removed by washing thoroughly with 250 μ l sterile phosphate buffered saline (PBS) (Thermo Fisher Scientific, Waltham, MA, USA) three times. To obtain the total biofilm biomass, the washed biofilms

were fixed with 250 µl bouin's solution (Sigma Aldrich) at 22 ° Celsius on the benchtop for 30 minutes. The fixative was washed three times with 200 µl sterile PBS three times and then stained with 250 µl .01% crystal violet (Sigma Aldrich) in water for 30 minutes at 22 °Celsius on the bench. Finally, the unstained crystal violet was removed by washing three times with PBS and then the biomass was solubilized with 250 µl of 100% ethanol (Sigma Aldrich). The amount of biofilm biomass was quantified using nm absorbance with a Synergy H1(BioTek) spectrophotometry machine. All raw absorbance values were adjusted by removing the background values obtained from the cell-free TSB controls. The conditions had average and standard deviation calculated.

For quantification of total viable cells in the biofilm, the biofilms were formed as previously described. After 16 hours growth at 22 ° Celsius, planktonic cells were removed by washing thoroughly with 250 µl PBS. The XTT cell viability kit (Cell Signaling Technologies, Danvers, MA, USA) was then performed according to manufacturer's instructions. The plates were read in the Synergy H1 spectrophotometry machine after 5-hour incubation in the dark.

For the crystal violet and XTT reduction assays, the biofilm synergy scores were calculated as previously reported for dual species biofilms. For each pairwise comparison, the synergy scores were reported as the difference between the average plus standard deviation for the single species biofilm and average minus standard deviation of the dual species biofilm.

$$(1) \text{ Biofilm synergy} = (\text{Average}_{\text{DualSpecies}} - \text{Standard Deviation}_{\text{DualSpecies}}) - (\text{Average}_{\text{SingleSpecies}} + \text{Standard Deviation}_{\text{SingleSpecies}})$$

4.5.12 Statistics

Unpaired, nonparametric Mann Whitney statistical tests were used to compare the adjusted OD₅₉₀ and OD₄₅₀ values between the total biofilm biomass and total viable cells in the dual vs single species biofilms.

4.5.13 Data availability

Assemblies are available from NCBI under BioProject: [PRJNA497126](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA497126).

4.6 Acknowledgments

This work was supported by a United States Agency for International Development award (award number 3220-29047) to SA, CAB, and GD. JHK received support from the Washington University Institute of Clinical and Translational Sciences grant UL1TR000448, sub-award KL2TR000450 from the National Center for Advancing Translational Sciences of the National Institutes of Health. RFP received support from the Monsanto Excellence Fund Graduate Fellowship. AWD received support from the Institutional Program Unifying Population and Laboratory-Based Sciences Burroughs Welcome Fund grant to Washington University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies. We thank Mario Feldman and his laboratory for providing strains of *A. baumannii* ATCC-17978 (17978) and *A. baumannii* ATCC-17978 Δ pgl. Additionally, we thank Barbara Murray and her laboratory for providing strains *E. faecium* TX82 and TX5645. The authors thank Center for Genome Sciences & Systems Biology staff, Eric Martin, Brian Koebbe, and Jessica Hoisington-López for technical support and sequencing expertise.

4.7 References

1. Crofts TS, Gasparri AJ, Dantas G. Next-generation approaches to understand and combat the antibiotic resistome. *Nat Rev Microbiol.* 2017;15(7):422-34. doi: 10.1038/nrmicro.2017.28. PubMed PMID: 28392565.
2. Stone PW, Gupta A, Loughrey M, Della-Latta P, Cimiotti J, Larson E, Rubenstein D, Saiman L. Attributable costs and length of stay of an extended-spectrum beta-lactamase-producing *Klebsiella pneumoniae* outbreak in a neonatal intensive care unit. *Infection control and hospital epidemiology.* 2003;24(8):601-6. Epub 2003/08/28. doi: 10.1086/502253. PubMed PMID: 12940582.
3. Cosgrove SE, Kaye KS, Eliopoulos GM, Carmeli Y. Health and economic outcomes of the emergence of third-generation cephalosporin resistance in *Enterobacter* species. *Arch Intern Med.* 2002;162(2):185-90. Epub 2002/02/13. PubMed PMID: 11802752.
4. Carmeli Y, Eliopoulos G, Mozaffari E, Samore M. Health and economic outcomes of vancomycin-resistant enterococci. *Arch Intern Med.* 2002;162(19):2223-8. Epub 2002/10/24. PubMed PMID: 12390066.
5. Allcock S, Young EH, Holmes M, Gurdasani D, Dougan G, Sandhu MS, Solomon L, Torok ME. Antimicrobial resistance in human populations: challenges and opportunities. *Glob Health Epidemiol Genom.* 2017;2:e4. Epub 2017/12/26. doi: 10.1017/ghg.2017.4. PubMed PMID: 29276617; PMCID: PMC5732576.
6. Dortet L, Poirel L, Nordmann P. Worldwide dissemination of the NDM-type carbapenemases in Gram-negative bacteria. *BioMed research international.*

2014;2014:249856. Epub 2014/05/03. doi: 10.1155/2014/249856. PubMed PMID: 24790993; PMCID: PMC3984790.

7. Wang R, van Dorp L, Shaw LP, Bradley P, Wang Q, Wang X, Jin L, Zhang Q, Liu Y, Rieux A, Dorai-Schneiders T, Weinert LA, Iqbal Z, Didelot X, Wang H, Balloux F. The global distribution and spread of the mobilized colistin resistance gene *mcr-1*. *Nat Commun*. 2018;9(1):1179. Epub 2018/03/23. doi: 10.1038/s41467-018-03205-z. PubMed PMID: 29563494; PMCID: PMC5862964.

8. Carrer A, Poirel L, Yilmaz M, Akan OA, Feriha C, Cuzon G, Matar G, Honderlick P, Nordmann P. Spread of OXA-48-encoding plasmid in Turkey and beyond. *Antimicrobial agents and chemotherapy*. 2010;54(3):1369-73. Epub 2010/01/21. doi: 10.1128/AAC.01312-09. PubMed PMID: 20086157; PMCID: PMC2825965.

9. Potter RF, D'Souza AW, Dantas G. The rapid spread of carbapenem-resistant Enterobacteriaceae. *Drug Resist Updat*. 2016;29:30-46. doi: 10.1016/j.drug.2016.09.002. PubMed PMID: 27912842; PMCID: PMC5140036.

10. de Man TJB, Lutgring JD, Lonsway DR, Anderson KF, Kiehlbauch JA, Chen L, Walters MS, Sjolund-Karlsson M, Rasheed JK, Kallen A, Halpin AL. Genomic Analysis of a Pan-Resistant Isolate of *Klebsiella pneumoniae*, United States 2016. *MBio*. 2018;9(2). Epub 2018/04/05. doi: 10.1128/mBio.00440-18. PubMed PMID: 29615503; PMCID: PMC5885025.

11. Sonnevend A, Ghazawi A, Hashmey R, Haidermota A, Girgis S, Alfaresi M, Omar M, Paterson DL, Zowawi HM, Pal T. Multihospital Occurrence of Pan-Resistant *Klebsiella pneumoniae* Sequence Type 147 with an ISEcp1-Directed *blaOXA-181*

Insertion in the mgrB Gene in the United Arab Emirates. *Antimicrob Agents Chemother.* 2017;61(7). Epub 2017/04/26. doi: 10.1128/AAC.00418-17. PubMed PMID: 28438945; PMCID: PMC5487649.

12. CDC. Antibiotic Resistance Threats in the United States, 2013. <http://www.cdc.gov/drugresistance/pdf/ar-threats-2013-508.pdf>: Centers for Disease Control and Prevention, 2013.

13. O'Neill J. Tackling Drug-Resistant Infections Globally: Final Report and Recommendations. *Review on Antimicrobial Resistance*, 2016.

14. Mora M, Mahnert A, Koskinen K, Pausan MR, Oberauner-Wappis L, Krause R, Perras AK, Gorkiewicz G, Berg G, Moissl-Eichinger C. Microorganisms in Confined Habitats: Microbial Monitoring and Control of Intensive Care Units, Operating Rooms, Cleanrooms and the International Space Station. *Front Microbiol.* 2016;7:1573. doi: 10.3389/fmicb.2016.01573. PubMed PMID: 27790191; PMCID: PMC5061736.

15. Weiner LM, Webb AK, Limbago B, Dudeck MA, Patel J, Kallen AJ, Edwards JR, Sievert DM. Antimicrobial-Resistant Pathogens Associated With Healthcare-Associated Infections: Summary of Data Reported to the National Healthcare Safety Network at the Centers for Disease Control and Prevention, 2011-2014. *Infection control and hospital epidemiology.* 2016;37(11):1288-301. Epub 2016/10/22. doi: 10.1017/ice.2016.174. PubMed PMID: 27573805.

16. Hidron AI, Edwards JR, Patel J, Horan TC, Sievert DM, Pollock DA, Fridkin SK, National Healthcare Safety Network T, Participating National Healthcare Safety Network F. NHSN annual update: antimicrobial-resistant pathogens associated with healthcare-

associated infections: annual summary of data reported to the National Healthcare Safety Network at the Centers for Disease Control and Prevention, 2006-2007. *Infection control and hospital epidemiology*. 2008;29(11):996-1011. Epub 2008/10/25. doi: 10.1086/591861. PubMed PMID: 18947320.

17. Rice LB. Federal funding for the study of antimicrobial resistance in nosocomial pathogens: no ESKAPE. *J Infect Dis*. 2008;197(8):1079-81. Epub 2008/04/19. doi: 10.1086/533452. PubMed PMID: 18419525.

18. Lax S, Gilbert JA. Hospital-associated microbiota and implications for nosocomial infections. *Trends Mol Med*. 2015;21(7):427-32. Epub 2015/04/25. doi: 10.1016/j.molmed.2015.03.005. PubMed PMID: 25907678.

19. Magill SS, Edwards JR, Bamberg W, Beldavs ZG, Dumyati G, Kainer MA, Lynfield R, Maloney M, McAllister-Hollod L, Nadle J, Ray SM, Thompson DL, Wilson LE, Fridkin SK, Emerging Infections Program Healthcare-Associated I, Antimicrobial Use Prevalence Survey T. Multistate point-prevalence survey of health care-associated infections. *N Engl J Med*. 2014;370(13):1198-208. Epub 2014/03/29. doi: 10.1056/NEJMoa1306801. PubMed PMID: 24670166; PMCID: PMC4648343.

20. Renner LD, Zan J, Hu LI, Martinez M, Resto PJ, Siegel AC, Torres C, Hall SB, Slezak TR, Nguyen TH, Weibel DB. Detection of ESKAPE Bacterial Pathogens at the Point of Care Using Isothermal DNA-Based Assays in a Portable Degas-Actuated Microfluidic Diagnostic Assay Platform. *Appl Environ Microbiol*. 2017;83(4). Epub 2016/12/18. doi: 10.1128/AEM.02449-16. PubMed PMID: 27986722; PMCID: PMC5288812.

21. Wendt C, Dietze B, Dietz E, Ruden H. Survival of *Acinetobacter baumannii* on dry surfaces. *Journal of clinical microbiology*. 1997;35(6):1394-7. Epub 1997/06/01. PubMed PMID: 9163451; PMCID: PMC229756.
22. Byappanahalli MN, Nevers MB, Korajkic A, Staley ZR, Harwood VJ. Enterococci in the environment. *Microbiol Mol Biol Rev*. 2012;76(4):685-706. Epub 2012/12/04. doi: 10.1128/MMBR.00023-12. PubMed PMID: 23204362; PMCID: PMC3510518.
23. Hu H, Johani K, Gosbell IB, Jacombs AS, Almatroudi A, Whiteley GS, Deva AK, Jensen S, Vickery K. Intensive care unit environmental surfaces are contaminated by multidrug-resistant bacteria in biofilms: combined results of conventional culture, pyrosequencing, scanning electron microscopy, and confocal laser microscopy. *J Hosp Infect*. 2015;91(1):35-44. doi: 10.1016/j.jhin.2015.05.016. PubMed PMID: 26187533.
24. Mehta Y, Gupta A, Todi S, Myatra S, Samaddar DP, Patil V, Bhattacharya PK, Ramasubban S. Guidelines for prevention of hospital acquired infections. *Indian J Crit Care Med*. 2014;18(3):149-63. Epub 2014/04/05. doi: 10.4103/0972-5229.128705. PubMed PMID: 24701065; PMCID: PMC3963198.
25. WHO. Global Action Plan on Antimicrobial Resistance. World Health Organization, 2015.
26. ResistanceMap [Internet]. The Center for Disease Dynamics, Economics & Policy. 2017.
27. Laxminarayan R, Duse A, Wattal C, Zaidi AK, Wertheim HF, Sumpradit N, Vlieghe E, Hara GL, Gould IM, Goossens H, Greko C, So AD, Bigdeli M, Tomson G, Woodhouse W, Ombaka E, Peralta AQ, Qamar FN, Mir F, Kariuki S, Bhutta ZA, Coates

A, Bergstrom R, Wright GD, Brown ED, Cars O. Antibiotic resistance-the need for global solutions. *The Lancet Infectious diseases*. 2013;13(12):1057-98. Epub 2013/11/21. doi: 10.1016/S1473-3099(13)70318-9. PubMed PMID: 24252483.

28. Saleem AF, Ahmed I, Mir F, Ali SR, Zaidi AK. Pan-resistant *Acinetobacter* infection in neonates in Karachi, Pakistan. *J Infect Dev Ctries*. 2009;4(1):30-7. Epub 2010/02/05. PubMed PMID: 20130376.

29. Pesesky MW, Hussain T, Wallace M, Wang B, Andleeb S, Burnham CA, Dantas G. KPC and NDM-1 genes in related *Enterobacteriaceae* strains and plasmids from Pakistan and the United States. *Emerging infectious diseases*. 2015;21(6):1034-7. Epub 2015/05/20. doi: 10.3201/eid2106.141504. PubMed PMID: 25988236; PMCID: 4451916.

30. Potter RF, D'Souza AW, Wallace MA, Shupe A, Patel S, Gul D, Kwon JH, Beatty W, Andleeb S, Burnham CD, Dantas G. *Superficieibacter electus* gen. nov., sp. nov., an Extended-Spectrum beta-Lactamase Possessing Member of the *Enterobacteriaceae* Family, Isolated From Intensive Care Unit Surfaces. *Frontiers in microbiology*. 2018;9:1629. Epub 2018/08/07. doi: 10.3389/fmicb.2018.01629. PubMed PMID: 30079059; PMCID: PMC6062592.

31. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30(14):2068-9. Epub 2014/03/20. doi: 10.1093/bioinformatics/btu153. PubMed PMID: 24642063.

32. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J. Roary: rapid large-scale prokaryote pan genome

- analysis. *Bioinformatics*. 2015;31(22):3691-3. Epub 2015/07/23. doi:
10.1093/bioinformatics/btv421. PubMed PMID: 26198102; PMCID: PMC4817141.
33. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014;30(9):1312-3. doi:
10.1093/bioinformatics/btu033. PubMed PMID: 24451623; PMCID: PMC3998144.
34. Mostowy R, Croucher NJ, Andam CP, Corander J, Hanage WP, Marttinen P. Efficient Inference of Recent and Ancestral Recombination within Bacterial Populations. *Mol Biol Evol*. 2017;34(5):1167-82. doi: 10.1093/molbev/msx066. PubMed PMID: 28199698; PMCID: PMC5400400.
35. Kleinheinz KA, Joensen KG, Larsen MV. Applying the ResFinder and VirulenceFinder web-services for easy identification of acquired antibiotic resistance and *E. coli* virulence genes in bacteriophage and prophage nucleotide sequences. *Bacteriophage*. 2014;4(1):e27943. Epub 2014/02/28. doi: 10.4161/bact.27943. PubMed PMID: 24575358; PMCID: PMC3926868.
36. CLSI. Performance standards for antimicrobial susceptibility testing; 26th informational supplement. M100 S26:2016.: Clinical and Laboratory Standards Institute; 2016.
37. Roach DJ, Burton JN, Lee C, Stackhouse B, Butler-Wu SM, Cookson BT, Shendure J, Salipante SJ. A Year of Infection in the Intensive Care Unit: Prospective Whole Genome Sequencing of Bacterial Clinical Isolates Reveals Cryptic Transmissions and Novel Microbiota. *PLoS Genet*. 2015;11(7):e1005413. doi:
10.1371/journal.pgen.1005413. PubMed PMID: 26230489; PMCID: PMC4521703.

38. Břinda K, Callendrello A, Cowley L, Charalampous T, Lee RS, MacFadden DR, Kucherov G, O'Grady J, Baym M, Hanage WP. Lineage calling can identify antibiotic resistant clones within minutes. *bioRxiv*. 2018.
39. Kumarasamy KK, Toleman MA, Walsh TR, Bagaria J, Butt F, Balakrishnan R, Chaudhary U, Doumith M, Giske CG, Irfan S, Krishnan P, Kumar AV, Maharjan S, Mushtaq S, Noorie T, Paterson DL, Pearson A, Perry C, Pike R, Rao B, Ray U, Sarma JB, Sharma M, Sheridan E, Thirunarayan MA, Turton J, Upadhyay S, Warner M, Welfare W, Livermore DM, Woodford N. Emergence of a new antibiotic resistance mechanism in India, Pakistan, and the UK: a molecular, biological, and epidemiological study. *The Lancet Infectious diseases*. 2010;10(9):597-602. Epub 2010/08/14. doi: 10.1016/S1473-3099(10)70143-2. PubMed PMID: 20705517; PMCID: PMC2933358.
40. Russotto V, Cortegiani A, Raineri SM, Giarratano A. Bacterial contamination of inanimate surfaces and equipment in the intensive care unit. *J Intensive Care*. 2015;3:54. doi: 10.1186/s40560-015-0120-5. PubMed PMID: 26693023; PMCID: PMC4676153.
41. Hota B. Contamination, disinfection, and cross-colonization: are hospital surfaces reservoirs for nosocomial infection? *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*. 2004;39(8):1182-9. Epub 2004/10/16. doi: 10.1086/424667. PubMed PMID: 15486843.
42. Sillanpaa J, Nallapareddy SR, Singh KV, Prakash VP, Fothergill T, Ton-That H, Murray BE. Characterization of the *ebp(fm)* pilus-encoding operon of *Enterococcus faecium* and its role in biofilm formation and virulence in a murine model of urinary tract

infection. *Virulence*. 2010;1(4):236-46. Epub 2010/08/03. PubMed PMID: 20676385; PMCID: PMC2910428.

43. Iwashkiw JA, Seper A, Weber BS, Scott NE, Vinogradov E, Stratilo C, Reiz B, Cordwell SJ, Whittal R, Schild S, Feldman MF. Identification of a general O-linked protein glycosylation system in *Acinetobacter baumannii* and its role in virulence and biofilm formation. *PLoS Pathog*. 2012;8(6):e1002758. doi: 10.1371/journal.ppat.1002758. PubMed PMID: 22685409; PMCID: PMC3369928.

44. Sydnor ER, Perl TM. Hospital epidemiology and infection control in acute-care settings. *Clinical microbiology reviews*. 2011;24(1):141-73. Epub 2011/01/15. doi: 10.1128/CMR.00027-10. PubMed PMID: 21233510; PMCID: PMC3021207.

45. Kampfer P, Glaeser SP, Raza MW, Abbasi SA, Perry JD. *Pseudocitrobacter* gen. nov., a novel genus of the Enterobacteriaceae with two new species *Pseudocitrobacter faecalis* sp. nov., and *Pseudocitrobacter anthropi* sp. nov, isolated from fecal samples from hospitalized patients in Pakistan. *Syst Appl Microbiol*. 2014;37(1):17-22. Epub 2013/11/05. doi: 10.1016/j.syapm.2013.08.003. PubMed PMID: 24182752.

46. Bhatti MD, Kalia A, Sahasrabhojane P, Kim J, Greenberg DE, Shelburne SA. Identification and Whole Genome Sequencing of the First Case of *Kosakonia radicincitans* Causing a Human Bloodstream Infection. *Front Microbiol*. 2017;8:62. doi: 10.3389/fmicb.2017.00062. PubMed PMID: 28174569; PMCID: PMC5258702.

47. Pecora ND, Li N, Allard M, Li C, Albano E, Delaney M, Dubois A, Onderdonk AB, Bry L. Genomically Informed Surveillance for Carbapenem-Resistant Enterobacteriaceae in a Health Care System. *mBio*. 2015;6(4):e01030. Epub

2015/07/30. doi: 10.1128/mBio.01030-15. PubMed PMID: 26220969; PMCID: PMC4551976.

48. Cella E, Ciccozzi M, Lo Presti A, Fogolari M, Azarian T, Prospero M, Salemi M, Equestre M, Antonelli F, Conti A, Cesaris M, Spoto S, Incalzi RA, Coppola R, Dicuonzo G, Angeletti S. Multi-drug resistant *Klebsiella pneumoniae* strains circulating in hospital setting: whole-genome sequencing and Bayesian phylogenetic analysis for outbreak investigations. *Scientific reports*. 2017;7(1):3534. Epub 2017/06/16. doi: 10.1038/s41598-017-03581-4. PubMed PMID: 28615687; PMCID: PMC5471223.

49. Saharman YR, Karuniawati A, Sedono R, Aditjaningsih D, Sudarmono P, Goessens WHF, Klaassen CHW, Verbrugh HA, Severin JA. Endemic carbapenem-nonsusceptible *Acinetobacter baumannii-calcoaceticus* complex in intensive care units of the national referral hospital in Jakarta, Indonesia. *Antimicrob Resist Infect Control*. 2018;7:5. Epub 2018/01/19. doi: 10.1186/s13756-017-0296-7. PubMed PMID: 29344351; PMCID: PMC5767053.

50. Ning NZ, Liu X, Bao CM, Chen SM, Cui EB, Zhang JL, Huang J, Chen FH, Li T, Qu F, Wang H. Molecular epidemiology of bla OXA-23 -producing carbapenem-resistant *Acinetobacter baumannii* in a single institution over a 65-month period in north China. *BMC infectious diseases*. 2017;17(1):14. Epub 2017/01/07. doi: 10.1186/s12879-016-2110-1. PubMed PMID: 28056839; PMCID: PMC5217423.

51. Freitas AR, Coque TM, Novais C, Hammerum AM, Lester CH, Zervos MJ, Donabedian S, Jensen LB, Francia MV, Baquero F, Peixe L. Human and swine hosts share vancomycin-resistant *Enterococcus faecium* CC17 and CC5 and *Enterococcus*

faecalis CC2 clonal clusters harboring Tn1546 on indistinguishable plasmids. *Journal of clinical microbiology*. 2011;49(3):925-31. Epub 2011/01/14. doi: 10.1128/JCM.01750-

10. PubMed PMID: 21227995; PMCID: PMC3067689.

52. Freitas AR, Novais C, Ruiz-Garbajosa P, Coque TM, Peixe L. Dispersion of multidrug-resistant *Enterococcus faecium* isolates belonging to major clonal complexes in different Portuguese settings. *Appl Environ Microbiol*. 2009;75(14):4904-8. Epub 2009/05/19. doi: 10.1128/AEM.02945-08. PubMed PMID: 19447948; PMCID: PMC2708421.

53. Weingarten RA, Johnson RC, Conlan S, Ramsburg AM, Dekker JP, Lau AF, Khil P, Odom RT, Deming C, Park M, Thomas PJ, Program NCS, Henderson DK, Palmore TN, Segre JA, Frank KM. Genomic Analysis of Hospital Plumbing Reveals Diverse Reservoir of Bacterial Plasmids Conferring Carbapenem Resistance. *MBio*. 2018;9(1). Epub 2018/02/14. doi: 10.1128/mBio.02011-17. PubMed PMID: 29437920; PMCID: PMC5801463.

54. Arias CA, Murray BE. The rise of the *Enterococcus*: beyond vancomycin resistance. *Nat Rev Microbiol*. 2012;10(4):266-78. doi: 10.1038/nrmicro2761. PubMed PMID: 22421879; PMCID: PMC3621121.

55. Zhao S, Lieberman TD, Poyet M, Kauffman KM, Gibbons SM, Groussin M, Xavier RJ, Alm EJ. Adaptive Evolution within Gut Microbiomes of Healthy People. *Cell Host Microbe*. 2019;25(5):656-67 e8. Epub 2019/04/28. doi: 10.1016/j.chom.2019.03.007. PubMed PMID: 31028005.

56. Soto-Giron MJ, Rodriguez RL, Luo C, Elk M, Ryu H, Hoelle J, Santo Domingo JW, Konstantinidis KT. Biofilms on Hospital Shower Hoses: Characterization and Implications for Nosocomial Infections. *Appl Environ Microbiol*. 2016;82(9):2872-83. Epub 2016/03/13. doi: 10.1128/AEM.03529-15. PubMed PMID: 26969701; PMCID: PMC4836434.
57. Kizny Gordon AE, Mathers AJ, Cheong EYL, Gottlieb T, Kotay S, Walker AS, Peto TEA, Crook DW, Stoesser N. The Hospital Water Environment as a Reservoir for Carbapenem-Resistant Organisms Causing Hospital-Acquired Infections-A Systematic Review of the Literature. *Clin Infect Dis*. 2017;64(10):1435-44. Epub 2017/02/16. doi: 10.1093/cid/cix132. PubMed PMID: 28200000.
58. Gorrie CL, Mirceta M, Wick RR, Edwards DJ, Thomson NR, Strugnell RA, Pratt NF, Garlick JS, Watson KM, Pilcher DV, McGloughlin SA, Spelman DW, Jenney AWJ, Holt KE. Gastrointestinal Carriage Is a Major Reservoir of *Klebsiella pneumoniae* Infection in Intensive Care Patients. *Clin Infect Dis*. 2017;65(2):208-15. doi: 10.1093/cid/cix270. PubMed PMID: 28369261; PMCID: PMC5850561.
59. Barrasa-Villar JI, Aibar-Remon C, Prieto-Andres P, Mareca-Donate R, Moliner-Lahoz J. Impact on Morbidity, Mortality, and Length of Stay of Hospital-Acquired Infections by Resistant Microorganisms. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*. 2017;65(4):644-52. Epub 2017/05/05. doi: 10.1093/cid/cix411. PubMed PMID: 28472416.
60. Nowak J, Zander E, Stefanik D, Higgins PG, Roca I, Vila J, McConnell MJ, Cisneros JM, Seifert H, MagicBullet Working Group WP. High incidence of pandrug-

resistant *Acinetobacter baumannii* isolates collected from patients with ventilator-associated pneumonia in Greece, Italy and Spain as part of the MagicBullet clinical trial. *The Journal of antimicrobial chemotherapy*. 2017;72(12):3277-82. Epub 2017/09/30. doi: 10.1093/jac/dkx322. PubMed PMID: 28961773; PMCID: PMC5890771.

61. Wailan AM, Sartor AL, Zowawi HM, Perry JD, Paterson DL, Sidjabat HE. Genetic Contexts of blaNDM-1 in Patients Carrying Multiple NDM-Producing Strains. *Antimicrobial agents and chemotherapy*. 2015;59(12):7405-10. Epub 2015/09/24. doi: 10.1128/AAC.01319-15. PubMed PMID: 26392493; PMCID: 4649221.

62. Habimana O, Heir E, Langsrud S, Asli AW, Moretro T. Enhanced surface colonization by *Escherichia coli* O157:H7 in biofilms formed by an *Acinetobacter calcoaceticus* isolate from meat-processing environments. *Appl Environ Microbiol*. 2010;76(13):4557-9. Epub 2010/05/11. doi: 10.1128/AEM.02707-09. PubMed PMID: 20453142; PMCID: PMC2897464.

63. Giaouris E, Chorianopoulos N, Doulgeraki A, Nychas GJ. Co-culture with *Listeria monocytogenes* within a dual-species biofilm community strongly increases resistance of *Pseudomonas putida* to benzalkonium chloride. *PloS one*. 2013;8(10):e77276. Epub 2013/10/17. doi: 10.1371/journal.pone.0077276. PubMed PMID: 24130873; PMCID: PMC3795059.

64. Makovcova J, Babak V, Kulich P, Masek J, Slany M, Cincarova L. Dynamics of mono- and dual-species biofilm formation and interactions between *Staphylococcus aureus* and Gram-negative bacteria. *Microb Biotechnol*. 2017;10(4):819-32. Epub

2017/04/13. doi: 10.1111/1751-7915.12705. PubMed PMID: 28401747; PMCID: PMC5481519.

65. Mohamed JA, Huang DB. Biofilm formation by enterococci. *J Med Microbiol.* 2007;56(Pt 12):1581-8. Epub 2007/11/24. doi: 10.1099/jmm.0.47331-0. PubMed PMID: 18033823.

66. Wong D, Nielsen TB, Bonomo RA, Pantapalangkoor P, Luna B, Spellberg B. Clinical and Pathophysiological Overview of Acinetobacter Infections: a Century of Challenges. *Clin Microbiol Rev.* 2017;30(1):409-47. Epub 2016/12/16. doi: 10.1128/CMR.00058-16. PubMed PMID: 27974412; PMCID: PMC5217799.

67. Alvarez-Fraga L, Perez A, Rumbo-Feal S, Merino M, Vallejo JA, Ohneck EJ, Edelmann RE, Beceiro A, Vazquez-Ucha JC, Valle J, Actis LA, Bou G, Poza M. Analysis of the role of the LH92_11085 gene of a biofilm hyper-producing *Acinetobacter baumannii* strain on biofilm formation and attachment to eukaryotic cells. *Virulence.* 2016;7(4):443-55. Epub 2016/02/09. doi: 10.1080/21505594.2016.1145335. PubMed PMID: 26854744; PMCID: PMC4871663.

68. Schweppe DK, Harding C, Chavez JD, Wu X, Ramage E, Singh PK, Manoil C, Bruce JE. Host-Microbe Protein Interactions during Bacterial Infection. *Chem Biol.* 2015;22(11):1521-30. Epub 2015/11/10. doi: 10.1016/j.chembiol.2015.09.015. PubMed PMID: 26548613; PMCID: PMC4756654.

69. Montealegre MC, Singh KV, Somarajan SR, Yadav P, Chang C, Spencer R, Sillanpaa J, Ton-That H, Murray BE. Role of the Emp Pilus Subunits of *Enterococcus faecium* in Biofilm Formation, Adherence to Host Extracellular Matrix Components, and

Experimental Infection. *Infect Immun*. 2016;84(5):1491-500. doi: 10.1128/IAI.01396-15. PubMed PMID: 26930703; PMCID: PMC4862714.

70. Pour NK, Dusane DH, Dhakephalkar PK, Zamin FR, Zinjarde SS, Chopade BA. Biofilm formation by *Acinetobacter baumannii* strains isolated from urinary tract infection and urinary catheters. *FEMS Immunol Med Microbiol*. 2011;62(3):328-38. Epub 2011/05/17. doi: 10.1111/j.1574-695X.2011.00818.x. PubMed PMID: 21569125.

71. Lim SY, Teh CSJ, Thong KL. Biofilm-Related Diseases and Omics: Global Transcriptional Profiling of *Enterococcus faecium* Reveals Different Gene Expression Patterns in the Biofilm and Planktonic Cells. *OMICS*. 2017;21(10):592-602. Epub 2017/10/20. doi: 10.1089/omi.2017.0119. PubMed PMID: 29049010.

72. Rampelotto PH, Sereia AFR, de Oliveira LFV, Margis R. Exploring the Hospital Microbiome by High-Resolution 16S rRNA Profiling. *Int J Mol Sci*. 2019;20(12). Epub 2019/06/28. doi: 10.3390/ijms20123099. PubMed PMID: 31242612.

73. Lee K, Lee KM, Kim D, Yoon SS. Molecular Determinants of the Thickened Matrix in a Dual-Species *Pseudomonas aeruginosa* and *Enterococcus faecalis* Biofilm. *Appl Environ Microbiol*. 2017;83(21). Epub 2017/08/27. doi: 10.1128/AEM.01182-17. PubMed PMID: 28842537; PMCID: PMC5648906.

74. Westblade LF, Garner OB, MacDonald K, Bradford C, Pincus DH, Mochon AB, Jennemann R, Manji R, Bythrow M, Lewinski MA, Burnham CA, Ginocchio CC. Assessment of Reproducibility of Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry for Bacterial and Yeast Identification. *Journal of clinical*

microbiology. 2015;53(7):2349-52. Epub 2015/05/01. doi: 10.1128/JCM.00187-15.
PubMed PMID: 25926486; PMCID: PMC4473194.

75. McElvania TeKippe E, Burnham CA. Evaluation of the Bruker Biotyper and VITEK MS MALDI-TOF MS systems for the identification of unusual and/or difficult-to-identify microorganisms isolated from clinical specimens. *European journal of clinical microbiology & infectious diseases* : official publication of the European Society of Clinical Microbiology. 2014;33(12):2163-71. Epub 2014/06/26. doi: 10.1007/s10096-014-2183-y. PubMed PMID: 24962194.

76. Branda JA, Rychert J, Burnham CA, Bythrow M, Garner OB, Ginocchio CC, Jennemann R, Lewinski MA, Manji R, Mochon AB, Procop GW, Richter SS, Sercia LF, Westblade LF, Ferraro MJ. Multicenter validation of the VITEK MS v2.0 MALDI-TOF mass spectrometry system for the identification of fastidious gram-negative bacteria. *Diagn Microbiol Infect Dis*. 2014;78(2):129-31. Epub 2013/12/11. doi: 10.1016/j.diagmicrobio.2013.08.013. PubMed PMID: 24321357.

77. Manji R, Bythrow M, Branda JA, Burnham CA, Ferraro MJ, Garner OB, Jennemann R, Lewinski MA, Mochon AB, Procop GW, Richter SS, Rychert JA, Sercia L, Westblade LF, Ginocchio CC. Multi-center evaluation of the VITEK(R) MS system for mass spectrometric identification of non-Enterobacteriaceae Gram-negative bacilli. *European journal of clinical microbiology & infectious diseases* : official publication of the European Society of Clinical Microbiology. 2014;33(3):337-46. Epub 2013/09/11. doi: 10.1007/s10096-013-1961-2. PubMed PMID: 24019163.

78. Richter SS, Sercia L, Branda JA, Burnham CA, Bythrow M, Ferraro MJ, Garner OB, Ginocchio CC, Jennemann R, Lewinski MA, Manji R, Mochon AB, Rychert JA, Westblade LF, Procop GW. Identification of Enterobacteriaceae by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry using the VITEK MS system. *European journal of clinical microbiology & infectious diseases* : official publication of the European Society of Clinical Microbiology. 2013;32(12):1571-8. Epub 2013/07/03. doi: 10.1007/s10096-013-1912-y. PubMed PMID: 23818163.
79. Hink T, Burnham CA, Dubberke ER. A systematic evaluation of methods to optimize culture-based recovery of *Clostridium difficile* from stool specimens. *Anaerobe*. 2013;19:39-43. Epub 2012/12/19. doi: 10.1016/j.anaerobe.2012.12.001. PubMed PMID: 23247066; PMCID: 4146438.
80. Baym M, Kryazhimskiy S, Lieberman TD, Chung H, Desai MM, Kishony R. Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS One*. 2015;10(5):e0128036. Epub 2015/05/23. doi: 10.1371/journal.pone.0128036. PubMed PMID: 26000737; PMCID: PMC4441430.
81. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-20. Epub 2014/04/04. doi: 10.1093/bioinformatics/btu170. PubMed PMID: 24695404; PMCID: PMC4103590.
82. Schmieder R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS One*. 2011;6(3):e17288. Epub 2011/03/17. doi: 10.1371/journal.pone.0017288. PubMed PMID: 21408061; PMCID: PMC3052304.

83. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 2012;19(5):455-77. Epub 2012/04/18. doi: 10.1089/cmb.2012.0021. PubMed PMID: 22506599; PMCID: PMC3342519.
84. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUASt: quality assessment tool for genome assemblies. *Bioinformatics.* 2013;29(8):1072-5. doi: 10.1093/bioinformatics/btt086. PubMed PMID: 23422339; PMCID: PMC3624806.
85. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 2016;17(1):132. Epub 2016/06/22. doi: 10.1186/s13059-016-0997-x. PubMed PMID: 27323842; PMCID: PMC4915045.
86. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 2007;35(9):3100-8. Epub 2007/04/25. doi: 10.1093/nar/gkm160. PubMed PMID: 17452365; PMCID: PMC1888812.
87. Yoon SH, Ha SM, Kwon S, Lim J, Kim Y, Seo H, Chun J. Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *Int J Syst Evol Microbiol.* 2017;67(5):1613-7. doi: 10.1099/ijsem.0.001755. PubMed PMID: 28005526; PMCID: PMC5563544.

88. Richter M, Rossello-Mora R, Oliver Glockner F, Peplies J. JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. *Bioinformatics*. 2016;32(6):929-31. Epub 2015/11/19. doi: 10.1093/bioinformatics/btv681. PubMed PMID: 26576653; PMCID: PMC5939971.
89. Janda JM, Abbott SL. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*. 2007;45(9):2761-4. Epub 2007/07/13. doi: 10.1128/JCM.01228-07. PubMed PMID: 17626177; PMCID: PMC2045242.
90. Richter M, Rossello-Mora R. Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences of the United States of America*. 2009;106(45):19126-31. Epub 2009/10/27. doi: 10.1073/pnas.0906412106. PubMed PMID: 19855009; PMCID: PMC2776425.
91. Loytynoja A. Phylogeny-aware alignment with PRANK. *Methods Mol Biol*. 2014;1079:155-70. doi: 10.1007/978-1-62703-646-7_10. PubMed PMID: 24170401.
92. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res*. 2016;44(W1):W242-5. doi: 10.1093/nar/gkw290. PubMed PMID: 27095192; PMCID: PMC4987883.
93. Cheng L, Connor TR, Siren J, Aanensen DM, Corander J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol Biol Evol*. 2013;30(5):1224-8. doi: 10.1093/molbev/mst028. PubMed PMID: 23408797; PMCID: PMC3670731.

94. Knaus BJ, Grunwald NJ. vcfr: a package to manipulate and visualize variant call format data in R. *Mol Ecol Resour.* 2017;17(1):44-53. Epub 2016/07/13. doi: 10.1111/1755-0998.12549. PubMed PMID: 27401132.
95. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003;13(11):2498-504. doi: 10.1101/gr.1239303. PubMed PMID: 14597658; PMCID: PMC403769.
96. Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer. *Bioinformatics.* 2011;27(7):1009-10. Epub 2011/02/01. doi: 10.1093/bioinformatics/btr039. PubMed PMID: 21278367; PMCID: PMC3065679.
97. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. *BMC Bioinformatics.* 2009;10:421. doi: 10.1186/1471-2105-10-421. PubMed PMID: 20003500; PMCID: PMC2803857.

Chapter 5: *In silico* analysis of *Gardnerella* genomospecies detected in the setting of bacterial vaginosis

5.1 Abstract

Gardnerella vaginalis is implicated as one of the causative agents of bacterial vaginosis, but it can also be isolated from the vagina of healthy women. Previous efforts to study *G. vaginalis* identified 4-6 clades but average nucleotide identity analysis indicates that *G. vaginalis* may be multiple species. Recently, *Gardnerella* was determined to be 13 genomospecies, with *Gardnerella piottii*, *Gardnerella leopoldii*, and *Gardnerella swidsinkii* delineated as separate species. We accessed 103 publicly available genomes annotated as *G. vaginalis*. We performed comprehensive taxonomic and phylogenomic analysis to quantify the number of species called *G. vaginalis*, the similarity of their core-genes, and their burden of their accessory genes. We additionally analyzed publicly available metatranscriptomic datasets of bacterial vaginosis to determine if the newly delineated genomospecies are present, and to identify putative conserved features of *Gardnerella* pathogenesis. *Gardnerella* could be classified into 8-14 genomospecies depending on the *in silico* classification tools used. Consensus classification identified nine different *Gardnerella* genomospecies, here annotated as GS01-GS09. The genomospecies could be readily distinguished by the phylogeny of their shared genes and burden of accessory genes. All of the new genomospecies were identified in metatranscriptomes of bacterial vaginosis. Multiple *Gardnerella* genomospecies operating in isolation or in concert with one another may be responsible for bacterial vaginosis. These results have important implications for future efforts to understand the evolution of the *Gardnerella* genomospecies, host-pathogen interactions

of the genomospecies during bacterial vaginosis, diagnostic assay development for bacterial vaginosis, and metagenomic investigations of the vaginal microbiota.

5.2 Introduction

Bacterial vaginosis (BV) is a common infectious disease of women, often caused by *Gardnerella vaginalis* (1, 2). However, *G. vaginalis* has also been identified in healthy women without BV (3, 4). One explanation is that certain strains of *G. vaginalis* are more pathogenic than others. Genome-based taxonomic methods, which have delineated novel species in other genera, have scarcely been applied to *G. vaginalis*. Importantly, one recent investigation found that average nucleotide identity (ANI) values between different *G. vaginalis* subgroups were below the species cutoff of 96%, indicating *G. vaginalis* may be multiple species (5). Recently, using ANI and digital DNA-DNA hybridization assays, it was found that 13 different *Gardnerella* genomospecies may currently be annotated as *G. vaginalis* (6). Three of these species were fully elucidated using phenotypic assays and termed *Gardnerella piovii*, *Gardnerella leopoldii*, and *Gardnerella swidsinkii* (7).

Historically, delineation of new bacteria taxa has relied on phenotypic differences between strains such as chemical analysis or biochemical utilization characteristics, with the laborious DNA-DNA hybridization assay representing the gold standard analysis for species-level determination (7). Using 16S rRNA gene sequencing, a cut off of 97% is typically used to delineate bacteria, but given that recognized different species can have values greater than 97% similarity, whole-genome data such as ANI are often also used (7, 8). ANI values $\geq 96\%$ are used as thresholds for species-level designations (9). In the absence of an isolated organism, putative novel species can be determined

by genetic content alone but are termed genomospecies. In addition, recognized species can also be re-classified such as *Escherichia hermannii* and *Salmonella subterranea* to *Atlantibacter hermannii* and *Atlantibacter subterranea*, respectively (10). Given the previous analysis on *Gardnerella*, we used multiple *in silico* taxonomic classification tools to bin publicly available *Gardnerella* genomes into different genomospecies and then performed comparative analysis between the genomospecies. To address the knowledge gap regarding the taxonomic diversity and relatedness within genomes currently classified as *G. vaginalis*, we performed a retrospective comparative analysis using 103 publicly available genomes as well as BV metatranscriptomes. Based on the observation that multiple *G. vaginalis* genomes may be related by ANI values less than the species cut-off of 96%, we hypothesized that multiple distinct genomospecies have been collapsed into a single *G. vaginalis* species annotation. Further, we hypothesized that these genomospecies could be distinguished by the relatedness of their shared genes and the differential burden of their accessory genes.

5.3 Results

5.3.1 *In silico* tool-dependent classification of *G. vaginalis* into eight to fourteen genomospecies

01- *G. vaginalis* ATCC 14018(T) 06- *G. swidsinkii* GS 9838-1(T) 11- GED7760B
 02- JCP8108 07- JCP8481A 12- CMW7778B
 03- JCP8017A 08- UMB1686 13- KA00225
 04- *G. piovii* UGENT 18.01(T) 09- 6119V5 14- NR010
 05- *G. leopoldii* UGENT 06.41(T) 10- 1500E

— Same genomospecies

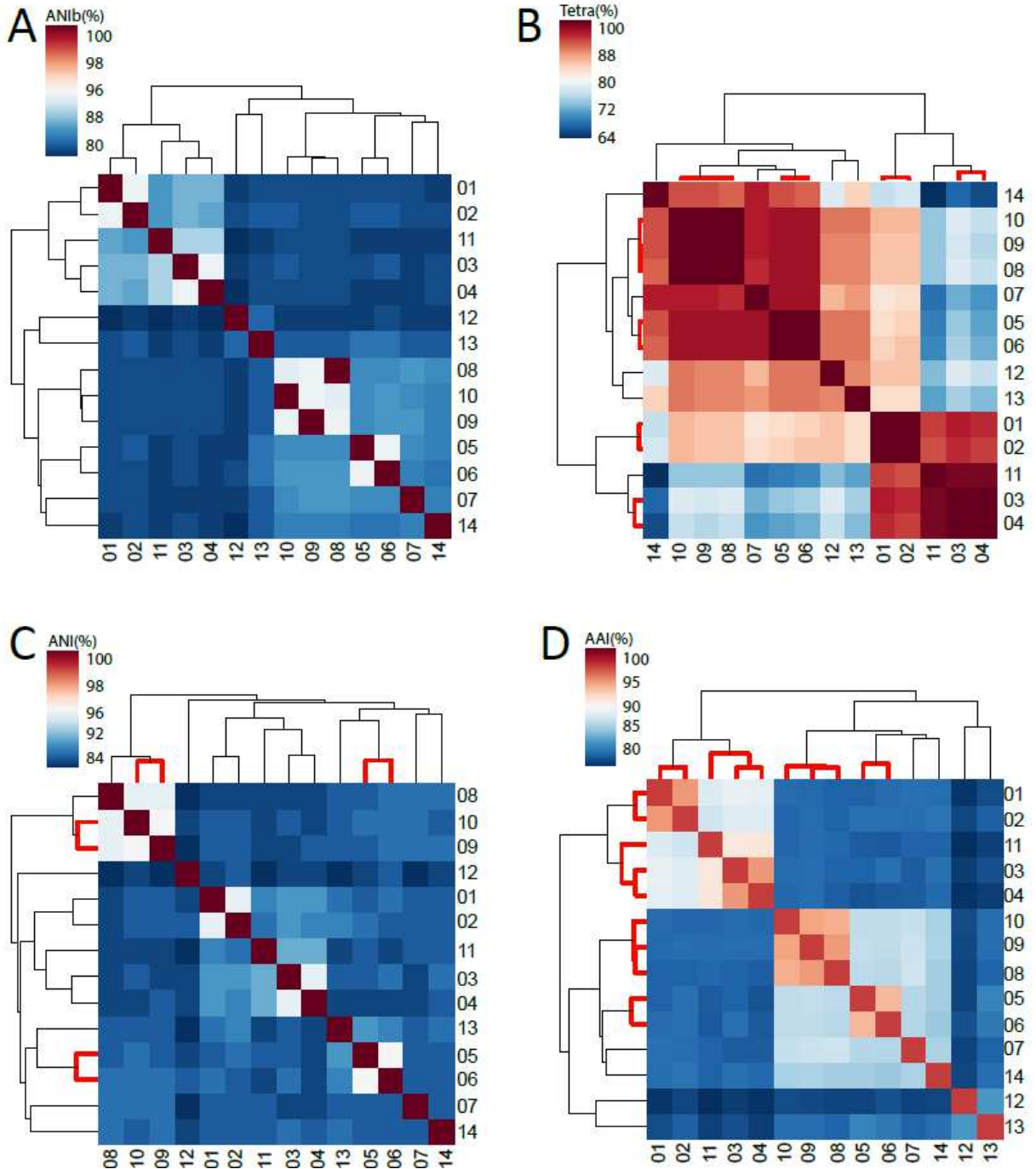


Figure 5.3.1 Different *in silico* taxonomic tools produce 8-14 *Gardnerella* genomospecies

Heatmaps with hierarchal clustering of the pairwise values for JSpeciesWS ANIb (A), JSpeciesWS tetranucleotide frequency (B), Kostas Lab ANI (C), and Kostas Lab AAI (D) for the 14 representative *Gardnerella* strains. All strains are assigned a numerical identifier. Red lines in the clustering pattern indicate groups of 2 or more genomes that represent the same genomospecies for that tool. The ANI (A & C) and AAI (D) plots are colored so that the taxonomic cutoff of 96% or 90% respectively are white. Pairwise values above this cutoff in the red spectrum and values below in the blue spectrum.

We began analysis by using pyANI with the mummer nucleotide alignment method to determine pairwise ANI values between the 103 genomes obtained from NBCI. This analysis indicated that in addition to the 13 genomospecies delineated by Vaneechoutte et al (6), strain NR010 may represent a 14th genomospecies as it did not have any ANI values $\geq 96\%$ to any other genome. From this we found that *Gardnerella* may contain a maximum of 14 genomospecies. We then used multiple publicly available tools for verification. Therefore, to further assess if genomes annotated as the species *G. vaginalis* represent multiple genomospecies, we used two additional different ANI platforms, tetranucleotide frequency, and AAI tools to delineate the genomes into genomospecies (9, 14, 29, 30). We chose 13 genomes from the recent delineation of *G. vaginalis* including the type genomes for *G. vaginalis*, *G. piovii*, *G. leopoldii*, and *G. swidsinkii*, as well as NR010 (6) (Table 1). Our results showed that the number of annotated genomospecies ranged from 8-14, depending on the classification tools employed (Figure 5.3.1). ANI with the BLAST nucleotide alignment method (ANIb) from JSpeciesWS indicated that these genomes represented 14 unique genomospecies (Figure 5.3.1A). In comparison, classification by tetranucleotide frequency from JSpeciesWS, found that the 14 genomes are 9 genomospecies (Figure 5.3.1B). Tetranucleotide frequency-based classification found that the type strains *G. leopoliddi* UGENT 06.41 (T) and *G. swidsinkii* GS 9838-1 (T) may be the same genomospecies. A separate ANI classifier (ANI calculator from the Kostas lab) indicated that the 14 genomes instead represented 12 genomospecies, and again that the type strains *G. leopoliddi* UGENT 06.41 (T) and *G. swidsinkii* GS 9838-1 (T) may be the same genomospecies (Figure 5.3.1C). Finally, an AAI classifier (AAI calculator from the

Kostas lab) produced the most conservative estimate of the number of genomospecies, identifying 8 genomospecies from the 14 genomes (Figure 5.3.1D). The AAI classifications were concordant with the JSpeciesWS tetranucleotide frequency-based classifications in that *G. vaginalis* ATCC 14018 (T)/JCP8108, *G. piotti* UGENT 18.01 (T)/JCP8017A, and UMB1686/6119V5/1500E were the same genomospecies, respectively. The AAI calculator was the only tool that considered GED7760B the same genomospecies as *G. piotti* UGENT 18.01 (T)/JCP8017A.

We adopted a conservative consensus approach for genomospecies classification for the remainder of our analysis. Specifically, if two or more of the aforementioned tools indicated that the genomes represent the same genomospecies then we counted them as the same. This method had exact concordance with the tetranucleotide frequency tool classification and yielded nine *Gardnerella* genomospecies (GS01-GS09) (Table 5.3.1). All comparative analyses and biological conclusions hereafter are based on these nine genomospecies.

5.3.2 Core-genome alignment support relatedness of the genomospecies into 8 clades

To gain further insight into the taxonomic structure of the *Gardnerella* genus, we determined the 200 core-genes (the loci present in 100% of strains) at 70% nucleotide identity with the pan-genome tool Roary and aligned these genes with PRANK to create a core-gene alignment. We used FastTree to construct an approximate maximum likelihood tree from the core-genome alignment, which depicted the evolutionary relationship between all genomes analyzed and provided a confidence value for every branch point (Figure 5.3.2A). The tree had 100% bootstrap support values at the major

branch points, indicating a high degree of confidence on the relatedness of the *Gardnerella* genomospecies to one another. Midpoint rooting of the tree in iTOL depicted a major split within the genus between GS01/GS02/GS06 and GS03/GS04/GS05/GS07/GS08/GS09. Lineage identification using FastGear/BAPS on the core-genome alignment identified eight major lineages that had almost exact concordance with the consensus delineation into genomospecies, except that GS02 and the single genome GED7760B (GS06) were determined to be in the same lineage (Figure 5.3.2A). FastGear initially assigns clusters with the BAPS software and then uses an additional allele comparison to produce more refined groups. The single genomes/genomospecies KA00225 (GS08) and NR010 (GS09) were counted as their own lineages, indicating that the allele frequencies between GS06 and GS02 may be similar enough compared to the background comparisons that they are linked into the same lineage.

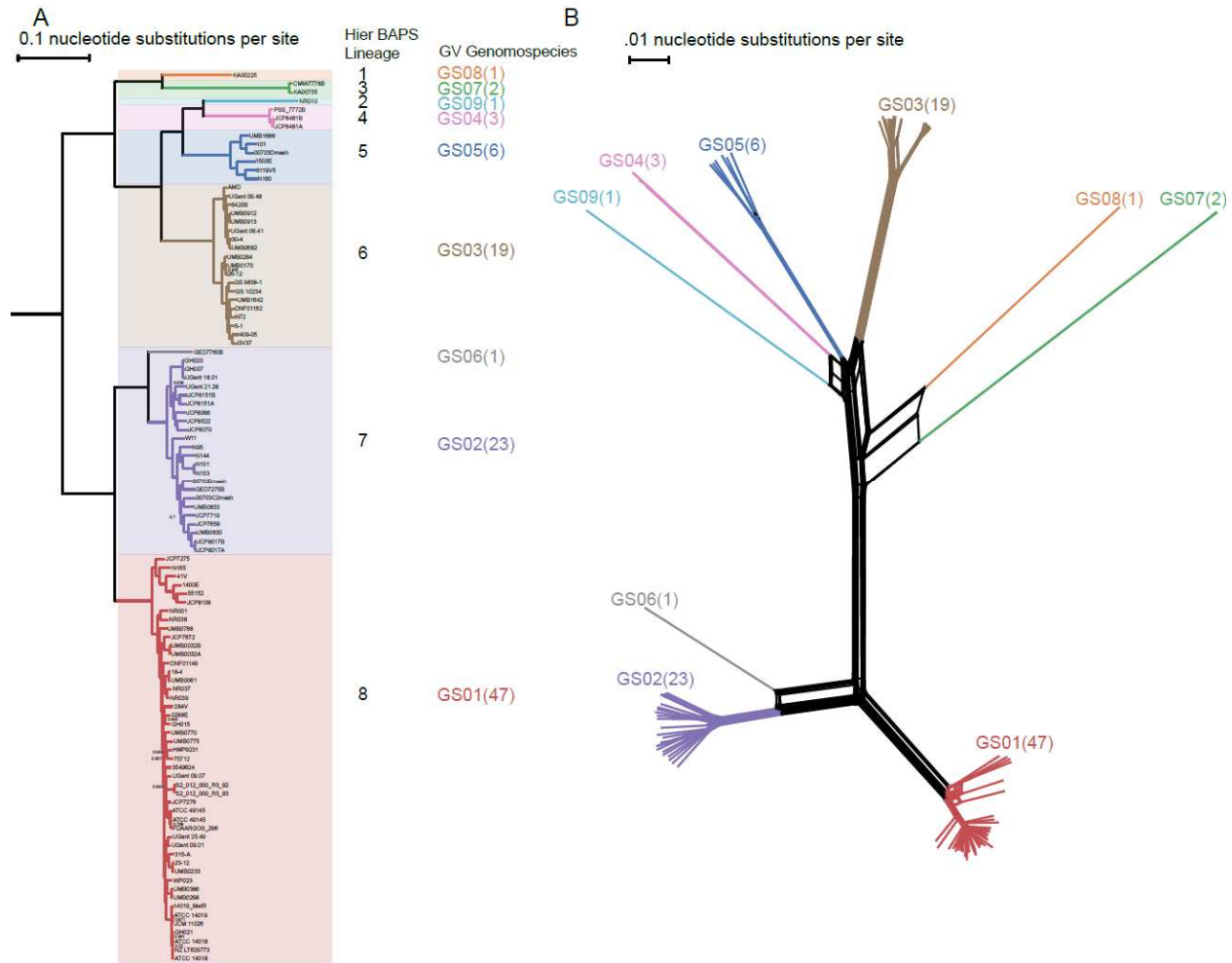


Figure 5.3.2 Core-genome phylogenetic analysis shows the genospecies fall into 9 distinct clusters

(A) Approximate maximum likelihood phylogenetic tree from PRANK alignment of the 200 core-genes identified by Roary with FastGear/BAPs lineages annotated adjacent to the tree. Groups of genomes that represent the same genospecies identified by the conservative consensus approach are colored. (B) Nearest neighbor network of the core-genome alignment with genospecies annotated as tip labels.

As a second method to view the relatedness of the genospecies, we visualized the core-genome alignment file as a nearest neighbor network in SplitsTrees (Figure 5.3.2B). The clustering pattern of the isolates were visually concordant with the maximum likelihood tree. Importantly, the isolates from GS07 and GS08 deviated away

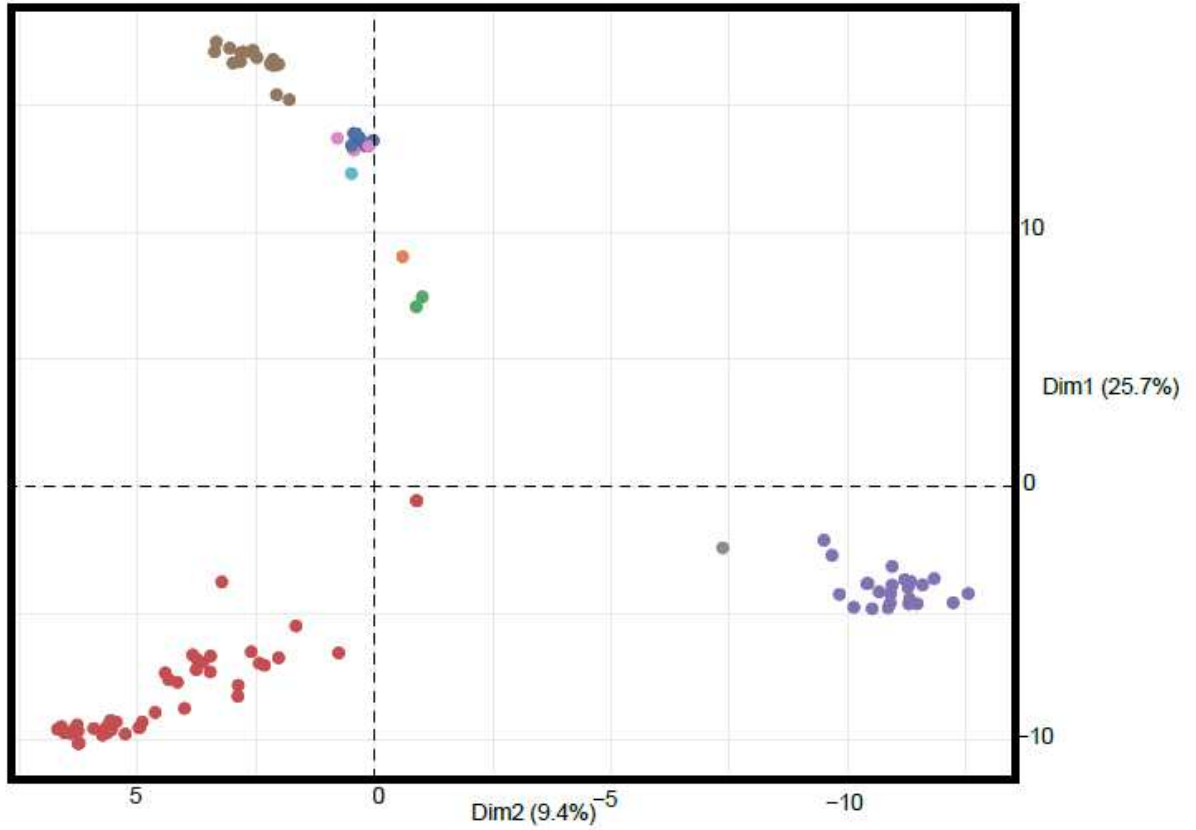
from the center of the network, providing additional evidence on their separation from the group containing GS03, GS04, GS05, and GS09 (Figure 5.3.2B).

5.3.3 *Gardnerella* genomospecies have distinct accessory gene repertoires

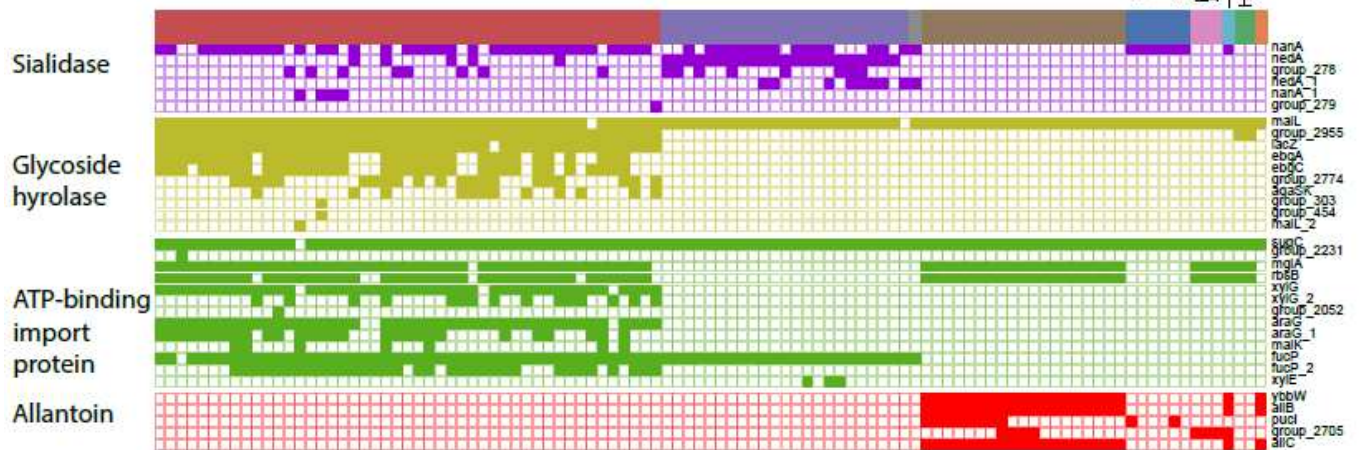
To understand the differential burden of accessory genes that may contribute to niche adaptation within the vaginal microenvironment and/or to BV pathology, we performed a principal component analysis on the presence/absence matrix of non-core genes identified by Roary. PERMANOVA using adonis2 from the vegan package in RStudio indicated that genomospecies and accessory genome composition were significantly ($p < .00001$) linked. Superposition of the genomospecies classification onto the principal component analysis plot demonstrated that the accessory gene content for GS04, GS05, and GS09 were remarkably similar but that there were large differences between the other genomospecies (Figure 5.3.3A). In particular, the three major genomospecies, GS01, GS02, and GS03 were situated in the periphery of the plot, demonstrating disparities between the gene repertoire within these genomospecies (Figure 5.3.3A).

A

GS01(47)
GS02(23)
GS03(19)
GS04(3)
GS05(6)
GS06(1)
GS07(2)
GS08(1)
GS09(1)



B



C

P value < 0.0001
**
P value < 0.001

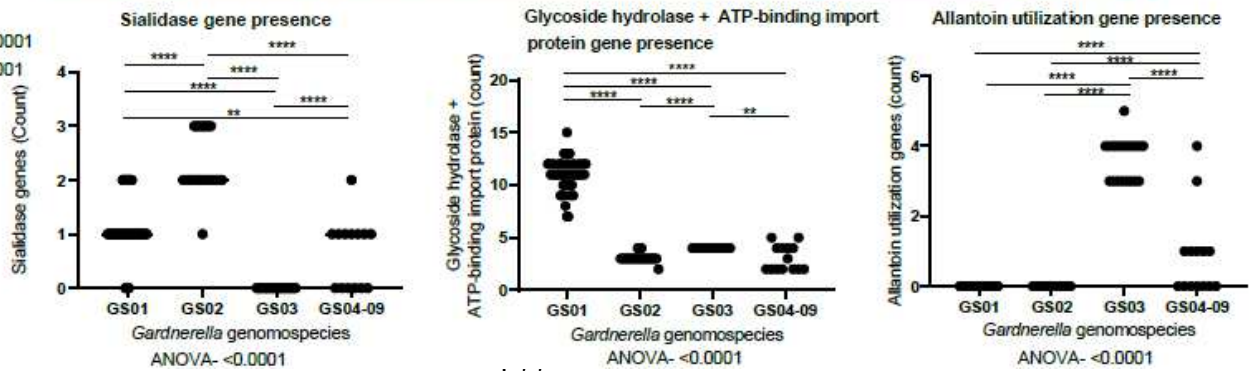


Figure 5.3.3 Accessory gene burden is different between the major genomospecies

(A) Principal component analysis of the accessory gene presence/absence matrix created by Roary with the genomes colored by their genomospecies. Each individual point represents the accessory gene content of a *Gardnerella* genome. Points closer to one another have similar accessory gene content. (B) Presence/absence matrix for accessory genes with putative sialidase, glycoside hydrolase, carbohydrate ATP-binding, and allantoin utilization roles adjacent to the phylogenetic tree from Figure 2A (distances not to scale) with genomospecies identity. Each filled square represents the presence of a given gene and a blank square represents the absence of that gene. Squares are colored based off predicted function. (C) Total counts for the number of genes identified with suspected function as sialidases, glycoside hydrolases/carbohydrate ATP-binding, and allantoin utilization abilities. Paired student T-test results are shown for all significantly different gene burdens between GS01, GS02, GS03, and GS04-09.

To gain insight into which genes may be driving this clustering pattern, we queried the pan-genome using the pan-genome association tool, Scoary, for genes that were differentially enriched within GS01, GS02, and GS03 (Figure 5.3.3B). For genes with putative function as sialidases, glycoside hydrolases, carbohydrate ATP-binding import protein, and allantoin metabolism, we viewed the presence/absence of the clusters in iTOL and quantified the gene burden in Prism. Interestingly, we found that genes annotated as sialidases were significantly ($p < 0.0001$) enriched in GS02 over GS01 and GS04-GS09 (Figure 5.3.3C). These genes were completely absent from GS03 genomes. Similarly, genes annotated as glycoside hydrolases and carbohydrate ATP-binding import proteins were significantly ($p < 0.0001$) enriched in GS01 (Figure 5.3.3C). Lastly, we found that genes involved in the uptake and usage of allantoin were enriched ($p < 0.0001$) in GS03, absent in GS01 and GS02, and sparsely present in the other genomospecies (Figure 5.3.3C). To understand overall differences in metabolic potential between the genomospecies, we submitted the pan-genome reference fasta to

EggNOG for COG annotation and quantified the number of COGs present in each genomospecies. The results were remarkably similar across all COGS except a notable increase in genes related to “Carbohydrate transport and metabolism” in GS01. Similarly, GS01 had a significantly ($p < 0.0001$) higher amount of carbohydrate utilization genes annotated by the CAZy database compared to the other genomospecies. In summary, we found that the different *Gardnerella* genomospecies could largely be distinguished by the presence/absence of their accessory genes and that accessory genes with specific functions were enriched or absent in certain genomospecies. These results indicated that different *Gardnerella* genomospecies had distinct gene repertoires, which may lead to niche separation within the vaginal environment.

5.3.4 Taxonomic signatures of novel genomospecies during BV

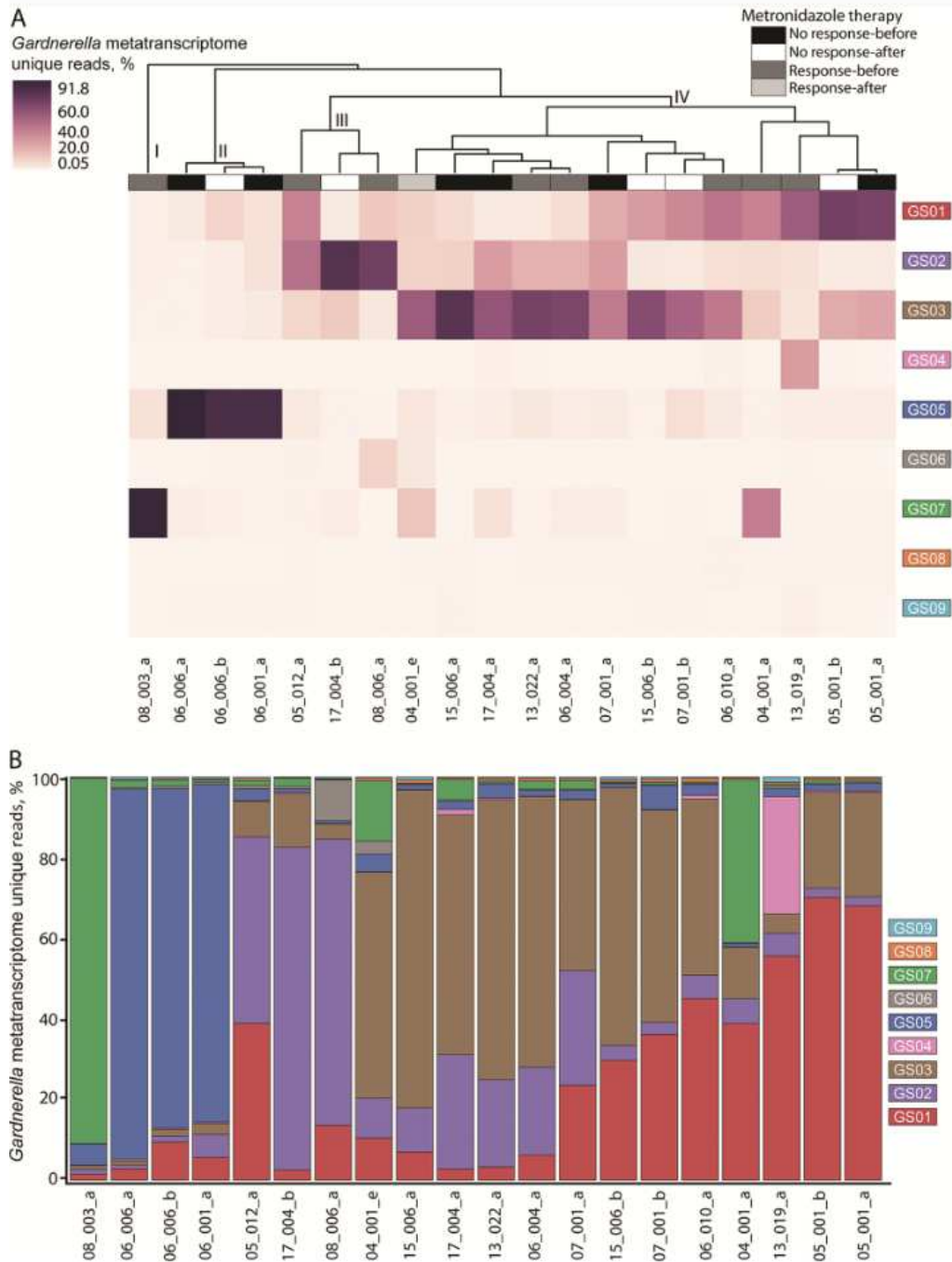


Figure 5.3.4 Newly elucidated genomospecies are identified in BV metatranscriptome samples

(A), Heatmap with hierarchal clustering from Centrifuge depicting the percentage of unique *Gardnerella* genomospecies-specific reads for each patient sample. Samples are labeled for clinical information related to efficacy of metronidazole treatment. Clusters of similar taxonomic profiles mentioned in the Results are labeled using roman numerals at the root of each cluster. (B), Stacked bar plots from Centrifuge showing the percentage of unique *Gardnerella* reads for each genomospecies within the metatranscriptome samples.

Metatranscriptomes are the genes that are expressed by a community of bacteria in any given environment. Given our improved resolution of *Gardnerella* into nine genomospecies, we wanted to investigate if any of the newly elucidated genomospecies could be identified in the metatranscriptomes of BV samples. To accomplish this, we used the short-read classifier Centrifuge, on metatranscriptome sequencing reads from BV samples of women before and after receipt of metronidazole therapy (26, 27).

We used Centrifuge to identify the percentage of *Gardnerella* reads that uniquely map to just one of the genomospecies within BV metatranscriptome samples for all of the genomospecies (26, 27). Metatranscriptome reads mapping uniquely to only a single genomospecies were identified for all genomospecies but GS04, GS06, GS08, and GS09 had a mean presence of 1.74%, 1.0%, 0.49%, and 0.41% across all samples. As our classification scheme was designed specifically to focus only on the *Gardnerella* genomospecies, in 3 of 20 samples, the largest value of unique reads was unmapped. 08_003_a had the largest number of *Gardnerella* specific unique reads, with only 15.01% unmapped. Hierarchical clustering of the genomospecies percent values showed that clustering was primarily driven by taxonomic signatures, rather than treatment outcomes to metronidazole therapy (Figure 5.3.4A). Visual interpretation of the heatmap showed 4 primary clusters of similar taxonomic profiles.

The first cluster containing only 08_003_a is largely dominated by a single genomospecies as 90.56% of the *Gardnerella* specific reads uniquely map to GS07 (Figure 5.3.4B). The second cluster contains 06_006_a, 06_006_b, and 06_001_a, and are notable for their abundance of GS05 since it composes 91.79%, 84.20%, and 83.89% of the *Gardnerella* specific reads in these samples (Figure 5.3.4B). The third

cluster contains 17_004_b, 08_006_a, and 05_012_a, and has high levels of GS02 metatranscriptome unique reads at 79.95%, 70.93%, 46.29%, respectively (Figure 5.3.4B). The fourth cluster containing 13 of 20 of the samples is notable for having the highest mean percent of GS01 (30.51%) and GS03 (46.18%). Unique reads for GS04 had the greatest prevalence within this cluster as sample 13_019_a contains 29.15% of GS04 unique reads (Figure 5.3.4B). These results indicate that the newly elucidated *Gardnerella* genomospecies can be identified as major contributors to *Gardnerella* specific metatranscriptome reads during BV.

5.3.5 Expression of translation machinery and putative virulence factors by *Gardnerella* during BV

Since metatranscriptomes provide a snapshot of the genes that are being transcribed, we finally wanted to investigate conserved features of *Gardnerella* gene expression during BV. To accomplish this, we used PanPhlAn to quantify gene coverage values for every gene in the pan-genome matrix created by Roary and the EggNOG COG annotation to identify enriched functions within highly expressed genes (one standard deviation above mean coverage levels across all samples). 194 of these 224 highly

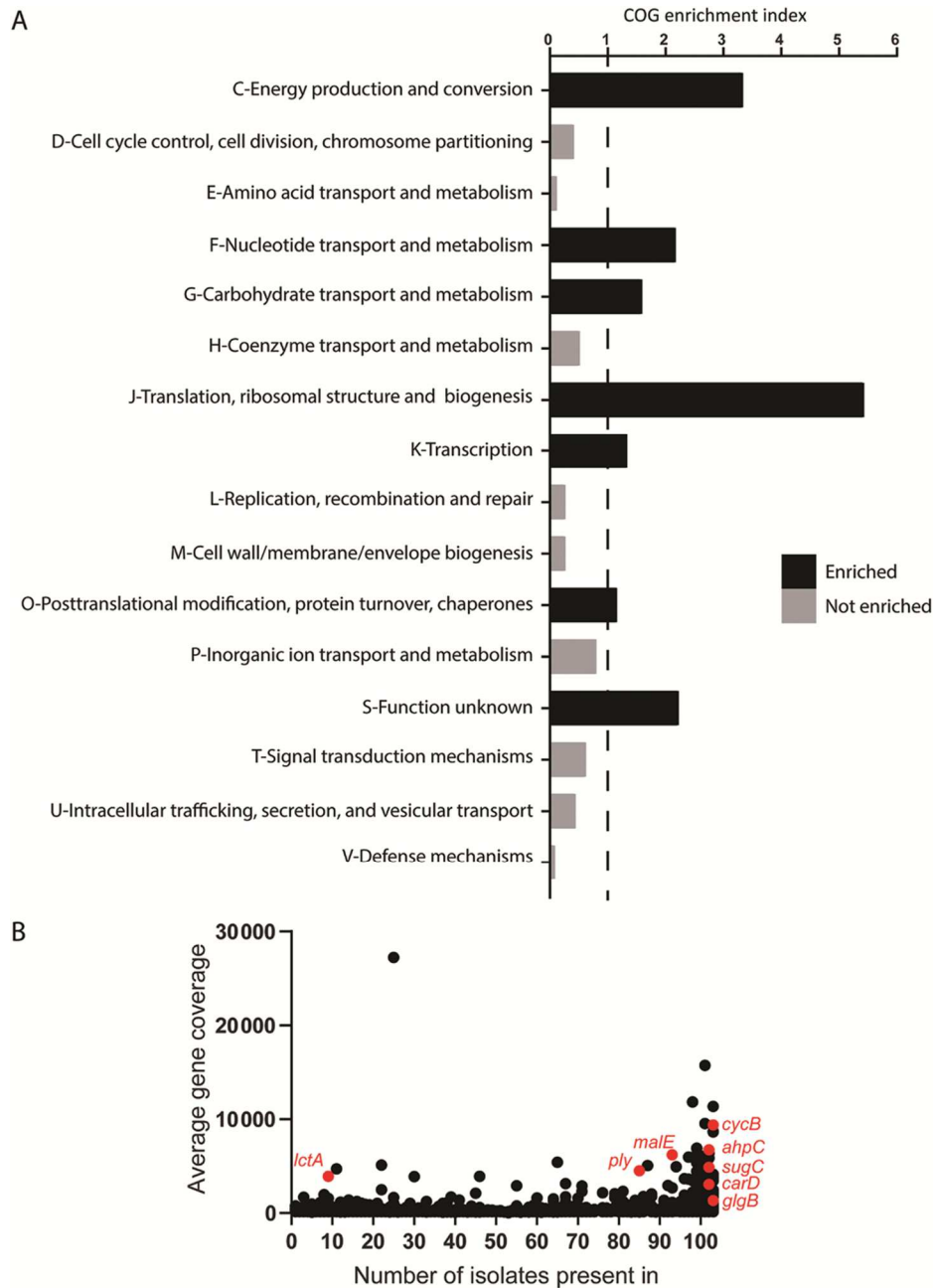


Figure 5.3.5 Gardnerella translation machinery and vaginolysin expression during BV

(A), Enrichment index scores for the COGs from EggNOG of the top 224 expressed Gardnerella pan-genome genes. Scores >1 are considered significantly enriched and are depicted in black; nonenriched COGs identified within the top 224 expressed genes are in gray. (B), Pan-genome plot depicting each of the 7402 genes as the number of isolates that harbor that gene (x value) and the PanPhlAn gene coverage value (y axis). Genes mentioned in article text are shown in red.

expressed genes had a COG annotation. 16 of 26 COGs were identified in these 194 genes (Figure 5.3.5A). Seven of 16 of these COGs had enrichment indices greater than one, indicating that they were disproportionately found amongst the highly expressed genes. The propagation and maintenance of proteins was found to be especially important, since COGs for transcription (K),

translation (J), and protein turnover (O) were all enriched within the highly expressed genes (Figure 5.3.5A). The other evident pattern was that COGs for carbohydrate transport/metabolism (G) and energy production/conversion (C) were also enriched.

Analysis of individual genes among the highly expressed group identified several candidates involved in BV pathogenesis, including the known vaginolysin toxin gene (*ply*) (Figure 5.3.5B) (31). Arguably the best studied pathogen in the Actinobacteria is *Mycobacterium tuberculosis*, and although *M. tuberculosis* and *G. vaginalis* exist in different human-associated environments, several of the highly expressed genes in BV transcriptomes are known virulence factors in the actinobacterial pathogen *M. tuberculosis*, suggesting a conserved importance in *G. vaginalis* (Figure 5.3.5B) (32-35). These include the transcriptional regulator *carD*, the trehalose import protein *sugC*, the glycan-branching enzyme *glgB*, and the oxidative stress response gene *ahpC*. *cycB*, a gene involved in maltose binding, was also found to be highly expressed, consistent with the importance of carbohydrate metabolism in the COG enrichment analysis (36). Additionally, we found that the lantibiotic lactacin gene, *lctA*, was not strongly conserved in the *Gardnerella* pan-genome, but was expressed at a high level (37). This data identifies several genetic loci which may be a conserved feature of *Gardnerella* pathogenesis in BV.

5.4 Discussion

In microbial taxonomy, phylogenomic methods are being utilized with increasing frequency to specifically and accurately delineate new bacterial species from several previously known genera, including commensal and pathogenic members of *Klebsiella* and *Propionibacterium* (22, 38). These delineations can have important implications in

understanding the biology and clinical significance of these closely-related organisms, and their potential differential contributions to health and disease in various hosts. For instance, although previously believed to be a benign environmental species, *Klebsiella variicola* strains can cause higher bladder infection titers in a mouse model of urinary tract infections compared to the canonical pathogen *Klebsiella pneumoniae* (39). Similarly, presence of gene clusters encoding ABC transporters and phosphotransferase systems were differentially present within different genera of cutaneous propionibacteria, which may enable adaptation of *Cutibacterium*, *Pseudopropionibacterium*, and *Acidipropionibacterium* to different skin niches (38). Our goal was to use initial binning of *Gardnerella* genomes into genomospecies using a variety of available tools and then explore differences in phylogeny, gene content, and metatranscriptome presence to provide insights into the biology of *Gardnerella* during BV.

Several previous reports have used whole-genome sequencing to compare pathogenic and commensal *Gardnerella* strains but did not systematically use taxonomic tools to define clear genomospecies (5, 40-42). An early analysis between strains 409-05 (GS03), ATCC 14018 (GS01), and ATCC 14019 (GS01) found that 409-05 lacked mucin degrading sialidases (40). Our analysis corroborates this finding, since we did not identify any sialidases in GS03 genomes. A broader analysis of the core-genome similarity between 17 strains found that they could be classified into 4 separate clades (42). By applying taxonomic methods to compare 103 publicly available genomes annotated as *G. vaginalis*, we found exact concordance between these initial clades and our genomospecies, as Group-1 corresponds to GS01, Group-2 corresponds to

GS02, Group-3 corresponds to GS05, and Group-4 corresponds to GS03 (42). The differentiation of *G. vaginalis* into 4 clades had been recapitulated by alignment of just the cpn60 locus (5). Again, we found complete concordance between these earlier delineations and our genomospecies as the subgroup A isolates corresponding to GS03, subgroup B were GS02, subgroup C were GS01, and subgroup D were GS05. Importantly, this latter study used pairwise ANI analysis to determine that ANI values $\leq 95\%$ were found between the cpn60 group designations, suggesting that they constituted separate genomospecies (5). One report on comparative analysis of 37 *Gardnerella* isolates identified 6 clades based off of conserved gene similarity and distinct gene presence (43). This study, however, indicated that JCP8481A/JCP8481B (GS04) and 6119V4/00703Dmash (GS05) were both in clade 3A (43). Conversely, they suggested that JCP775/ATCC 14019 and 00703C2mash/JCP8070, GS01 and GS02 respectively, were in separate clades (43).

Recently, one analysis of 81 *Gardnerella* strains found that they represented 13 genomospecies, 3 of which were elucidated to be the new species *G. piovii*, *G. leopoldii*, and *G. swidsinkii*, using a combination of *in silico* tools and phenotypic assays (6). This report did not include NR010, which was consistently annotated as a separate genomospecies in the 4 taxonomic tools that we used, making the current maximum number of *Gardnerella* genomospecies as 14. However, we show that for the type strains within *G. vaginalis*, *G. piovii*, *G. leopoldii*, *G. swidsinkii*, and representatives of the other 10 genomospecies, conflicting taxonomic information can arise from ANI vs tetranucleotide frequency vs AAI tool use (9, 29, 30). The greatest discrepancies were between the JSpeciesWS ANIb method (14 species) and the Kostas Lab AAI tool (8

genomospecies). For the purposes of our study, we took the conservative consensus for the number of genomospecies across the 4 tools, which had exact concordance with the tetranucleotide frequency analysis in JSpeciesWS. Strikingly, in 3 of 4 of the tools tested, the type strains *G. leopoldii* UGENT 06.41 (T) and *G. swidsinkii* GS 9838-1 (T) were annotated as being the same genomospecies, conflicting the phenotypic results (6). It is possible that these two strains may represent different subspecies of the same *Gardnerella* species. The approximate maximum likelihood tree and nearest neighbor network both showed that the genomospecies can be readily distinguished by the similarity of their 200 core-genes. Similarly, the genomospecies had vastly different repertoires of accessory genomes, except the group that contained GS05, GS04, and GS09. These differences may be important for adaptation to the vaginal microenvironment or BV pathology since some of the genes driving this difference include those known for virulence (e.g., sialidases) (44, 45).

Given that GS07 and GS04-GS09 were unknown in these prior genomic studies and that there was occasionally ambiguity between previous group determinations, we used metatranscriptome sequencing reads from BV samples from women before and after metronidazole, to determine if the *Gardnerella* genomospecies could be implicated in BV pathogenesis (12). Similar to the original study, our results did not find any association between the presence of specific *Gardnerella* genomospecies and resistance to metronidazole, even with the improved taxonomic classification (12). However, we detected unique transcripts to all genomospecies in every sample. When we combined the taxonomic information to identify conserved features of *Gardnerella* pathogenesis in BV, we found enrichment for COGs involved in carbohydrate transport

and conversion into energy as well as propagation of protein machinery. Importantly, the known virulence factor *ply* and several genes implicated in pathogenesis of the Actinobacteria *M. tuberculosis* were some of the genes with the highest coverage values.

The major limitation of this investigation is that as a retrospective genomic analysis, we do not have immediate access to the isolates for species characterization.

Comprehensive analysis of differences in membrane lipids and biochemical utilizations would be necessary to correctly classify these nine genomospecies into species, with proper Latin nomenclature (6). Another limitation relates to the fact that our analysis includes metatranscriptome reads rather than metagenomic reads, and thus we are unable to accurately quantify absolute abundance of the different genomospecies within the BV samples, since it is possible that a genomospecies could compose a smaller overall fraction but express a large number of genes.

5.5 Materials and Methods

5.5.1 Publicly available genomes and metatranscriptome reads

4 genomes annotated as *Gardnerella* unclassified and 99 genomes annotated as *G. vaginalis* were retrieved from National Center for Biotechnology Information genomes in October 2018. The assembly file containing chromosome and plasmid components for all genomes were used for analysis. All genomes were re-annotated for open reading frames using prokka (11). Paired-end 2 x 100 bp Illumina reads from a metatranscriptomic investigation of bacterial vaginosis (BioProject accession number PRJEB21446) were retrieved from the Sequence Read Archive in October 2018 (12).

5.5.2 *In silico* taxonomic analysis

We initially used pyANI (<https://github.com/widdowquinn/pyani>) to obtain pairwise ANI values with the mummer nucleotide alignment method on all 103 genomes obtained from the National Center for Biotechnology Information. Representative genomes for the 14 different genomospecies were uploaded to JSpeciesWS in January 2019 and annotated with default conditions for the ANIb and Tetranucleotide frequency analysis (<http://jspecies.ribohost.com/jspeciesws/#home>) (13, 14). The same 14 genomes were uploaded to the ANI matrix software from the Kostas lab (<http://enve-omics.ce.gatech.edu/g-matrix/index>). The faa file from prokka, containing protein sequences for identified open reading frames, were uploaded to the Kostas lab (average amino acid identity) AAI matrix software in January 2019 (14). For the purposes of our investigation, to create genomospecies bins for downstream analysis we adopted a conservative consensus approach. Thereby, if two or more of the tools indicated that the genomes represent the same genomospecies, we then counted them as the same genomospecies.

5.5.3 Core-genome analysis

Roary was used to cluster the open reading frames in the *Gardnerella* cohort to identify the core-genome and accessory-genome at 70% identity (15). The 200 core-genes were aligned using PRANK (16). The core-genome alignment was converted into an approximate maximum likelihood tree with FastTree and lineages were identified using BAPS within FastGear (17-19). The newick file from FastTree was viewed as a midpoint rooted tree in iTOL with bootstrap support values as branch labels (20). To construct the nearest neighbor network, the core-genome alignment file was uploaded to SplitsTrees (21).

5.5.4 Accessory genome analysis

The gene presence/absence file constructed by Roary were removed of core genes and analyzed for principal components in RStudio using prcomp (22). The elucidated genomospecies were overlaid onto the genomes. To identify genes responsible for the distinct clustering pattern observed, we used Scoary to identify genes in the Roary pan-genome that are strongly associated with the 9 different genomospecies (23). The presence/absence matrix for genes annotated as sialidases, glycoside hydrolases, ATP-binding import proteins, or allantoin utilization were viewed as a binary matrix in iToL. Counts for the number of these genes within the different genomospecies were computed and viewed in Prism V8.

5.5.5 Cluster of orthologous groups (COGs) and gene of interest quantification

We uploaded the pan-genome reference file from Roary, which contains a representative gene for the 7,402 genes in the pan-genome database to EggNOG 4.5.1 in November 2018 to identify functional categories for all possible genes (24).

Normalized COG counts for each genomospecies were determined by dividing the number of genes for each individual COG annotation by the total number of genes that had any COG assigned. To identify all the genes with putative role in carbohydrate metabolism, we uploaded the pan-genome reference file from Roary to dbCAN which uses HMMER and DIAMOND to compare our query with the CAZy database (25).

5.5.6 Taxonomic metatranscriptome analysis

To determine the presence of the *Gardnerella* genomospecies within the metatranscriptome samples, we used the short-read classification program Centrifuge (26). Initially, we made a custom database by assigning the 5,971 total contigs from the

downloaded fasta files within the *Gardnerella* cohort to a specific genomospecies using our previously described conservative consensus approach. Therefore, our database contained all open reading frames and intergenic regions. Our classification scheme was designed to ignore the other members of the vaginal microbiota, so each read could be assigned as mapping to one or more of the *Gardnerella* genomospecies, mapping uniquely to just one genomospecies, or not mapping to any of the genomospecies. For the 20 samples used in our investigation we then computed the percentage of *Gardnerella* specific reads that uniquely mapped to an individual genomospecies by quantifying the number of unique reads per genomospecies divided by the total sum of unique reads that mapped to all genomospecies. The percentage matrix created by this analysis was hierarchically clustered using SciPy and viewed as a clustermap in seaborn. Additionally, the percentage values were viewed as a stacked barplot in matplotlib.

5.5.7 Metatranscriptome functional analysis

We built a pan-genome database in PanPhlAn using the presence/absence matrix previously identified by Roary (panphlan_pangenome_generation.py) (27). We mapped the *Gardnerella* specific transcriptome reads and quantified the coverage amount of every gene in the pan-genome for each metatranscriptome sample (http://panphlan_map.py). We used the mean coverage value for each gene across the twenty metatranscriptome samples. The top 224 expressed genes, defined as the mean plus standard deviation of the coverage level, were analyzed to identify any enriched COGs. To determine if the COGs were enriched within the metatranscriptome datasets, we computed an enrichment index using the below formula (28):

(Number of top expressed genes with COG_x/Number of top expressed genes with a COG)/ (Number of total genes with COG_x/Number of total genes with a COG)

If the index was >1 then it indicated that COG was enriched within the top expressed genes. Additionally, we used a quantitative assessment of coverage values to identify genes within the *Gardnerella* pan-genome that were expressed at a statistically meaningful percentage within the dataset. To accomplish this, we viewed each individual gene in the pan-genome by plotting on the X-axis the number of isolates that harbor the gene identified by Roary and on the Y-axis the mean coverage value from PanPhIAn analysis of the 20 metatranscriptome samples.

5.5.8 Statistical Analysis

ANOVA for number of sialidase, glycoside hydrolase, ATP-binding import proteins, allantoin utilization genes, and CAZy database hits in the different genomospecies were performed in GraphPad Prism V8. Paired student T-test between select groups were performed in GraphPad Prism V8. Permutational multivariate analysis of variance (PERMANOVA) was performed on the gene_presence_absence_matrix from Roary in RStudio using the adonis2 command from the Vegan (<https://cran.r-project.org/web/packages/vegan/vegan.pdf>) package.

5.6 Acknowledgments

We thank members of the Dantas lab for insightful discussions of the results and conclusions, especially Alaric W. D'Souza for his assistance with RStudio. This work is supported in part by awards to G.D. through the National Institute of Allergy and

Infectious Diseases, and the Eunice Kennedy Shriver National Institute of Child Health & Human Development, of the National Institutes of Health under award numbers R01AI123394 and R01HD092414, respectively. RFP was supported by a National Institute of General Medical Sciences training grant through award T32 GM007067 (PI: James Skeath) and the Monsanto/Bayer Excellence Fund graduate fellowship. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies. The funding sources did not have any influence on analysis. The authors have no conflicts of interest to disclose.

5.7 References

1. Bagnall P, Rizzolo D. Bacterial vaginosis: A practical review. *JAAPA*. 2017;30(12):15-21. doi: 10.1097/01.JAA.0000526770.60197.fa. PubMed PMID: 29135564.
2. Gardner HL, Dukes CD. *Haemophilus vaginalis* vaginitis: a newly defined specific infection previously classified non-specific vaginitis. *Am J Obstet Gynecol*. 1955;69(5):962-76. PubMed PMID: 14361525.
3. Mikamo H, Sato Y, Hayasaki Y, Hua YX, Tamaya T. Vaginal microflora in healthy women with *Gardnerella vaginalis*. *J Infect Chemother*. 2000;6(3):173-7. doi: 10.1007/s101560000008. PubMed PMID: 11810560.
4. Hickey RJ, Forney LJ. *Gardnerella vaginalis* does not always cause bacterial vaginosis. *J Infect Dis*. 2014;210(10):1682-3. doi: 10.1093/infdis/jiu303. PubMed PMID: 24855684; PMCID: PMC4334793.
5. Schellenberg JJ, Paramel Jayaprakash T, Withana Gamage N, Patterson MH, Vaneechoutte M, Hill JE. *Gardnerella vaginalis* Subgroups Defined by cpn60

Sequencing and Sialidase Activity in Isolates from Canada, Belgium and Kenya. PLoS One. 2016;11(1):e0146510. doi: 10.1371/journal.pone.0146510. PubMed PMID: 26751374; PMCID: PMC4709144.

6. Vaneechoutte M, Guschin A, Van Simaey L, Gansemans Y, Van Nieuwerburgh F, Cools P. Emended description of *Gardnerella vaginalis* and description of *Gardnerella leopoldii* sp. nov., *Gardnerella piovii* sp. nov. and *Gardnerella swidsinskii* sp. nov., with delineation of 13 genomic species within the genus *Gardnerella*. Int J Syst Evol Microbiol. 2019. doi: 10.1099/ijsem.0.003200. PubMed PMID: 30648938.

7. Tindall BJ, Rossello-Mora R, Busse HJ, Ludwig W, Kämpfer P. Notes on the characterization of prokaryote strains for taxonomic purposes. Int J Syst Evol Microbiol. 2010;60(Pt 1):249-66. doi: 10.1099/ijse.0.016949-0. PubMed PMID: 19700448.

8. Medini D, Serruto D, Parkhill J, Relman DA, Donati C, Moxon R, Falkow S, Rappuoli R. Microbiology in the post-genomic era. Nat Rev Microbiol. 2008;6(6):419-30. doi: 10.1038/nrmicro1901. PubMed PMID: 18475305.

9. Ciuffo S, Kannan S, Sharma S, Badretdin A, Clark K, Turner S, Brover S, Schoch CL, Kimchi A, DiCuccio M. Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. Int J Syst Evol Microbiol. 2018;68(7):2386-92. doi: 10.1099/ijsem.0.002809. PubMed PMID: 29792589.

10. Hata H, Natori T, Mizuno T, Kanazawa I, Eldesouky I, Hayashi M, Miyata M, Fukunaga H, Ohji S, Hosoyama A, Aono E, Yamazoe A, Tsuchikane K, Fujita N, Ezaki T. Phylogenetics of family Enterobacteriaceae and proposal to reclassify *Escherichia hermannii* and *Salmonella subterranea* as *Atlantibacter hermannii* and *Atlantibacter*

- subterranea gen. nov., comb. nov. *Microbiol Immunol*. 2016;60(5):303-11. doi: 10.1111/1348-0421.12374. PubMed PMID: 26970508.
11. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30(14):2068-9. doi: 10.1093/bioinformatics/btu153. PubMed PMID: 24642063.
 12. Deng ZL, Gottschick C, Bhujji S, Masur C, Abels C, Wagner-Dobler I. Metatranscriptome Analysis of the Vaginal Microbiota Reveals Potential Mechanisms for Protection against Metronidazole in Bacterial Vaginosis. *mSphere*. 2018;3(3). doi: 10.1128/mSphereDirect.00262-18. PubMed PMID: 29875146; PMCID: PMC5990888.
 13. Richter M, Rossello-Mora R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A*. 2009;106(45):19126-31. doi: 10.1073/pnas.0906412106. PubMed PMID: 19855009; PMCID: PMC2776425.
 14. Richter M, Rossello-Mora R, Oliver Glockner F, Peplies J. JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. *Bioinformatics*. 2016;32(6):929-31. doi: 10.1093/bioinformatics/btv681. PubMed PMID: 26576653.
 15. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*. 2015;31(22):3691-3. doi: 10.1093/bioinformatics/btv421. PubMed PMID: 26198102; PMCID: PMC4817141.
 16. Loytynoja A. Phylogeny-aware alignment with PRANK. *Methods Mol Biol*. 2014;1079:155-70. doi: 10.1007/978-1-62703-646-7_10. PubMed PMID: 24170401.

17. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One*. 2010;5(3):e9490. doi: 10.1371/journal.pone.0009490. PubMed PMID: 20224823; PMCID: PMC2835736.
18. Cheng L, Connor TR, Siren J, Aanensen DM, Corander J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol Biol Evol*. 2013;30(5):1224-8. doi: 10.1093/molbev/mst028. PubMed PMID: 23408797; PMCID: PMC3670731.
19. Mostowy R, Croucher NJ, Andam CP, Corander J, Hanage WP, Marttinen P. Efficient Inference of Recent and Ancestral Recombination within Bacterial Populations. *Mol Biol Evol*. 2017;34(5):1167-82. doi: 10.1093/molbev/msx066. PubMed PMID: 28199698; PMCID: PMC5400400.
20. Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*. 2007;23(1):127-8. doi: 10.1093/bioinformatics/btl529. PubMed PMID: 17050570.
21. Huson DH. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*. 1998;14(1):68-73. PubMed PMID: 9520503.
22. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, Jenney A, Connor TR, Hsu LY, Severin J, Brisse S, Cao H, Wilksch J, Gorrie C, Schultz MB, Edwards DJ, Nguyen KV, Nguyen TV, Dao TT, Mensink M, Minh VL, Nhu NT, Schultz C, Kuntaman K, Newton PN, Moore CE, Strugnell RA, Thomson NR. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. *Proc Natl Acad Sci U S A*.

2015;112(27):E3574-81. doi: 10.1073/pnas.1501049112. PubMed PMID: 26100894; PMCID: PMC4500264.

23. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* 2016;17(1):238. doi: 10.1186/s13059-016-1108-8. PubMed PMID: 27887642; PMCID: PMC5124306.

24. Huerta-Cepas J, Szklarczyk D, Heller D, Hernandez-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, von Mering C, Bork P. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 2019;47(D1):D309-D14. doi: 10.1093/nar/gky1085. PubMed PMID: 30418610; PMCID: PMC6324079.

25. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, Busk PK, Xu Y, Yin Y. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* 2018;46(W1):W95-W101. doi: 10.1093/nar/gky418. PubMed PMID: 29771380; PMCID: PMC6031026.

26. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* 2016;26(12):1721-9. doi: 10.1101/gr.210641.116. PubMed PMID: 27852649; PMCID: PMC5131823.

27. Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT, Tett A, Morrow AL, Segata N. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat Methods.* 2016;13(5):435-8. doi: 10.1038/nmeth.3802. PubMed PMID: 26999001.

28. Murray GL, Tsyganov K, Kostoulias XP, Bulach DM, Powell D, Creek DJ, Boyce JD, Paulsen IT, Peleg AY. Global Gene Expression Profile of *Acinetobacter baumannii* During Bacteremia. *J Infect Dis.* 2017;215(suppl_1):S52-S7. doi: 10.1093/infdis/jiw529. PubMed PMID: 28375520.
29. Noble PA, Citek RW, Ogunseitan OA. Tetranucleotide frequencies in microbial genomes. *Electrophoresis.* 1998;19(4):528-35. doi: 10.1002/elps.1150190412. PubMed PMID: 9588798.
30. Konstantinidis KT, Tiedje JM. Towards a genome-based taxonomy for prokaryotes. *J Bacteriol.* 2005;187(18):6258-64. doi: 10.1128/JB.187.18.6258-6264.2005. PubMed PMID: 16159757; PMCID: PMC1236649.
31. Gelber SE, Aguilar JL, Lewis KL, Ratner AJ. Functional and phylogenetic characterization of Vaginolysin, the human-specific cytolysin from *Gardnerella vaginalis*. *J Bacteriol.* 2008;190(11):3896-903. doi: 10.1128/JB.01965-07. PubMed PMID: 18390664; PMCID: PMC2395025.
32. Weiss LA, Harrison PG, Nickels BE, Glickman MS, Campbell EA, Darst SA, Stallings CL. Interaction of CarD with RNA polymerase mediates *Mycobacterium tuberculosis* viability, rifampin resistance, and pathogenesis. *J Bacteriol.* 2012;194(20):5621-31. doi: 10.1128/JB.00879-12. PubMed PMID: 22904282; PMCID: PMC3458692.
33. Kalscheuer R, Weinrick B, Veeraraghavan U, Besra GS, Jacobs WR, Jr. Trehalose-recycling ABC transporter LpqY-SugA-SugB-SugC is essential for virulence of *Mycobacterium tuberculosis*. *Proc Natl Acad Sci U S A.* 2010;107(50):21761-6. doi: 10.1073/pnas.1014642108. PubMed PMID: 21118978; PMCID: PMC3003129.

34. Dkhar HK, Gopalsamy A, Loharch S, Kaur A, Bhutani I, Saminathan K, Bhagyaraj E, Chandra V, Swaminathan K, Agrawal P, Parkesh R, Gupta P. Discovery of Mycobacterium tuberculosis alpha-1,4-glucan branching enzyme (GlgB) inhibitors by structure- and ligand-based virtual screening. *J Biol Chem*. 2015;290(1):76-89. doi: 10.1074/jbc.M114.589200. PubMed PMID: 25384979; PMCID: PMC4281769.
35. Master SS, Springer B, Sander P, Boettger EC, Deretic V, Timmins GS. Oxidative stress response genes in Mycobacterium tuberculosis: role of *ahpC* in resistance to peroxynitrite and stage-specific survival in macrophages. *Microbiology*. 2002;148(Pt 10):3139-44. doi: 10.1099/00221287-148-10-3139. PubMed PMID: 12368447.
36. Kamionka A, Dahl MK. Bacillus subtilis contains a cyclodextrin-binding protein which is part of a putative ABC-transporter. *FEMS Microbiol Lett*. 2001;204(1):55-60. doi: 10.1111/j.1574-6968.2001.tb10862.x. PubMed PMID: 11682178.
37. Furgerson Ihnken LA, Chatterjee C, van der Donk WA. In vitro reconstitution and substrate specificity of a lantibiotic protease. *Biochemistry*. 2008;47(28):7352-63. doi: 10.1021/bi800278n. PubMed PMID: 18570436; PMCID: PMC2574596.
38. Scholz CF, Kilian M. The natural history of cutaneous propionibacteria, and reclassification of selected species within the genus Propionibacterium to the proposed novel genera Acidipropionibacterium gen. nov., Cutibacterium gen. nov. and Pseudopropionibacterium gen. nov. *Int J Syst Evol Microbiol*. 2016;66(11):4422-32. doi: 10.1099/ijsem.0.001367. PubMed PMID: 27488827.
39. Potter RF, Lainhart W, Twentyman J, Wallace MA, Wang B, Burnham CA, Rosen DA, Dantas G. Population Structure, Antibiotic Resistance, and Uropathogenicity of

Klebsiella variicola. MBio. 2018;9(6). doi: 10.1128/mBio.02481-18. PubMed PMID: 30563902; PMCID: PMC6299229.

40. Yeoman CJ, Yildirim S, Thomas SM, Durkin AS, Torralba M, Sutton G, Buhay CJ, Ding Y, Dugan-Rocha SP, Muzny DM, Qin X, Gibbs RA, Leigh SR, Stumpf R, White BA, Highlander SK, Nelson KE, Wilson BA. Comparative genomics of *Gardnerella vaginalis* strains reveals substantial differences in metabolic and virulence potential. PLoS One. 2010;5(8):e12411. doi: 10.1371/journal.pone.0012411. PubMed PMID: 20865041; PMCID: PMC2928729.

41. Harwich MD, Jr., Alves JM, Buck GA, Strauss JF, 3rd, Patterson JL, Oki AT, Girerd PH, Jefferson KK. Drawing the line between commensal and pathogenic *Gardnerella vaginalis* through genome analysis and virulence studies. BMC Genomics. 2010;11:375. doi: 10.1186/1471-2164-11-375. PubMed PMID: 20540756; PMCID: PMC2890570.

42. Ahmed A, Earl J, Retchless A, Hillier SL, Rabe LK, Cherpes TL, Powell E, Janto B, Eutsey R, Hiller NL, Boissy R, Dahlgren ME, Hall BG, Costerton JW, Post JC, Hu FZ, Ehrlich GD. Comparative genomic analyses of 17 clinical isolates of *Gardnerella vaginalis* provide evidence of multiple genetically isolated clades consistent with subspeciation into genovars. J Bacteriol. 2012;194(15):3922-37. doi: 10.1128/JB.00056-12. PubMed PMID: 22609915; PMCID: PMC3416530.

43. Cornejo OE, Hickey RJ, Suzuki H, Forney LJ. Focusing the diversity of *Gardnerella vaginalis* through the lens of ecotypes. Evol Appl. 2018;11(3):312-24. doi: 10.1111/eva.12555. PubMed PMID: 29632552; PMCID: PMC5881158.

44. Govinden G, Parker JL, Naylor KL, Frey AM, Anumba DOC, Stafford GP. Inhibition of sialidase activity and cellular invasion by the bacterial vaginosis pathogen *Gardnerella vaginalis*. *Arch Microbiol*. 2018;200(7):1129-33. doi: 10.1007/s00203-018-1520-4. PubMed PMID: 29777255; PMCID: PMC6096708.
45. Hardy L, Jespers V, Van den Bulck M, Buyze J, Mwambarangwe L, Musengamana V, Vaneechoutte M, Crucitti T. The presence of the putative *Gardnerella vaginalis* sialidase A gene in vaginal specimens is associated with bacterial vaginosis biofilm. *PLoS One*. 2017;12(2):e0172522. doi: 10.1371/journal.pone.0172522. PubMed PMID: 28241058; PMCID: PMC5328246.

Chapter 6: Phenotypic and genotypic characterization of linezolid-resistant *Enterococcus faecium* from the USA and Pakistan

6.1 Abstract

Linezolid is an important therapeutic option for the treatment of infections caused by vancomycin-resistant *Enterococcus*. Linezolid is a synthetic antimicrobial and resistance to this antimicrobial agent remains relatively rare. As a result, data on the comparative genomics of linezolid resistance determinants in *Enterococcus faecium* is relatively sparse. To address this knowledge gap in *E. faecium*, we deployed phenotypic antibiotic susceptibility testing and Illumina whole-genome on hospital surface (environmental) and clinical isolates from the United States and Pakistan. We found complete concordance between isolate source country and mechanism of linezolid resistance, with all the United States isolates possessing a 23S rRNA gene mutation and the Pakistan isolates harboring 2-3 acquired antibiotic resistance genes. These resistance genes include the recently elucidated efflux pumps *optrA* and *poxtA* and a novel *cfr*-like variant. Although there was no difference in the linezolid MIC between the United States and Pakistan isolates, there was a significant difference in the geometric mean of the MIC between the Pakistan isolates that had two versus three of the acquired antibiotic resistance genes. In five of the Pakistan *E. faecium* that possessed all three of the resistance genes, we found no difference in the local genetic context of *poxtA* and the *cfr*-like gene, but we identified different genetic contexts surrounding *optrA*. These results demonstrate that *E. faecium* from different geographical regions employ alternative strategies to counter selective pressure of increasing clinical linezolid use.

6.2 Introduction

Enterococcus faecium is a common gut commensal organism and an increasingly important cause of nosocomial infection.(1) One feature implicated in the success of *E. faecium* as a pathogen is its repertoire of acquired antibiotic resistance genes (ARGs) that enable evasion of antimicrobial therapy.(1) As an example, treatment of *E. faecium* infections with vancomycin has facilitated proliferation of the *vanA* gene cassette throughout *E. faecium*.(2) Due to the increase in vancomycin resistant Gram-positive pathogens, newer therapeutics, notably the oxazolidinones linezolid and tedizolid, have become important therapeutic agents for treating infections caused by this organism.(3) Accordingly, sporadic resistance to linezolid has been identified in cohorts of *E. faecium* and other Gram-positive bacteria.(4-6) These include vertically transmitted mutations in the linezolid target, the 23s rRNA gene sequence, and alterations in the ribosomal proteins L3, L4, and L22.(7-9) Acquired plasmid-borne antimicrobial resistance genes (ARGs), including the 23S rRNA methyltransferases *cfr* and *cfr(B)*, have been previously identified in *E. faecium*.(10-12) Newly identified efflux pump genes, *optrA* and *poxA*, have also been described in *E. faecium*.(13, 14)

Despite the identification of vertically and horizontally transferable linezolid resistance determinants, a comprehensive genomic survey of linezolid resistant *E. faecium* isolates has not been performed. Additionally, there is a gap in knowledge on the relationship of established linezolid resistance determinants and their encoded phenotypic susceptibility to the newest oxazolidinone, tedizolid. To address this, we performed whole-genome sequencing and comparative analysis on 41 newly sequenced isolates from the United States and 8 newly sequenced isolates from Pakistan. To increase the number of isolates for analysis, we supplemented these data with 52 publicly available

genomes of *E. faecium* isolated from the same locations in the US and Pakistan. Our results indicate that the mechanism of linezolid resistance is more strongly associated with geography rather than *E. faecium* clade/phylogeny in this cohort, with resistant isolates from the US harboring the G2576T SNP in 23S rRNA loci and resistant isolates from Pakistan encoding combinations of *poxtA*, *optrA*, and a *cfr*-like ARGs.

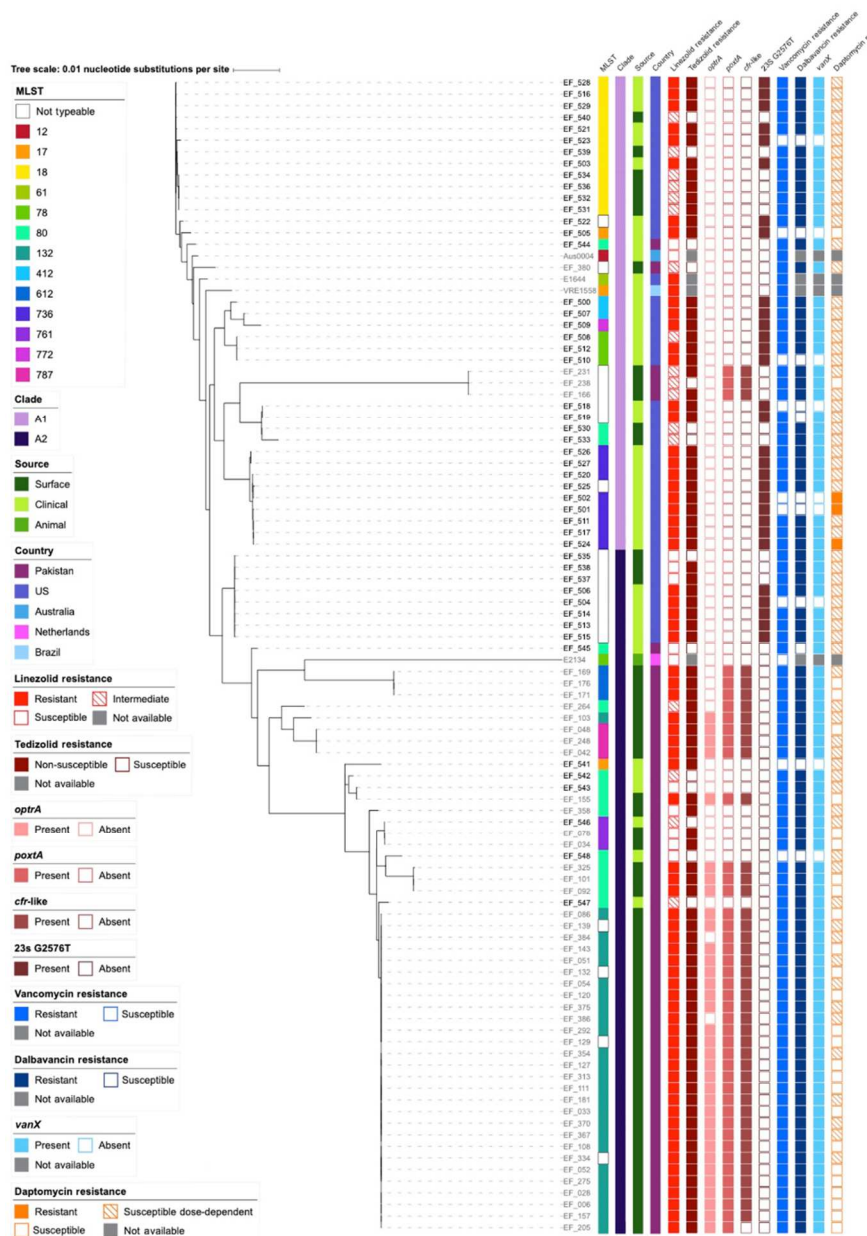
6.3 Results

6.3.1 Acquired linezolid resistance genes (*optrA*, *poxtA*, and *cfr*-like) were found exclusively in the *E. faecium* isolates recovered from Pakistan, regardless of clade.

We accessed banked environmental and clinical isolates of linezolid non-susceptible *E. faecium* isolates from the United States and Pakistan as well as several known linezolid susceptible isolates from both locations to perform a genomic analysis of linezolid resistance determinants.

Figure 6.3.1 Recombination-free phylogenetic tree including MLST, country, source, resistance, resistance gene and mutation data.

Linezolid resistance in US isolates was attributed solely to the G2576T mutation of the 23S rRNA gene sequence. In contrast, linezolid resistance in Pakistan isolates resulted from different combinations of the acquired resistance genes *optrA*, *poxtA* and a *cfr*-like gene. Vancomycin resistance was observed in 90.6% (87/96) of the isolates and dalbavancin resistance was observed in 88.5% (85/96). Daptomycin resistance was observed in 3.13% (3/96) of the isolates with an additional 68.8% (66/96) classified as susceptible dose-dependent.



We used Illumina whole-genome sequencing to construct draft-genomes for 49 isolates and obtained 52 publicly available *E. faecium* genomes isolated from the same locations in the US and

Pakistan. We used Kirby-Bauer disk diffusion and gradient diffusion methods in conjunction with CLSI interpretive guidelines to assign phenotypic resistance criteria to linezolid (resistant, intermediate, or susceptible) and tedizolid (using *Enterococcus faecalis* breakpoints for non-susceptible or susceptible). Initially, we constructed a core-genome phylogenetic tree on the 1691 core-genes between all genomes. Phylogenetic comparison of the cohort to reference isolates from *E. faecium* clades A1, A2, and B, determined that all isolates in the cohort belong to Clades A1 and A2, characteristic of human pathogens.(15) To gain further resolution on the relatedness of the *E. faecium* isolates, we excluded the Clade B isolate E1007 and constructed a recombination-free phylogenetic tree using parSNP (Figure 6.3.1). The phylogeny of the isolates was generally geographically stratified, as 80.4% (33/41) *E. faecium* from the United States were in Clade A1 and 90.9% (50/55) *E. faecium* from Pakistan were in Clade A2. The isolate cohort represented eleven identifiable multi-locus sequence types. 70.7% (29/41) of the US isolates were resistant to linezolid and of these, 100% (29/29) were positive for the G2576T 23S rRNA SNP using bowtie2 alignment of Illumina reads to the Aus0004 reference sequence (Figure 6.3.1).(16) A comparable amount of the *E. faecium* isolates from Pakistan, 72.7% (40/55), were also resistant to linezolid, however in contrast 97.5% (39/40) of these isolates were positive for an acquired linezolid resistance gene identified by ResFinder or prokka but negative for the G2576T SNP. The canonical 23S rRNA-methyltransferase, *cfr*, was not identified in our isolates, however a variant of the *cfr* family was annotated by prokka in 76.4% (42/55) *E. faecium* isolates from Pakistan (Figure 6.3.1). BLASTP query and comparison to previously characterized sequences of the *cfr* gene, the *cfr(B)* variant, and the ancestral *rlmN* gene

determined that the *cfr*-like gene shared 64% identity over 95% of query length with the original *cfr* gene and 65% identity over 97% of the length of *cfr(B)*. An identity of 74.9% over 99.7% was previously used to classify *cfr(B)* as unique from *cfr*, therefore the gene we have described fits within the category of other emerging *cfr*-like family members.(17, 18) 78.2% (43/55) and 61.8% (34/55) of the isolates from Pakistan contained the linezolid ABC transporters *poxtA* and *optrA*, respectively. 76.7% (33/43) of the isolates with gene-based resistance harbored all three of the resistance genes identified in the cohort. 20.9% (9/43) of the isolates harbored only *poxtA* and the *cfr*-like gene, and 2.32% (1/43) harbored only *optrA* and *poxtA*. 90.6% (87/96) and 88.5% (85/96) of the isolates were resistant to vancomycin and dalbavancin respectively. Only 3.12% (3/96) isolates were non-susceptible to daptomycin, another therapeutic commonly used to treat VRE in the US, however, an additional 68.8% (66/96) were tested to have minimum inhibitory concentration values in the susceptible-dose dependent classification range. These results indicate that while clade A1 and clade A2 *E. faecium* isolates can be found in both the United States and Pakistan, there is a differential burden in the mechanism of linezolid resistance between the surveyed isolates from these locations.

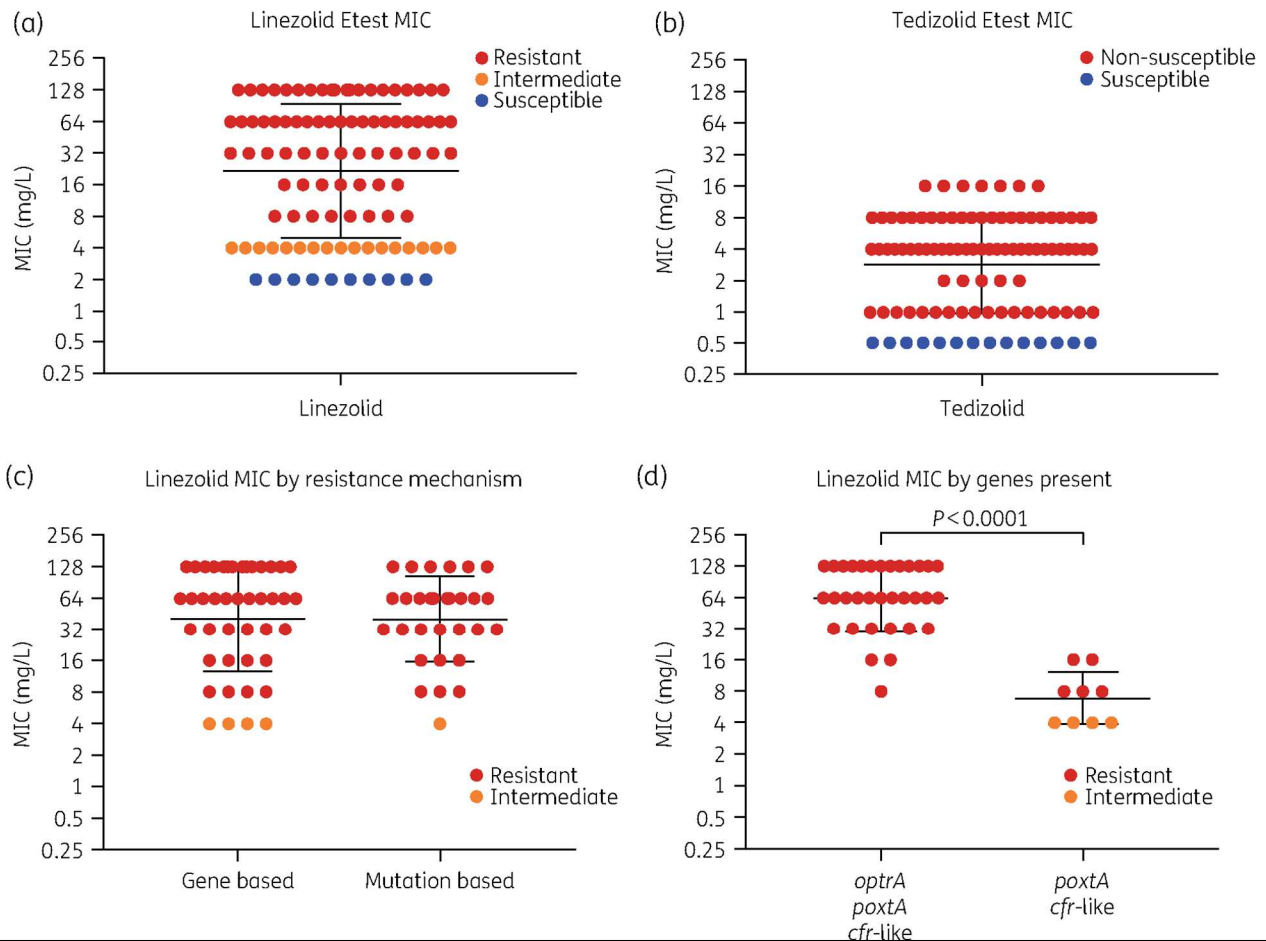


Figure 6.3.2. Linezolid and tedizolid MICs and comparisons by basis of resistance mechanism.

The geometric mean MIC of linezolid (a) is higher than the geometric mean MIC of tedizolid (b) at 21.83 and 2.87 mg/L, respectively. There was no difference in linezolid resistance between isolates with gene- or mutation-based resistance mechanisms (c). However, isolates that harboured *poxtA* and *cfr*-like genes had significantly lower levels of linezolid resistance than those that harboured all three linezolid resistance genes (d); statistical analysis was done using the unpaired t-test in Prism v8. Please note, y-axis values for all graphs are log₂ scaled for visual acuity.

6.3.2 Linezolid resistance differs by genes present, not by mechanism.

The geometric mean MIC for linezolid (21.83 mg/L) was greater than the geometric mean MIC for tedizolid (2.87 mg/L) (Figure 6.3.2ab). There was minimal difference between the geometric mean MIC of isolates with gene-based resistance (40.75 mg/L) and isolates with mutation-based resistance (40.32 mg/L) (Figure 6.3.2c). However, the geometric mean MIC of isolates with all three observed resistance genes (64 mg/L) was

significantly greater ($P < .0001$) than the geometric mean MIC of isolates that harbored only *poxtA* and the *cfr*-like gene (6.86 mg/L) (Figure 6.3.2d). Our results demonstrate that while tedizolid resistance and linezolid resistance may be related, there are several instances in our cohort where they are independent of one another. 22.9% (22/96) of the isolates were neither susceptible to both antibiotics nor resistant to linezolid and non-susceptible to tedizolid. Of these, 40.9% (9/22) of isolates had intermediate linezolid resistance but were susceptible to tedizolid, 36.4% (8/22) of isolates were linezolid intermediate and non-susceptible to tedizolid, and 22.7% (5/22) were susceptible to linezolid but non-susceptible to tedizolid. The previously identified 23S rRNA G2505A linezolid resistance mutation was not identified within the isolates from our cohort.(16, 19) However, heterogeneity at site 1232 in the aligned 23S rRNA gene of *E. faecium* Aus0004 was observed in all isolates from our cohort (with >17% frequency in 76 isolates). This site has not previously been associated with linezolid resistance and the mutation was observed in both linezolid resistant and susceptible isolates, therefore it likely does not contribute to phenotypic linezolid resistance. Within the population of *E. faecium* that contained the G2576T mutation at >17%, there was not a correlation between frequency of the G2576T SNP and phenotypic linezolid resistance.

6.3.3 Different genetic platforms of *optrA* in linezolid resistant *E. faecium* from Pakistan

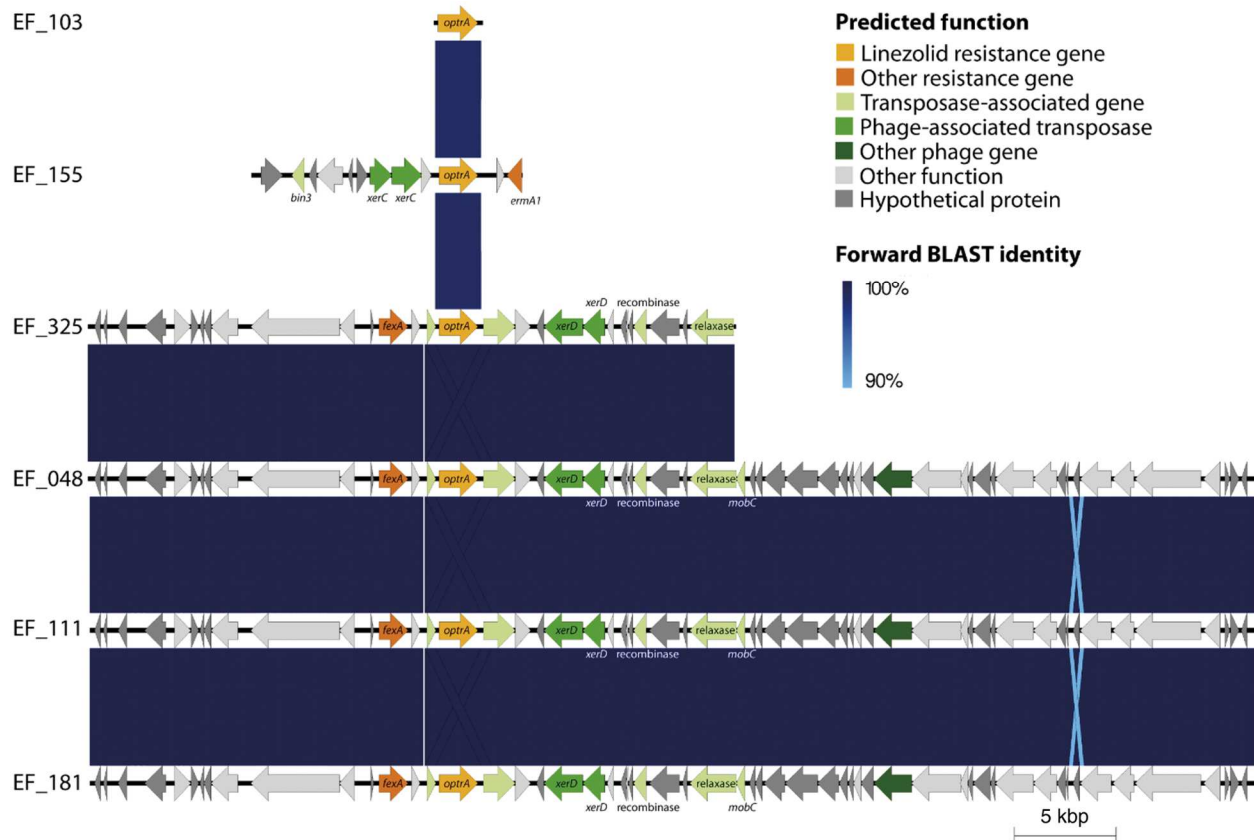


Figure 6.3.3. Genetic context of *optrA* in isolates that harbour *optrA*, *cfr*-like and *poxtA* genes.

In isolates EF_325, EF_048, EF_111 and EF_181, *optrA* is downstream of the resistance gene *fexA* and in isolate EF_155 it is upstream of an *ermA1* resistance gene. These contexts are similar to those that *optrA* was in when it was first identified. However, the mobile elements surrounding *optrA* in our isolates differ from those previously identified. *optrA*'s location near mobile elements may allow it to be transferable.

We used EasyFig to analyze the genetic context of *optrA*, *poxtA*, and the *cfr*-like gene in five isolates that harbored all three genes (Figure 6.3.3). The visualized genetic context of *optrA* was identical in Pakistan isolates EF_48, EF_111, and EF_181, as well as partially in EF_325. These segments harbored a *fexA* phenicol resistance gene adjacent to *optrA*. The context of *optrA* in EF_155 differed from the others and contained the *ermA1* methyltransferase gene. The *optrA* contigs also contained several transposase-associated and phage-associated transposase genes, which could enable horizontal transfer of the *optrA* gene. The contig from EF_103 contained only the *optrA*

gene. In all cases *poxxA* was assembled on a short contig with no other flanking genes, and the genetic context around the *cfr*-like gene was identical in the isolates we observed.

6.4 Discussion

The molecular epidemiology of linezolid resistance in VRE is largely uncharacterized, but linezolid resistance is rapidly increasing.(5) Consistent with earlier reports on the distribution of isolates in *E. faecium* clades, all of our isolates were in the A1 or A2 group.(15) Nearly 72% (69/96) of the isolates in this study were linezolid resistant, with an additional 18% (17/96) having intermediate linezolid resistance. Additionally, 85% (82/96) of the isolates were non-susceptible to tedizolid, with much lower MIC values than observed for linezolid, as has been previously observed in linezolid resistant *E. faecium* from Germany.(20) In our cohort, linezolid resistance can be attributed to a combination of resistance genes or the G2576T mutation in the 23S rRNA gene. While the resistance mechanism differs between geographic locations – with resistance in the strains recovered from Pakistan containing gene-mediated resistance determinants and US isolates harboring 23S rRNA gene mutation(s), both groups displayed similar phenotypic MIC distributions. Possibly due to differences between short-read Illumina and longer-read Sanger sequencing, we did not observe a correlation between the MIC to linezolid and the proportion of the G2576T mutation 23S rRNA allele, as has been identified previously.(21) Limiting linezolid use may partly curtail the spread of resistance, as the G2576T resistance mutation can arise in pathogens due to prolonged drug exposure and the *cfr*, *opxA*, and *poxxA* resistance genes identified have historically been capable of horizontal transfer through situation on mobile genetic elements.(14,

22-24) Tedizolid holds promise for treatment of multidrug resistant infections.(25) However, we found that 100% (69/69) of linezolid resistant isolates were also non-susceptible to tedizolid, and 47% (8/17) of linezolid intermediate isolates were tedizolid non-susceptible. Unexpectedly, 5 isolates were linezolid susceptible but tedizolid non-susceptible, although the MIC distributions for these isolates were near the resistance breakpoint for both antimicrobials. The MIC breakpoints published by the Clinical and Laboratory Standards Institute for non-susceptibility to tedizolid are lower than for linezolid based on pharmacokinetic and pharmacodynamic properties.(26) Future investigations to examine tedizolid specific resistance determinants and suitable breakpoints specifically for *E. faecium* are warranted.(25)

To the best of our knowledge, the *cfr* 23S rRNA methyltransferase family and the *optrA* and *poxtA* efflux pumps are the only known acquired ARGs against linezolid.(10, 13, 14) These genes can also confer resistance to other antibiotics, including chloramphenicol and clindamycin, complicating treatment options. *cfr*, *cfr(b)*, *cfr(c)*, and unnamed *cfr*-like genes have previously been identified in linezolid resistant strains of *Staphylococcus aureus*, *Clostridium difficile* (now *Clostridioides difficile*), *Enterococcus spp.*, *E. faecalis*, and *E. faecium*.(10, 11, 27-30) Interestingly, these genes do not appear restricted to pathogens but can be found in a diverse number of Gram-positive species, indicating that multiple opportunities for horizontal gene transfer may arise.(18) Previously, *cfr* and its variants have been identified in isolates from countries including the US, Germany, Spain, Italy, China, France, Denmark, and the United Kingdom, but to the best of our knowledge this is the first report from Pakistan. In all isolates that we observed the *cfr*-like gene, we also identified *poxtA* or both *poxtA* and *optrA*. Among isolates that only

harbored the *cfr*-like gene and *poxtA*, the geometric mean MIC (6.86 mg/L) was approximately ten times lower than that of those that harbored all three identified resistance genes (64 mg/L), with one of the two-gene isolates achieving only intermediate resistance. The genes *optrA* and *cfr* have previously been reported co-localized on plasmids in hospital borne vancomycin resistant *E. faecium*(31). Upon its discovery, there was doubt as to whether *cfr(B)* granted the same resistance phenotype in *Enterococcus* as it does in *Staphylococcus* or if the *cfr*-like gene from *C. difficile* also confers antibiotic resistance.(11): (32) Additionally, a recent study using a mouse peritonitis model found that tedizolid underperformed linezolid and daptomycin in bacterial clearance of *cfr(B)* positive *E. faecium*.(33) Treatment of *cfr(B)*-positive *E. faecium* infection with linezolid garnered 86% survival in a mouse peritonitis model, despite presenting MICs that would suggest linezolid resistance.(33) Our data, coupled with these observations, suggests that the relative contribution of the *cfr*-like gene to phenotypic resistance may be less significant than that of other resistance genes and could be attributed to significant genotypic divergence from the canonical *cfr* gene. These phenotypic discrepancies may be exacerbated by synergistic effects occurring between the *optrA* and *poxtA* transporters and the *cfr*-like methyltransferase that are not occurring when *poxtA* and the *cfr*-like gene contribute to resistance in the absence of *optrA*. Therefore, while it is possible the *cfr*-like gene, *poxtA*, and *optrA* contribute equally to linezolid resistance, further investigation is necessary to determine their individual impacts on the observed resistance phenotypes.

Notably *optrA* resided in different contexts within our isolates. Comparing the ARG genetic contexts of isolates randomly selected from different branches of the

phylogenetic tree, we found several isolates with contexts similar to those which *optrA* was originally identified in - having either the *fexA* phenicol exporter gene upstream of *optrA* or an *ermA* antibiotic resistance gene downstream of *optrA* (Figure 6.3.3).(23) However, the mobile elements identified in our isolates (several of which are phage-associated) differed from those previously observed near *optrA*. Although the limitations of short read sequencing prevented us from obtaining longer genetic contexts of the *poxtA* and *cfr*-like genes, *poxtA*, *optrA*, and *cfr* variants have previously been observed near mobilizing elements, with *the cfr* variants and *optrA* residing on plasmids.(10, 11, 14, 23)

This study aimed to characterize the molecular epidemiology and investigate the differential burden of linezolid resistance mechanisms in *E. faecium* from two geographically distinct locations. We found that all US obtained isolates have the 23S rRNA G2576T mutation while isolates from Pakistan harbor combinations of a *cfr*-like gene, *optrA*, and *poxtA*. While geometric mean MIC values for these groups did not differ greatly (40.75 mg/L for gene-based resistance and 40.32 mg/L for mutation-based resistance), there was a difference between isolates that harbored *poxtA* and *optrA* compared to those isolates that had all 3 putative ARGs. Daptomycin is the antimicrobial agent evaluated in this study with the highest rate of susceptibility based on *in vitro* testing; 3.12% (3/96) isolates in this study are phenotypically susceptible, however, 68.8% (66/96) isolates are susceptible-dose dependent to daptomycin. Of note, daptomycin therapy is not a viable option for pulmonary infections, but *Enterococcus* spp. are very uncommon causes of pneumonia.(34, 35) Additionally, in the case of isolate EF_524, therapeutic options would be extremely limited as the

isolate is resistant to linezolid, tedizolid, vancomycin, dalbavancin, daptomycin, and ampicillin – the primary antibiotics available for *Enterococcus* infection treatment. In five isolates that harbored all three ARGs, *optrA* was observed in different genetic contexts, while the *cfr*-like gene and *poxtA* were observed in similar contexts or were assembled on contigs which were too short to identify flanking genes. The major limitation of this study is that by using Illumina sequencing, we are unable to resolve plasmid versus chromosomal segments. The use of long-read sequencing may further provide context for the genetic environment surrounding *cfr*, *poxtA*, and *optrA* in the isolates from Pakistan. Nevertheless, our results indicate that *E. faecium* isolates can use distinct genetic strategies to achieve comparable *in vitro* linezolid resistance. Continued investigation of linezolid resistance in *E. faecium* and antibiotic stewardship of linezolid are advised to prevent the spread of resistance against this last-resort antibiotic.

6.5 Materials and Methods

6.5.1 Linezolid non-susceptible *E. faecium* cohort

To understand the genotypic mechanism for linezolid resistance in two different geographies, we analyzed a collection of banked linezolid intermediate and linezolid resistant *E. faecium* isolates recovered from cultures of environmental or clinical specimens between 2012-2018. Inclusion criteria include phenotypic resistance or intermediate resistance to linezolid using the ETest gradient diffusion assay (bioMerieux, Durham, NC). We accessed 44 banked linezolid non-susceptible environmental *E. faecium* and 3 linezolid susceptible isolates from 2015-2016 that were sequenced in a previous analysis (BioProject PRJNA497126) from longitudinal surveillance of hospital surfaces in Pakistan. We newly sequenced 4 linezolid non-

susceptible and 4 linezolid susceptible isolates collected from a previous analysis of clinical isolates obtained in 2012-2013 from two hospitals in Pakistan.(36) We additionally accessed 30 clinical isolates of linezolid non-susceptible *E. faecium* banked from the Barnes-Jewish Hospital clinical microbiology laboratory from 2015-2018. Finally, we accessed 8 environmental linezolid non-susceptible and 3 linezolid susceptible *E. faecium* isolates obtained from environmental surfaces in Barnes-Jewish Hospital during 2017-2018. *E. faecium* Aus0004 (Clade A1 reference), *E. faecium* E2134 (Clade A2 reference), *E. faecium* E1007 (Clade B reference) were obtained from a previous genomic analysis of *Enterococcus* evolution.(15) The linezolid resistant isolate due to 23S rRNA G2576T mutation, *E. faecium* VRE1558, and linezolid resistant isolate due to a 23S rRNA G2505A mutation, *E. faecium* E1644, were also included in phylogenetic analysis.(19, 37)

6.5.2 Illumina whole-genome sequencing and genomic analysis

Stock cultures of the *E. faecium* sequenced in this investigation were recovered from freezer vials and streaked out onto blood agar (Hardy Diagnostics). ~10 colonies were suspended into 1 mL of nuclease free water. Genomic DNA was extracted using the QIAamp BiOstic Bacteremia DNA kit (Qiagen, Germantown, MD, USA). Genomic DNA was sequenced with Illumina whole-genome sequencing, producing short read sequences. Illumina adapter sequences were removed using Trimmomatic (version 0.38), and sequence contamination was removed with DeconSeq (version 0.4.3).(38, 39) The processed reads were assembled into contigs using SPAdes (version 3.13.0).(40) Isolates sequenced in this paper, as well as previously sequences isolates (including outgroups E1007, Aus0004, and E2134 – used for clade identification, and

VRE1558 and E1644 – positive for 23S rRNA mutations G2576T and G2505A respectively), were annotated with Prokka (version 1.12).(41) MLST was also determined using BLAST similarity (<https://github.com/tseemann/mlst>). Core-genome analysis was performed with Roary (version 3.12.0) on the .gff files from prokka. The core-genome alignment with PRANK was converted to an approximate maximum-likelihood tree in FastTree (version 2.1.9). After determination that all of the isolates were from Clades A1 or A2 we removed the Clade B genome from analysis and performed parSNP (version 1.2) on the fasta files of the isolates.(42) The newick file for both trees were viewed in iTOL.(43)

6.5.3 Antibiotic susceptibility testing

Pure cultures of isolates had phenotypic antibiotic resistance determined using Kirby Bauer disk diffusion assays and gradient diffusion (i.e. Etest) assays. Both assays were performed according to the manufacturers' instructions. The results were interpreted using the CLSI M100 criteria for *Enterococcus*.(44) Linezolid (BD, Franklin Lakes, NJ, USA) and Vancomycin (Hardy Diagnostics, Santa Maria, CA) were tested using Kirby-Bauer Disks. Strains were classified linezolid susceptible at or above 23 mm, intermediate at 21-22 mm, and resistant at or below 20 mm. Isolates were classified vancomycin susceptible at or above 17 mm, intermediate at 15-16 mm, and resistant at or below 14 mm. We additionally tested linezolid (bioMerieux), daptomycin (bioMerieux), dalbavancin (Liofilchem, Waltham, MA, USA), and tedizolid (Liofilchem) using quantitative gradient diffusion assay and interpreted the MIC value in accordance with 2019 CLSI standards. Strains were classified linezolid susceptible at or below 2 mg/L, intermediate at 4 mg/L, and resistant at or above 8 mg/L. Strains were classified

daptomycin susceptible at or below 1 mg/L, susceptible dose-dependent at 2-4 mg/L, and resistant at or above 8 mg/L. Isolates were classified dalbavancin susceptible at or below 0.25 mg/L.(44) As there is currently an absence of *E. faecium* breakpoints for tedizolid, we used the *E. faecalis* breakpoint criteria for our cohort; strains were classified tedizolid susceptible at or below 0.5 mg/L and non-susceptible above 0.5 mg/L. Statistical analysis performed in Figure 2d was done using the unpaired t-test in Prism v8. All interpretation of Etest MIC values were performed with clinical accuracy and read appropriately. Reported Etest MIC values were rounded up to the nearest doubling dilution.

6.5.4 *In silico* oxazolidinone resistant determinant identification

ResFinder annotation of known resistance genes was used to identify isolates that harbored *optrA*, *poxtA*, and *VanX*.(45) We used Roary to assemble the pan-genome of the isolates and found that a *cfr*-like gene had been annotated in the genes_presence_absence output of the program.(46) The gene sequence was compared to *cfr* and variant *cfr(B)* sequences using BLAST.(11, 47)

Following published suggestions for determining linezolid resistance mutations, the reads of processed isolates were aligned using Bowtie2 to a reference 23S rRNA sequence of Aus0004.(16) The 23S rRNA sequence of Aus0004 (NCBI Reference Sequence: NR_103056.1) did not harbor any of the mutations associated with linezolid resistance. SNPs that did not match the Aus0004 reference sequence were identified using a custom python3 script. From this alignment, the site of the SNP that correlated to the G2576T mutation (using *E. coli* numbering) responsible for linezolid resistance was identified. Isolates found to be positive for the mutation by this method had the SNP

in at least 50% of reads. To identify all isolates that had the G2576T mutation at any frequency, a second script was run to extract isolates with a SNP at the respective site. All isolates having the mutation at a frequency of at least 17% of reads, which is regarded as the minimum frequency for phenotypic linezolid resistance, were considered to be resistant by ribosomal mutation.(16) Other published mutations responsible for linezolid resistance were sought out but not identified in any of the isolates; these included the G2505 23S rRNA gene mutation and mutations in the L3, L4, and L22 proteins.(19, 48, 49)

6.5.5 Data availability:

All genomes sequenced in this study have been uploaded to the NCBI WGS database associated with BioProject PRJNA517335.

6.6 Acknowledgments

We thank members of the Dantas lab for insightful discussions of the results and conclusions. The authors thank Edison Family Center for Genome Sciences & Systems Biology staff, Eric Martin, Brian Koebbe, Jessica Hoisington-Lopez, and MariaLynn Jaeger for technical support. This work was supported by a United States Agency for International Development award (award number 3220-29047) to S.A., C.A.B., and G.D. This work is supported in part by awards to G.D. through the National Institute of Allergy and Infectious Diseases and the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health (NIH) under award numbers R01AI123394, and R01HD092414, respectively. R.F.P. received support from the Monsanto Excellence Fund Graduate Fellowship. A.W.D received support from the

Institutional Program Unifying Population and Laboratory-Based Sciences Burroughs Wellcome Fund grant to Washington University. I would also like to thank Kate Wardenburg for her continued contributions.

6.7 References

1. Miller WR, Munita JM, Arias CA. Mechanisms of antibiotic resistance in enterococci. *Expert Rev Anti Infect Ther.* 2014;12(10):1221-36. doi: 10.1586/14787210.2014.956092. PubMed PMID: 25199988; PMCID: PMC4433168.
2. Freitas AR, Tedim AP, Francia MV, Jensen LB, Novais C, Peixe L, Sanchez-Valenzuela A, Sundsfjord A, Hegstad K, Werner G, Sadowy E, Hammerum AM, Garcia-Migura L, Willems RJ, Baquero F, Coque TM. Multilevel population genetic analysis of vanA and vanB *Enterococcus faecium* causing nosocomial outbreaks in 27 countries (1986-2012). *J Antimicrob Chemother.* 2016;71(12):3351-66. doi: 10.1093/jac/dkw312. PubMed PMID: 27530756.
3. Bozdogan B, Appelbaum PC. Oxazolidinones: activity, mode of action, and mechanism of resistance. *Int J Antimicrob Agents.* 2004;23(2):113-9. doi: 10.1016/j.ijantimicag.2003.11.003. PubMed PMID: 15013035.
4. Auckland C, Teare L, Cooke F, Kaufmann ME, Warner M, Jones G, Bamford K, Ayles H, Johnson AP. Linezolid-resistant enterococci: report of the first isolates in the United Kingdom. *J Antimicrob Chemother.* 2002;50(5):743-6. PubMed PMID: 12407134.
5. Bi R, Qin T, Fan W, Ma P, Gu B. The emerging problem of linezolid-resistant enterococci. *J Glob Antimicrob Resist.* 2018;13:11-9. doi: 10.1016/j.jgar.2017.10.018. PubMed PMID: 29101082.

6. Kumar S, Bandyopadhyay M, Chatterjee M, Mukhopadhyay P, Poddar S, Banerjee P. The first linezolid-resistant *Enterococcus faecium* in India: High level resistance in a patient with no previous antibiotic exposure. *Avicenna J Med*. 2014;4(1):13-6. doi: 10.4103/2231-0770.127416. PubMed PMID: 24678466; PMCID: PMC3952390.
7. Stefani S, Bongiorno D, Mongelli G, Campanile F. Linezolid Resistance in *Staphylococci*. *Pharmaceuticals (Basel)*. 2010;3(7):1988-2006. doi: 10.3390/ph3071988. PubMed PMID: 27713338; PMCID: PMC4036669.
8. Ikonomidis A, Grapsa A, Pavlioglou C, Demiri A, Batarli A, Panopoulou M. Accumulation of multiple mutations in linezolid-resistant *Staphylococcus epidermidis* causing bloodstream infections; in silico analysis of L3 amino acid substitutions that might confer high-level linezolid resistance. *J Chemother*. 2016;28(6):465-8. doi: 10.1080/1120009X.2015.1119373. PubMed PMID: 27077930.
9. Dong W, Chochua S, McGee L, Jackson D, Klugman KP, Vidal JE. Mutations within the *rplD* Gene of Linezolid-Nonsusceptible *Streptococcus pneumoniae* Strains Isolated in the United States. *Antimicrob Agents Chemother*. 2014;58(4):2459-62. doi: 10.1128/AAC.02630-13. PubMed PMID: 24492357; PMCID: PMC4023712.
10. Morales G, Picazo JJ, Baos E, Candel FJ, Arribi A, Pelaez B, Andrade R, de la Torre MA, Fereres J, Sanchez-Garcia M. Resistance to linezolid is mediated by the *cfr* gene in the first report of an outbreak of linezolid-resistant *Staphylococcus aureus*. *Clin Infect Dis*. 2010;50(6):821-5. doi: 10.1086/650574. PubMed PMID: 20144045.
11. Deshpande LM, Ashcraft DS, Kahn HP, Pankey G, Jones RN, Farrell DJ, Mendes RE. Detection of a New *cfr*-Like Gene, *cfr(B)*, in *Enterococcus faecium* Isolates

- Recovered from Human Specimens in the United States as Part of the SENTRY Antimicrobial Surveillance Program. *Antimicrob Agents Chemother.* 2015;59(10):6256-61. doi: 10.1128/AAC.01473-15. PubMed PMID: 26248384; PMCID: PMC4576063.
12. Doern CD, Park JY, Gallegos M, Alspaugh D, Burnham CA. Investigation of Linezolid Resistance in Staphylococci and Enterococci. *J Clin Microbiol.* 2016;54(5):1289-94. doi: 10.1128/JCM.01929-15. PubMed PMID: 26935728; PMCID: PMC4844726.
13. Wang Y, Lv Y, Cai J, Schwarz S, Cui L, Hu Z, Zhang R, Li J, Zhao Q, He T, Wang D, Wang Z, Shen Y, Li Y, Fessler AT, Wu C, Yu H, Deng X, Xia X, Shen J. A novel gene, *optrA*, that confers transferable resistance to oxazolidinones and phenicols and its presence in *Enterococcus faecalis* and *Enterococcus faecium* of human and animal origin. *J Antimicrob Chemother.* 2015;70(8):2182-90. doi: 10.1093/jac/dkv116. PubMed PMID: 25977397.
14. Antonelli A, D'Andrea MM, Brenciani A, Galeotti CL, Morroni G, Pollini S, Varaldo PE, Rossolini GM. Characterization of *poxTA*, a novel phenicol-oxazolidinone-tetracycline resistance gene from an MRSA of clinical origin. *J Antimicrob Chemother.* 2018;73(7):1763-9. doi: 10.1093/jac/dky088. PubMed PMID: 29635422.
15. Lebreton F, Manson AL, Saavedra JT, Straub TJ, Earl AM, Gilmore MS. Tracing the Enterococci from Paleozoic Origins to the Hospital. *Cell.* 2017;169(5):849-61 e13. doi: 10.1016/j.cell.2017.04.027. PubMed PMID: 28502769; PMCID: PMC5499534.
16. Beukers AG, Hasman H, Hegstad K, van Hal SJ. Recommendations To Address the Difficulties Encountered When Determining Linezolid Resistance from Whole-

- Genome Sequencing Data. *Antimicrob Agents Chemother.* 2018;62(8). doi: 10.1128/AAC.00613-18. PubMed PMID: 29844046; PMCID: PMC6105777.
17. Hansen LH, Vester B. A cfr-like gene from *Clostridium difficile* confers multiple antibiotic resistance by the same mechanism as the cfr gene. *Antimicrob Agents Chemother.* 2015;59(9):5841-3. doi: 10.1128/AAC.01274-15. PubMed PMID: 26149991; PMCID: PMC4538495.
18. Vester B. The cfr and cfr-like multiple resistance genes. *Res Microbiol.* 2018;169(2):61-6. doi: 10.1016/j.resmic.2017.12.003. PubMed PMID: 29378339.
19. Prystowsky J, Siddiqui F, Chosay J, Shinabarger DL, Millichap J, Peterson LR, Noskin GA. Resistance to linezolid: characterization of mutations in rRNA and comparison of their occurrences in vancomycin-resistant enterococci. *Antimicrob Agents Chemother.* 2001;45(7):2154-6. doi: 10.1128/AAC.45.7.2154-2156.2001. PubMed PMID: 11408243; PMCID: PMC90620.
20. Klupp EM, Both A, Belmar Campos C, Buttner H, Konig C, Christopeit M, Christner M, Aepfelbacher M, Rohde H. Tedizolid susceptibility in linezolid- and vancomycin-resistant *Enterococcus faecium* isolates. *Eur J Clin Microbiol Infect Dis.* 2016;35(12):1957-61. doi: 10.1007/s10096-016-2747-0. PubMed PMID: 27525679.
21. Chacko KI, Sullivan MJ, Beckford C, Altman DR, Ciferri B, Pak TR, Sebra R, Kasarskis A, Hamula CL, van Bakel H. Genetic Basis of Emerging Vancomycin, Linezolid, and Daptomycin Heteroresistance in a Case of Persistent *Enterococcus faecium* Bacteremia. *Antimicrob Agents Chemother.* 2018;62(4). doi: 10.1128/AAC.02007-17. PubMed PMID: 29339387; PMCID: PMC5913925.

22. Bourgeois-Nicolaos N, Massias L, Couson B, Butel MJ, Andremont A, Doucet-Populaire F. Dose dependence of emergence of resistance to linezolid in *Enterococcus faecalis* in vivo. *J Infect Dis.* 2007;195(10):1480-8. doi: 10.1086/513876. PubMed PMID: 17436228.
23. He T, Shen Y, Schwarz S, Cai J, Lv Y, Li J, Fessler AT, Zhang R, Wu C, Shen J, Wang Y. Genetic environment of the transferable oxazolidinone/phenicol resistance gene *optrA* in *Enterococcus faecalis* isolates of human and animal origin. *J Antimicrob Chemother.* 2016;71(6):1466-73. doi: 10.1093/jac/dkw016. PubMed PMID: 26903276.
24. Toh SM, Xiong L, Arias CA, Villegas MV, Lolans K, Quinn J, Mankin AS. Acquisition of a natural resistance gene renders a clinical strain of methicillin-resistant *Staphylococcus aureus* resistant to the synthetic antibiotic linezolid. *Mol Microbiol.* 2007;64(6):1506-14. doi: 10.1111/j.1365-2958.2007.05744.x. PubMed PMID: 17555436; PMCID: PMC2711439.
25. Zhanel GG, Love R, Adam H, Golden A, Zelenitsky S, Schweizer F, Gorityala B, Lagace-Wiens PR, Rubinstein E, Walkty A, Gin AS, Gilmour M, Hoban DJ, Lynch JP, 3rd, Karlowsky JA. Tedizolid: a novel oxazolidinone with potent activity against multidrug-resistant gram-positive pathogens. *Drugs.* 2015;75(3):253-70. doi: 10.1007/s40265-015-0352-7. PubMed PMID: 25673021.
26. Bensaci M, Flanagan S, Sandison T. Determination of Tedizolid susceptibility interpretive criteria for gram-positive pathogens according to clinical and laboratory standards institute guidelines. *Diagn Microbiol Infect Dis.* 2018;90(3):214-20. doi: 10.1016/j.diagmicrobio.2017.10.023. PubMed PMID: 29277464.

27. Diaz L, Kiratisin P, Mendes RE, Panesso D, Singh KV, Arias CA. Transferable plasmid-mediated resistance to linezolid due to cfr in a human clinical isolate of *Enterococcus faecalis*. *Antimicrob Agents Chemother*. 2012;56(7):3917-22. doi: 10.1128/AAC.00419-12. PubMed PMID: 22491691; PMCID: PMC3393385.
28. Inkster T, Coia J, Meunier D, Doumith M, Martin K, Pike R, Imrie L, Kane H, Hay M, Wiuff C, Wilson J, Deighan C, Hopkins KL, Woodford N, Hill R. First outbreak of colonization by linezolid- and glycopeptide-resistant *Enterococcus faecium* harbouring the cfr gene in a UK nephrology unit. *J Hosp Infect*. 2017;97(4):397-402. doi: 10.1016/j.jhin.2017.07.003. PubMed PMID: 28698020.
29. Bender JK, Fleige C, Klare I, Fiedler S, Mischnik A, Mutters NT, Dingle KE, Werner G. Detection of a cfr(B) Variant in German *Enterococcus faecium* Clinical Isolates and the Impact on Linezolid Resistance in *Enterococcus* spp. *PLoS One*. 2016;11(11):e0167042. doi: 10.1371/journal.pone.0167042. PubMed PMID: 27893790; PMCID: PMC5125667.
30. Candela T, Marvaud JC, Nguyen TK, Lambert T. A cfr-like gene cfr(C) conferring linezolid resistance is common in *Clostridium difficile*. *Int J Antimicrob Agents*. 2017;50(3):496-500. doi: 10.1016/j.ijantimicag.2017.03.013. PubMed PMID: 28663118.
31. Lazaris A, Coleman DC, Kearns AM, Pichon B, Kinnevey PM, Earls MR, Boyle B, O'Connell B, Brennan GI, Shore AC. Novel multiresistance cfr plasmids in linezolid-resistant methicillin-resistant *Staphylococcus epidermidis* and vancomycin-resistant *Enterococcus faecium* (VRE) from a hospital outbreak: co-location of cfr and optrA in VRE. *J Antimicrob Chemother*. 2017;72(12):3252-7. doi: 10.1093/jac/dkx292. PubMed PMID: 28961986.

32. Schwarz S, Wang Y. Nomenclature and functionality of the so-called cfr gene from *Clostridium difficile*. *Antimicrob Agents Chemother*. 2015;59(4):2476-7. doi: 10.1128/AAC.04893-14. PubMed PMID: 25762794; PMCID: PMC4356762.
33. Singh KV, Arias CA, Murray BE. Efficacy of Tedizolid Against Enterococci and Staphylococci, including cfr+ strains, in a Mouse Peritonitis Model. *Antimicrob Agents Chemother*. 2019. doi: 10.1128/AAC.02627-18. PubMed PMID: 30670435.
34. Savini V, Gherardi G, Astolfi D, Polilli E, Dicuonzo G, D'Amario C, Fazii P, D'Antonio D. Insights into airway infections by enterococci: a review. *Recent Pat Antiinfect Drug Discov*. 2012;7(1):36-44. PubMed PMID: 22044357.
35. Silverman JA, Mortin LI, Vanpraagh AD, Li T, Alder J. Inhibition of daptomycin by pulmonary surfactant: in vitro modeling and clinical impact. *J Infect Dis*. 2005;191(12):2149-52. doi: 10.1086/430352. PubMed PMID: 15898002.
36. Pesesky MW, Hussain T, Wallace M, Wang B, Andleeb S, Burnham CA, Dantas G. KPC and NDM-1 genes in related Enterobacteriaceae strains and plasmids from Pakistan and the United States. *Emerg Infect Dis*. 2015;21(6):1034-7. doi: 10.3201/eid2106.141504. PubMed PMID: 25988236; PMCID: PMC4451916.
37. do Prado GVB, Marchi AP, Moreno LZ, Rizek C, Amigo U, Moreno AM, Rossi F, Guimaraes T, Levin AS, Costa SF. Virulence and resistance pattern of a novel sequence type of linezolid-resistant *Enterococcus faecium* identified by whole-genome sequencing. *J Glob Antimicrob Resist*. 2016;6:27-31. doi: 10.1016/j.jgar.2016.02.002. PubMed PMID: 27530835.

38. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-20. doi: 10.1093/bioinformatics/btu170. PubMed PMID: 24695404; PMCID: PMC4103590.
39. Schmieder R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS One*. 2011;6(3):e17288. doi: 10.1371/journal.pone.0017288. PubMed PMID: 21408061; PMCID: PMC3052304.
40. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012;19(5):455-77. doi: 10.1089/cmb.2012.0021. PubMed PMID: 22506599; PMCID: PMC3342519.
41. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30(14):2068-9. doi: 10.1093/bioinformatics/btu153. PubMed PMID: 24642063.
42. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol*. 2014;15(11):524. doi: 10.1186/PREACCEPT-2573980311437212. PubMed PMID: 25410596; PMCID: PMC4262987.
43. Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*. 2007;23(1):127-8. doi: 10.1093/bioinformatics/btl529. PubMed PMID: 17050570.
44. CLSI. Performance standards for antimicrobial susceptibility testing: Twenty-third Informational Supplement M100-S23: Clinical and Laboratory Standards Institute; 2013.

45. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother.* 2012;67(11):2640-4. doi: 10.1093/jac/dks261. PubMed PMID: 22782487; PMCID: PMC3468078.
46. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics.* 2015;31(22):3691-3. doi: 10.1093/bioinformatics/btv421. PubMed PMID: 26198102; PMCID: PMC4817141.
47. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403-10. doi: 10.1016/S0022-2836(05)80360-2. PubMed PMID: 2231712.
48. Mendes RE, Deshpande LM, Farrell DJ, Spanu T, Fadda G, Jones RN. Assessment of linezolid resistance mechanisms among *Staphylococcus epidermidis* causing bacteraemia in Rome, Italy. *J Antimicrob Chemother.* 2010;65(11):2329-35. doi: 10.1093/jac/dkq331. PubMed PMID: 20841419.
49. Roman F, Roldan C, Trincado P, Ballesteros C, Carazo C, Vindel A. Detection of linezolid-resistant *Staphylococcus aureus* with 23S rRNA and novel L4 riboprotein mutations in a cystic fibrosis patient in Spain. *Antimicrob Agents Chemother.* 2013;57(5):2428-9. doi: 10.1128/AAC.00208-13. PubMed PMID: 23459489; PMCID: PMC3632889.

Chapter 7: Pleiotropic effects of *pgsA2* mediated daptomycin resistance in *Corynebacterium*

7.1 Abstract

Daptomycin is an important drug of last resort in the fight against Gram-positive multidrug resistant bacteria. *Corynebacterium* are a genus of bacteria commonly found as skin commensals but also can be human pathogens. It has been found that *Corynebacterium striatum* can gain high level daptomycin resistance through null mutations in the phosphatidylglycerol synthase gene *pgsa2* leading to a complete reshuffling of the cell membrane with the notable absence of phosphatidylglycerol. However, the extent that this phenotype is able to occur across other *Corynebacterium* species and the effects that this development has on cellular physiology has not been explored. To address this gap in knowledge we curated a cohort of *Corynebacterium* isolates and phenotypically assayed their ability to develop daptomycin resistance before performing Illumina whole-genome sequencing. We then determined that a number of resistant-susceptible pairs had predicted alterations in *pgsa2* and that this occurred in one of the better studied species (*Corynebacterium striatum*) as well as rarer causes of infection. We also found that some pairs did not have these mutations, positing that multiple routes to daptomycin resistance may occur for *Corynebacterium*. Among phenotypic differences between daptomycin resistant and susceptible isolates, we found a number of compounds that had great antibacterial efficacy against daptomycin resistant *C. striatum* compared to daptomycin susceptible isolates. We also found changes in the cellular proteome between the resistant and susceptible isolate, with notable downregulation of nitrate reductase and a necessary cofactor. Consistent

with this finding, we determined that the daptomycin resistant isolates have impaired anaerobic growth relative to the susceptible. This work provides important information on the extent of daptomycin resistance across pathogenic *Corynebacterium* and on the effects that development has on changing bacterial physiology.

7.1 Introduction

Corynebacterium are a diverse genus of Gram-positive bacteria that include the industrial source of monosodium glutamate (*Corynebacterium glutamicum*) as well as human pathogens (including the causative agent of diphtheria *Corynebacterium diphtheriae*) but are predominantly found as skin commensals(1). One emerging pathogenic *Corynebacterium* is *Corynebacterium striatum*(2, 3). Prior clinical dogma has taught that *C. striatum* is purely a skin commensal and if found in isolation during cultures is a likely contaminant, however several studies have determined that *C. striatum* can be attributed as the causative infectious agent for different maladies (2, 3). Problematically, *C. striatum* are often multidrug resistant due to a high burden of acquired antibiotic resistance genes(2, 3). Therefore, drugs of last resort such as daptomycin are often the favored treatment.

Daptomycin is a lipopeptide antibiotic that is effective against growing and inert Gram-positive bacteria(4). Upon binding to the cell membrane, the lipid tail intercalates within the outer leaflet and multiple daptomycin molecules oligomerize to produce a pore which causes loss of membrane integrity and a non-lytic form of bacteria death(4). However, there have been many case reports detailing the overnight development of high level daptomycin resistance in *C. striatum*(5-7). Mechanistic work has determined that this phenotype is due to loss of function mutations in phosphatidylglycerol synthase

enzyme (*pgsA2*) in the resistant compared to wildtype isolates(5-7). Mutations in this gene are believed to be responsible for a complete loss of phosphatidylglycerol (PG) from the cell membrane(5-7). In order to maintain barrier integrity, this loss of PG occurs with a commensurate increase in the levels of phosphatidylinositol (PI) and glucuronosyl diacylglycerol(5-7)l.

While this phenomenon has been studied in *C. striatum*, there is a gap in knowledge on the extent of non-striatum *Corynebacterium* to develop high level daptomycin resistance. Additionally, there is a gap in knowledge on the effects that this massive rearrangement of lipid metabolism has on other aspects of bacterial physiology. To address this gap in knowledge we assembled a cohort of clinical *Corynebacterium* isolates from a variety of species and assed if they were capable of developing overnight daptomycin resistance and then performing Illumina whole-genome sequencing. We additionally performed proteomics and follow up experiments on an isolate of *C. striatum* and two isogenic resistant pairs.

7.3 Results

7.3.1 *In silico* species identification

We initially wanted to investigate if there was a phylogenetic signal for predictive ability of daptomycin resistant development and so we performed average nucleotide identity analysis on the assembled cohort of sequenced bacteria, type genomes from NCBI, and known *C. striatum* genomes. Initial MALDI-TOF MS analysis indicated that our cohort contained 23 different species, however ANI analysis was able to determine that only 16 of those had ANI >96% with type strains. Additionally, we found that the remaining isolates represented 19 novel genomospecies. 4/16 of the recognized species and 7/19 of the genomospecies had representative isolates that could develop high level

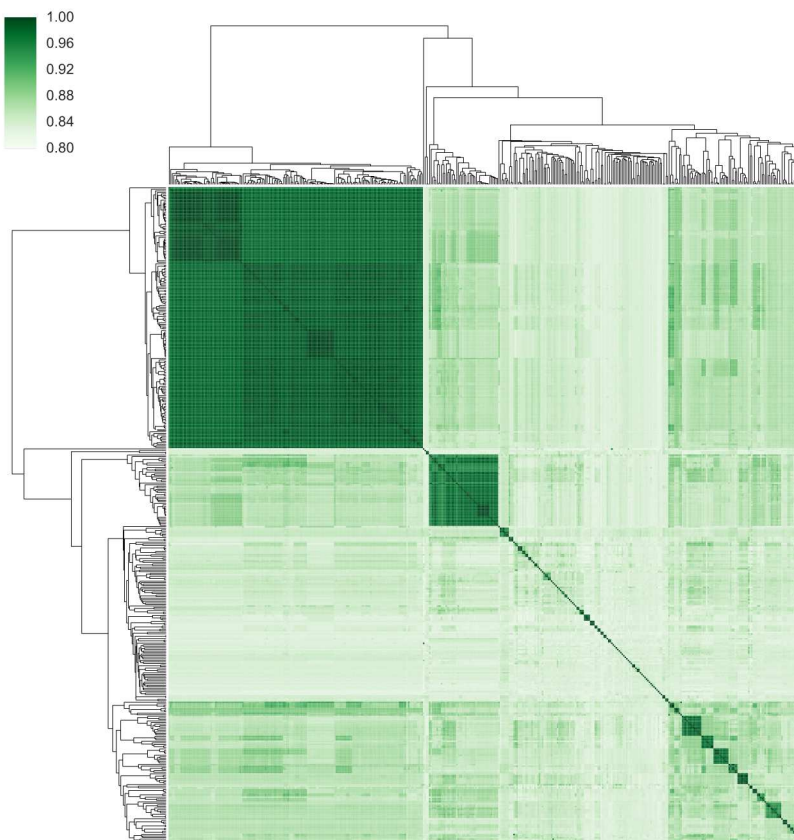


Figure 7.3.1 ANI heatmap for entire cohort. Dark green blocks indicate isolates of the same species, the large number of singletons highlight the diversity of *Corynebacterium*. Upper right corner depicts the ANI value scale.

daptomycin resistance.

7.3.2 Resistant mutation mapping

We used bowtie2 to align the reads from the resistant pairs to their respective susceptible genomes to identify genes that have SNPs. Similarly, we analyzed the protein amino acid sequence of the PgsA2 sequence in the resistant isolates to identify amino acid changes or

Species	Isolate A	Isolate B	Bowtie % Alignment	PgsA2 Mutation (Prokka)	<i>pgsA2</i> Mutation (VCF)
GS10	Cor_100	Cor_185	99.77%	none	none
GS11	Cor_086	Cor_116	99.70%	Truncation	none
GS12	Cor_127	Cor_128	99.72%	Truncation	1
GS2	Cor_183	Cor_134	99.76%	none	none
GS4	Cor_026	Cor_071	99.77%	none	none
GS8	Cor_096	Cor_120	99.81%	none	2
GS9	Cor_197	Cor_178	99.69%	none	none
<i>macginleyi</i>	Cor_002	Cor_067	99.52%	Truncation	2
<i>macginleyi</i>	Cor_087	Cor_117	99.67%	none	none
<i>simulans</i>	Cor_130	Cor_131	99.81%	Truncation	2
<i>striatum</i>	Cor_003	Cor_056	99.42%	Amino Acid Change	2
<i>striatum</i>	Cor_044	Cor_069	99.68%	Amino Acid Change	1
<i>striatum</i>	Cor_064	Cor_065	99.62%	Amino Acid Change	1
<i>striatum</i>	Cor_005	Cor_057	99.48%	Truncation	2
<i>striatum</i>	Cor_006	Cor_058	99.56%	Truncation	2
<i>striatum</i>	Cor_008	Cor_059	99.61%	Truncation	2
<i>striatum</i>	Cor_010	Cor_060	99.58%	Truncation	none
<i>striatum</i>	Cor_046	Cor_068	99.65%	Truncation	1

<i>striatum</i>	Cor_064	Cor_06	99.63%	Truncation	none
<i>striatum</i>	Cor_101	Cor_12	99.57%	Truncation	2
<i>striatum</i>	Cor_115	Cor_12	99.62%	Truncation	2
<i>striatum</i>	Cor_146	Cor_16	99.50%	Truncation	none
<i>striatum</i>	Cor_153	Cor_17	98.20%	Truncation	none
<i>striatum</i>	Cor_191	Cor_17	99.64%	Truncation	none
<i>striatum</i>	Cor_208	Cor_17	99.69%	Truncation	2
<i>striatum</i>	Cor_012	Cor_07	99.55%	none	4
<i>striatum</i>	Cor_099	Cor_12	99.53%	none	3
<i>striatum</i>	Cor_145	Cor_16	99.55%	none	2
<i>striatum</i>	Cor_148	Cor_16	99.59%	none	3
<i>striatum</i>	Cor_152	Cor_17	99.59%	none	2
<i>striatum</i>	Cor_141	Cor_16	99.60%	none	none
<i>ulcerans</i>	Cor_195	Cor_175	99.70%	Truncation	none

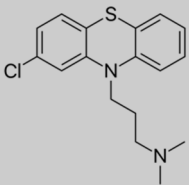
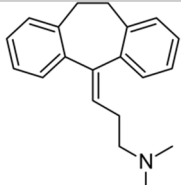
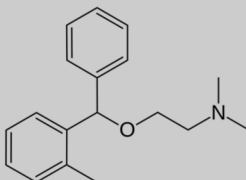
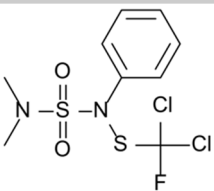
Table 7.3. 1 SNP analysis of susceptible-resistant pairs. Table courtesy of Kate Wardenburg.

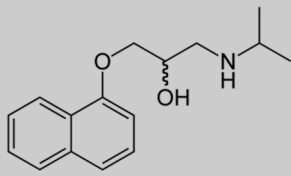
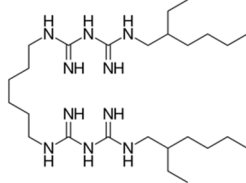
truncations. 13/31 had identified amino acid changes and polymorphisms in the nucleotide sequence, 5/31 has SNPs in *pgsa2* but no amino acid changes, and 7/31 had changes in amino

acid sequence but not SNPs. At present we are unable to explain these discrepancies between

nucleotide and amino acid sequence but it could possible be to development of heteroresistant populations in our sample being unable to be resolved by de-novo assembly. 6/31 did not have any amino acid or nucleotide changes in *pgsa2* which could indicate that *Corynebacterium* may have non *pgsa2* mediated mechanisms of resistance.

7.3.3 BiOLOG Chemical Sensitivity Screen

Compound	Normal Function	Compound Structure
Chlorpromazine	Anti-psychotic	
Amitriptyline	Anti-depressant	
Orphenadrine	Out of use for treatment of muscle pain and muscle control in Parkinson's patients	
Dichlofluandid	Fungicide often used on fruit	

D,L-Propranolol	Beta blocker to normalize cardiac rhythms		BiOLOG chemical sensitivity plates identified 6
Alexidine	Antiseptic used in mouthwash		compounds that had drastically increased

susceptibility against PR and IR compared to PS, indicating that they may be agents which damage the newly daptomycin resistant membrane. Interestingly 3/6 of these

Table 7.3.2 Structure and description of top BiOLOG hits that had differential activity against daptomycin resistant *C. striatum* compared to susceptible. Table courtesy of Kate Wardenburg.

compounds were previously FDA approved drugs

for neuropsychiatric disorders. 2/6 of the compounds had known antimicrobial activity but dichlofluanid is a fungicide rather than an antibiotic. There is structural similarity between 5/6 of the compounds as all but alexidine contain bulky aromatic benzene rings. Alexidine contains long hydrocarbon stretches at it ends which could intercalate within a membrane.

7.3.4 Proteomic identification of impaired nitrate reductase levels and anerobic growth assessment

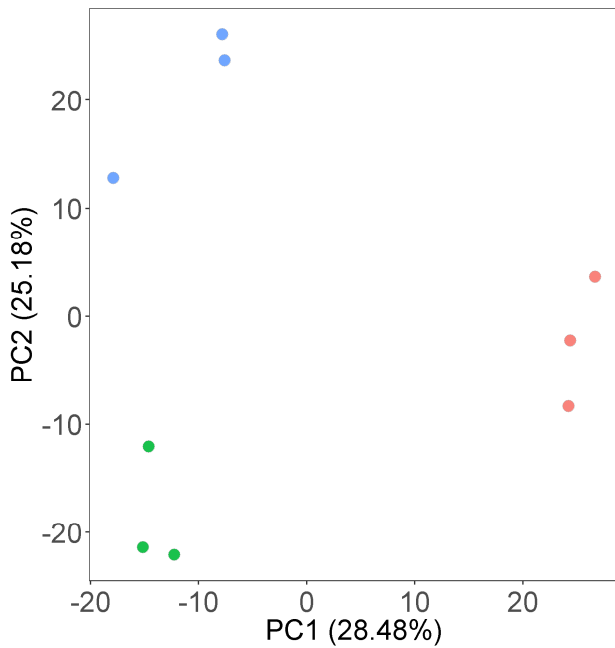


Figure 7.3.2 PCA of peptide fragments from proteomics. Analysis depicting that peptides clustered within sample and that there was a quantifiable difference between the susceptible versus resistant isolates but also between the

To assess global changes in gene expression as a result of *pgsa2* mutation we submitted triplicate cultures of the daptomycin susceptible *C. striatum* PS and its resistant isogenic clones PR and IR to the Washington University Proteomics core. Principal component analysis of the peptides indicate that protein expression profiles are similar within

replicate but that each strain has a distinct pattern of expression from one another, including PR and IR.

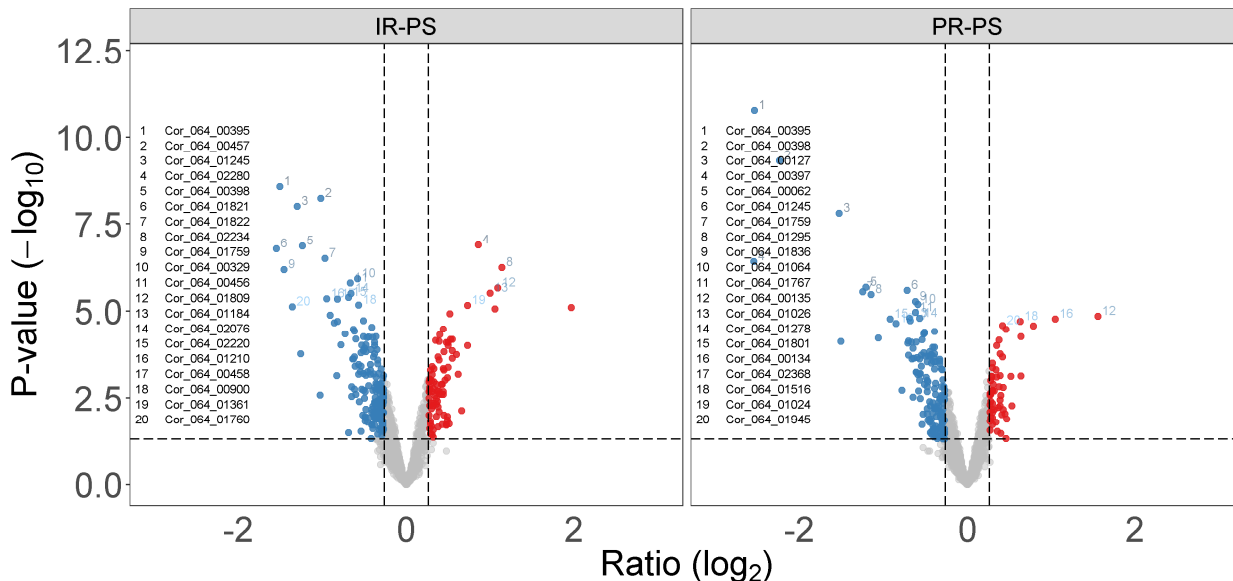


Figure 7.3.3 Volcano plot of differential abundant proteins. Comparison of peptide fragments between IR and PS and PR and PS were used to identify genes that had significantly different abundance between these conditions.

Interestingly, PS and PR are remarkably similar on the 1st PCA axis compared to IR, indicating that a number of protein changes may be attributed to the in vitro culture growth of IR. PS and IR both differ from PS on PC2 but interestingly IR is closer to PS on that axis. More than half of the diversity between these sets is explain by PC1 and PC2. Differential expression analysis was performed to identify significant protein differences between PS-PR and PS-IR. We then investigated the common proteins that are similarity downregulated or upregulated in these two comparisons. 20 proteins were commonly downregulated and 3 were upregulated

Gene	Prokka Annotation	COG	Function
Cor_064_00395	Fimbrial subunit type 1	M	cell wall anchor domain protein
Cor_064_00397	hypothetical protein	n/a	n/a
Cor_064_00398	hypothetical protein	M	Cna protein B-type domain
Cor_064_00457	hypothetical protein	C	phosphate acetyltransferase
Cor_064_00553	hypothetical protein	S	endonuclease exonuclease phosphatase
Cor_064_00656	Bifunctional ligase/repressor BirA	H	biotin acetyl-CoA-carboxylase ligase
Cor_064_00901	Molybdopterin synthase catalytic subunit 2	H	Molybdopterin
Cor_064_01024	Bifunctional protein PyrR	F	Also displays a weak uracil phosphoribosyltransferase activity which is not physiologically significant (By similarity)
Cor_064_01026	Dihydroorotase	F	dihydroorotase
Cor_064_01090	hypothetical protein	S	n/a
Cor_064_01210	Proton/glutamate-aspartate symporter	C	sodium dicarboxylate symporter
Cor_064_01245	hypothetical protein	S	Protein of unknown function (DUF3117)
Cor_064_01277	Urocanate hydratase	E	urocanate hydratase (EC 4.2.1.49)
Cor_064_01295	hypothetical protein	n/a	n/a
Cor_064_01759	Nitrate reductase alpha subunit	C	nitrate reductase, alpha subunit

Cor_064_01760	Respiratory nitrate reductase 2 beta chain	C	nitrate reductase beta
Cor_064_01767	hypothetical protein	P	ABC transporter, periplasmic molybdate-binding protein
Cor_064_01817	putative protein	E	Extracellular solute-binding protein, family 5
Cor_064_01922	Glutamine synthetase	E	glutamine synthetase
Cor_064_02220	hypothetical protein	J	UPF0176 protein

Table 7.3.3 Proteins that are commonly downregulated in PR and IR when compared against PS.

<i>Gene</i>	<i>Prokka Annotation</i>	<i>COG</i>	<i>Function</i>
Cor_064_01516	Manganese ABC transporter substrate-binding lipoprotein	P	transporter substrate-binding protein
Cor_064_01517	Zinc import ATP-binding protein ZnuC	P	ABC transporter
Cor_064_02052	hypothetical protein	P	ABC transporter

Table 7.3.4 Proteins that are commonly upregulated in PR and IR when compared against PS.

After delving deeper into the individual proteins that were downregulated by function, we learned that the metal ion molybdate is used to create a protein cofactor molybdopterin, which is a necessary cofactor for nitrate reductase activity, therefore linking Cor_064_01767, Cor_064_00901, Cor_064_01759, and Cor_064_01760 together. We hypothesized that due to the membrane restructuring during daptomycin resistance development these protein products are unable to properly localize to the membrane and are therefore degraded. Given that nitrate reductase is important for anaerobic growth, we did a semiquantitative growth assay by doing a 4 quadrant streak of PS, PR,

and IR, and then growing for 4 days. We found that PS grow up to the 4th quadrant and had lawns in quadrants 1-3 where as PR and IR only had single colonies in the 1st quadrant. We were going to do a more quantitative growth assay before being rudely interrupted by the SARS-2 novel coronavirus.

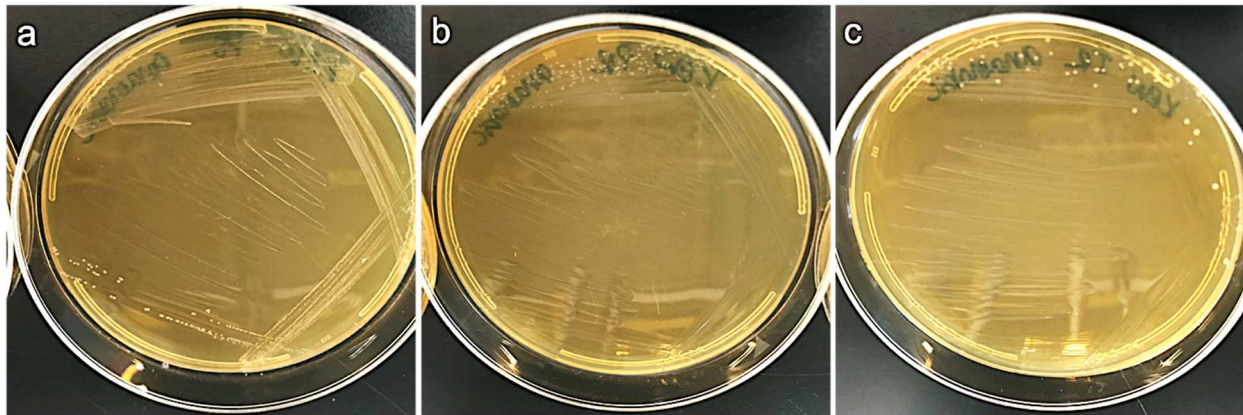


Figure 7.3.4 Four quadrant streak of PS (a), PR (b), and IR (c) under anaerobic conditions after 96 hours. Semiquantitative streaking depicts impaired growth for PR and IR as they are only able to grow single colonies in the first quadrant while PS can grow lawns in quadrants 1-3 and single colonies in quadrant 4.

7.4 Discussion

Given the importance of daptomycin as a therapeutic of last resort and the growing incidence of *Corynebacterium* attributed infections, the rapid development of antibiotic resistance capable by this pathogen is alarming. Previous efforts from the Dantas Lab and others have determined that in *Corynebacterium striatum*, the mechanism of phenotypic resistance is due to null mutations in *pgsa2* preventing presence of PG in the cell membrane. It is been demonstrated that some *Enterococcus faecium* and *Enterococcus faecalis* isolates are also capable of completely losing PG to gain daptomycin resistance, however a number of susceptible-resistant pairs did not do this- indicating that other mechanisms of resistance may occur(8). This gives strength to

our observation that a number of our *Corynebacterium* susceptible-resistant pairs do not have mutations in *pgsa2* allele. A diverse number of mutations have been identified in *Staphylococcus* and *Streptococcus* which lead can lead to different outcomes such as alterations in cell charge and thickening of cell wall to cause daptomycin resistance(9, 10). Further work is therefore warranted to investigate these possible mechanism of resistance in isolates without *pgsa2* mutations and also performing lipidomic analysis across the cohort to examine true presence of PG.

In addition to analyzing the mechanism of daptomycin resistance across this set of *Corynebacterium* species, we were also interested in more in depth analysis of the consequences of daptomycin resistance development in *C. striatum*. From a BiOLOG phenotypic microarray we were able to identify several compounds that had increased efficacy at preventing two daptomycin resistant strains (PR and IR) grow in rich media relative to an isogenic daptomycin sensitive strain (PS). Given the bulky aromatic moieties on 5/6 and the long hydrophobic chains of alexidine, it is possible that these compounds are able to more efficiently perturbate membrane integrity of *C. striatum* as it becomes daptomycin resistance and sheds PG. For *enterococcus*, it was determined that the loss of PG comes with a large decrease in membrane fluidity. In *S. aureus*, it was demonstrated that resistance to daptomycin can also increase resistance to other membrane intercalating agents, suggesting that alterations in phenotype may be species or mechanism of action specific(11). We did not see any increased resistance to the other cell envelope targeting antibiotics vancomycin or telavancin in our cohort. To the best of our knowledge there has only been one study analyzing global changes in microbial biology between daptomycin sensitive and resistant isolates but this was in

S. aureus. Interestingly, while we see similar changes in the cluster of orthologous group represented but not in the respective gene identities. One of our most interesting observations that genes involved in nitrate reductase and production of molybdenum were absent in the daptomycin resistance strains was not observed for *S. aureus*. The interpro page for nitrate reductase complex

(<https://www.ebi.ac.uk/interpro/entry/InterPro/IPR006468/>) indicates that there is a direct interaction between the complex and PG. Further work is therefore warranted to validate our observation that the daptomycin resistant isolates have a growth defect when grown anaerobically. Given the limitation of oxygen in different body sites this may also mean that resistant *C. striatum* isolates have growth defects in certain infection niches.

7.5 Materials & Methods

7.5.1 Clinical and computational cohort

We constructed a cohort of 198 clinical *Corynebacterium* isolates from patient samples obtained at Barnes-Jewish Hospital at Washington University in St. Louis, NorthShore at Northwestern, and Weill Cornell hospitals. To assay for daptomycin resistance evolution, 5mL tryptic soy broth (TSB) was inoculated with the wildtype isolate at .5 McFarland standard and a daptomycin Etest strip was cut in half and placed in the inoculum. The inoculum was incubated at 35°C for 24 to 48 hours, at which timepoints the media was checked for turbidity. Media with suspected growth was used as inoculum and streaked onto a blood agar plate to grow the resistant isolate. Phenotypic resistance was tested using a daptomycin Etest gradient diffusion strip.

Genomic DNA was isolated using bacteremia kit (Qiagen) and converted into Illumina sequencing libraries with the nextera protocol(12). Samples were pooled and sequenced on an Illumina NextSeq 2500 system. Reads were demultiplexed by barcode, had adapter content removed with trimmomatic, and had contaminating reads removed with deconseq(13, 14). Processed reads were assembled de novo using spades and had open reading frames annotated with prokka(15, 16). To confirm isolate species, we accessed publicly available NCBI type strain sequences of 86 recognized *Corynebacterium* species. To further classify the *C. striatum* isolates, we accessed publicly available sequences of 81 clinical *C. striatum* strains isolated in Beijing, China (17).

Sequenced isolates, type strains, and additional *C. striatum* genomes had average nucleotide analysis performed using pyANI. Heatmap was clustered hierarchically using seaborn. Confirmed *C. striatum* isolates had pan-genome identified using roary and the core-genome was clustered with prank(18). Alignment file was converted into a newick tree using fasttree and viewed with itol(19). For all resistant pairs we used bowtie2 to map the reads from the resistant isolate to the susceptible isolate and identify SNPs.

7.5.2 Proteomic characterization

Frozen stocks of the original *C. striatum* patient sensitive (PS) isolate and its isogenic daptomycin evolved patient resistant (PR) and *in vitro* resistant (IR) were streaked out on blood agar plates and triplicate single colonies were grown up overnight in 2 mL TSB. 1:1000 dilution of the overnight culture was added to 50 mL of TSB and grown aerobically to .05 OD600 in mid log phase. The samples were spun down and supernatant was removed. The samples were given to the Proteomics Core Laboratory

at Washington University in St. Louis for analysis with tandem mass tag. Peptides were quantified and mapped to the proteome for differential expression analysis. Cluster of orthologous group identification was performed on the PS proteome with EggNog v5(20).

7.5.3 BiOLOG chemical sensitivity assay

Frozen stocks of PS, PR, and IR were streaked out on blood agar overnight and a suspension was made from colonies in the fourth quadrant in TSB. The bacteria were all normalized to .05 OD600 and had 200 uL added to each well of pre loaded BiOLOG plates (PM11C, PM12B, PM13B, PM14A, PM15B, PM16A, PM17A, PM18C, PM19, and PM20B). The loaded plates were grown overnight at 37 °C and then assayed. Relative growth differences for PR and IR compared to PS were used to identify compounds that had the greatest ability to disrupt growth of the strains.

7.5.4 Anaerobic growth

Frozen stocks of PS, PR, and IR were streaked out on blood agar and placed in a 37 °C incubator in anaerobic chamber. Growth was monitored and the plates were removed after 4 days for picture assay.

7.6 Acknowledgments

We thank members of the Dantas lab for insightful discussions of the results and conclusions. The authors would like to thank Center for Genome Sciences & Systems Biology staff Brian Koebe and Eric Martin for operation of the High-Throughput Computing Facility. The authors additionally thank Center for Genome Sciences & Systems Biology staff Jessica Hoisington-Lopez and MariaLynn Jaeger for performing

the Illumina sequencing and demultiplexing. RFP was supported by a NIGMS training grant through award T32 GM007067 (PI: James Skeath) and the Monsanto Excellence Fund graduate fellowship. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication. I would also like to thank Kate Wardenburg for her continued contributions.

7.7 References

1. Burkovski A. The role of corynomycolic acids in *Corynebacterium*-host interaction. *Antonie Van Leeuwenhoek*. 2018;111(5):717-25. doi: 10.1007/s10482-018-1036-6. PubMed PMID: 29435693.
2. Datta P, Gupta V, Gupta M, Pal K, Chander J. *Corynebacterium striatum*: An emerging nosocomial pathogen. *Infect Disord Drug Targets*. 2020. doi: 10.2174/1871526520666200224103405. PubMed PMID: 32091348.
3. McMullen AR, Anderson N, Wallace MA, Shupe A, Burnham CA. When Good Bugs Go Bad: Epidemiology and Antimicrobial Resistance Profiles of *Corynebacterium striatum*, an Emerging Multidrug-Resistant, Opportunistic Pathogen. *Antimicrob Agents Chemother*. 2017;61(11). doi: 10.1128/AAC.01111-17. PubMed PMID: 28848008; PMCID: PMC5655097.
4. Heidary M, Khosravi AD, Khoshnood S, Nasiri MJ, Soleimani S, Goudarzi M. Daptomycin. *J Antimicrob Chemother*. 2018;73(1):1-11. doi: 10.1093/jac/dkx349. PubMed PMID: 29059358.

5. Goldner NK, Bulow C, Cho K, Wallace M, Hsu FF, Patti GJ, Burnham CA, Schlesinger P, Dantas G. Mechanism of High-Level Daptomycin Resistance in *Corynebacterium striatum*. *mSphere*. 2018;3(4). doi: 10.1128/mSphereDirect.00371-18. PubMed PMID: 30089649; PMCID: PMC6083094.
6. Ajmal S, Saleh OA, Beam E. Development of High-Grade Daptomycin Resistance in a Patient Being Treated for *Corynebacterium striatum* Infection. *Antimicrob Agents Chemother*. 2017;61(7). doi: 10.1128/AAC.00705-17. PubMed PMID: 28483949; PMCID: PMC5487685.
7. Hines KM, Waalkes A, Penewit K, Holmes EA, Salipante SJ, Werth BJ, Xu L. Characterization of the Mechanisms of Daptomycin Resistance among Gram-Positive Bacterial Pathogens by Multidimensional Lipidomics. *mSphere*. 2017;2(6). doi: 10.1128/mSphere.00492-17. PubMed PMID: 29242835; PMCID: PMC5729219.
8. Mishra NN, Bayer AS, Tran TT, Shamoo Y, Mileykovskaya E, Dowhan W, Guan Z, Arias CA. Daptomycin resistance in enterococci is associated with distinct alterations of cell membrane phospholipid content. *PLoS One*. 2012;7(8):e43958. doi: 10.1371/journal.pone.0043958. PubMed PMID: 22952824; PMCID: PMC3428275.
9. Garcia-de-la-Maria C, Xiong YQ, Pericas JM, Armero Y, Moreno A, Mishra NN, Rybak MJ, Tran TT, Arias CA, Sullam PM, Bayer AS, Miro JM. Impact of High-Level Daptomycin Resistance in the *Streptococcus mitis* Group on Virulence and Survivability during Daptomycin Treatment in Experimental Infective Endocarditis. *Antimicrob Agents Chemother*. 2017;61(5). doi: 10.1128/AAC.02418-16. PubMed PMID: 28264848; PMCID: PMC5404581.

10. Barros EM, Martin MJ, Selleck EM, Lebreton F, Sampaio JLM, Gilmore MS. Daptomycin Resistance and Tolerance Due to Loss of Function in *Staphylococcus aureus* *dsp1* and *asp23*. *Antimicrob Agents Chemother*. 2019;63(1). doi: 10.1128/AAC.01542-18. PubMed PMID: 30397055; PMCID: PMC6325204.
11. Fischer A, Yang SJ, Bayer AS, Vaezzadeh AR, Herzig S, Stenz L, Girard M, Sakoulas G, Scherl A, Yeaman MR, Proctor RA, Schrenzel J, Francois P. Daptomycin resistance mechanisms in clinically derived *Staphylococcus aureus* strains assessed by a combined transcriptomics and proteomics approach. *J Antimicrob Chemother*. 2011;66(8):1696-711. doi: 10.1093/jac/dkr195. PubMed PMID: 21622973; PMCID: PMC3133485.
12. Baym M, Kryazhimskiy S, Lieberman TD, Chung H, Desai MM, Kishony R. Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS One*. 2015;10(5):e0128036. doi: 10.1371/journal.pone.0128036. PubMed PMID: 26000737; PMCID: PMC4441430.
13. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-20. doi: 10.1093/bioinformatics/btu170. PubMed PMID: 24695404; PMCID: PMC4103590.
14. Schmieder R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS One*. 2011;6(3):e17288. doi: 10.1371/journal.pone.0017288. PubMed PMID: 21408061; PMCID: PMC3052304.
15. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its

- applications to single-cell sequencing. *J Comput Biol.* 2012;19(5):455-77. doi: 10.1089/cmb.2012.0021. PubMed PMID: 22506599; PMCID: PMC3342519.
16. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics.* 2014;30(14):2068-9. doi: 10.1093/bioinformatics/btu153. PubMed PMID: 24642063.
17. Wang X, Zhou H, Chen D, Du P, Lan R, Qiu X, Hou X, Liu Z, Sun L, Xu S, Ji X, Li H, Li D, Zhang J, Zeng H, Li Z. Whole-Genome Sequencing Reveals a Prolonged and Persistent Intrahospital Transmission of *Corynebacterium striatum*, an Emerging Multidrug-Resistant Pathogen. *J Clin Microbiol.* 2019;57(9). doi: 10.1128/JCM.00683-19. PubMed PMID: 31315959; PMCID: PMC6711910.
18. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics.* 2015;31(22):3691-3. doi: 10.1093/bioinformatics/btv421. PubMed PMID: 26198102; PMCID: PMC4817141.
19. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One.* 2010;5(3):e9490. doi: 10.1371/journal.pone.0009490. PubMed PMID: 20224823; PMCID: PMC2835736.
20. Huerta-Cepas J, Szklarczyk D, Heller D, Hernandez-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, von Mering C, Bork P. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 2019;47(D1):D309-D14. doi: 10.1093/nar/gky1085. PubMed PMID: 30418610; PMCID: PMC6324079.

Chapter 8: General Conclusions

In my thesis work I have investigated bacteria from a diverse range of Gram-negative (including *Proteus*, *Klebsiella*, *Acinetobacter*) and Gram-positive bacteria (*Gardnerella*, *Enterococcus*, *Corynebacterium*). Unifying the diverse goals of each chapter is the utility of whole-genome sequencing for improved taxonomy, understanding of intra-species diversity, and analysis of genes relevant for human infection. When combined with experiments designed to test these functional consequences of microbial diversity, including co-culture assays, antibiotic susceptibility testing, mouse infection, and human cohorts we can learn more about bacterial pathogens.

The use of whole-genome sequencing for improved taxonomic resolution was a major component of Chapters 3, 4, and 5. In Chapter 3 I started with *Klebsiella variicola*, which was known to be genetically dissimilar to *Klebsiella pneumoniae* and *Klebsiella quasipneumoniae*, but that distinction had not filtered down to the clinical laboratories until this project. An updated MALDI-TOF MS database then showed that a sizeable portion of what had been historically been called *K. pneumoniae* was actually *K. variicola*. Importantly, we demonstrate that some of these *K. variicola* strains were capable of causing greater infections in mice compared to the canonical pathogen *K. pneumoniae*. In contrast to the targeted taxonomic analysis in chapter, in chapter 4 I analyzed a diverse set of Gram-negative and Gram-positive bacteria that were collected in the same hospital. We found that certain bacteria were reliably identified by the non whole-genome based method of MALDI-TOF MS (ie. *Pseudomonas aeruginosa* and *Enterococcus faecium*) but a number of bacteria were misidentified. This sheds light on

the observation that rarer pathogens or environmental organisms may not be reliably identified by MALDI-TOF MS. Finally, in chapter 5 we start with one single species, *Gardnerella vaginalis*, but based off of past literature expected to identify multiple species being erroneously called as one. Thankfully our *in silico* efforts paid off and we determined that *G. vaginalis* may be considered 9 species. Further work is required to determine the true context of this genus within the bifidobacteriaceae.

The use of whole-genome sequencing for analysis of genes relevant to human infection was a major component of Chapters 2, 6, and 7. In Chapter 2 we found that a case of unidentified carbapenem resistance could be attributed to the presence of *bla*_{IMP-27} in a *Providencia rettgeri*. We test the functional consequences of this gene by demonstrating that *P. rettgeri* PR-1 and an additional *Proteus mirabilis* isolate PM187 are capable of conjugating resistance into *E. coli* J53. In chapter 6 we took advantage of the *E. faecium* isolates from Chapter 4 and added more isolates sequenced in the United States to study resistance to the critical antibiotic linezolid. Importantly, we found that the same phenotype of linezolid resistance can be attributed to two parallel but non overlapping mechanisms of resistance. In the United States cohort, all isolates had a G2576T SNP in their 23S rRNA loci, which has been demonstrated to prevent linezolid binding. In contrast, the Pakistan cohort all contained either 2 or 3 acquired ARGs. These ARGs represent two linezolid specific efflux pumps (*optrA* and *poxA*) as well as a novel *cfm* methyltransferase variant. Even within the Pakistan cohort we found the functional consequences of ARG carriage as isolates with all 3 resistance genes had a significantly higher minimum inhibitory concentration against linezolid compared to isolates that only had 2 ARGs. Further worrisome is the greater potential for these

ARGs to pass amongst pathogenic strains. Finally, in Chapter 7 we analyzed the functional consequences of antibiotic resistance in *Corynebacterium striatum* isolates that evolve high level daptomycin resistance. We initially used WGS to determine that a number of these isolates may evolve resistance via non *pgsa2* methods and then proteomics to determine that null mutations which confer *pgsa2* resistance drastically alter the repertoire of proteins that are produced. Importantly, we found impairment in nitrate reductase and a necessary cofactor which could impair the daptomycin resistant *C. striatum* isolates at growing anaerobically. We then tested that indeed the resistance strains were unable to grow as well as the susceptible strain under anerobic conditions.