Distribution Agreement

Signature:

Brooke Morgan Talbot Date

Genomic epidemiology of bacterial pathogen transmission, persistence, and resistance

By

Brooke Morgan Talbot
Doctor of Philosophy

Graduate Division of Biological and Biomedical Sciences
Population Biology, Ecology, and Evolution

____
Timothy D. Read
Advisor

____
Sarah Bowden
Committee Member

____
Katia Koelle
Committee Member

____
Anne Piantadosi
Committee Member

Accepted:

____
Kimberly Jacob Arriola, Ph.D, MPH
Dean of the James T. Laney School of Graduate Studies

____
Date

Genomic epidemiology of bacterial pathogen transmission, persistence, and resistance

By

Brooke Morgan Talbot

M.Sc., The George Washington University, 2018

Advisor: Timothy D. Read, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Graduate Division of Biological and Biomedical Sciences
Population Biology, Ecology, and Evolution

2024

Abstract

Genomic epidemiology of bacterial pathogen transmission, persistence, and resistance

By

Brooke Morgan Talbot

This dissertation broadens the application of evolutionary concepts within applied epidemiology to enhance and go beyond traditional case detection and diagnostics. It aims to identify the strengths and limitations of genomic approaches when paired with clinical and epidemiological data. I use methicillin-resistant *Staphylococcus aureus* (MRSA) bloodstream infections as a model for exploring within host and between host pathogen evolution and to test the capacity of single species comparative genomics to detect epidemiological linkages. I also expand on the relationship between genetic distance and spread using a metagenomic analysis of antimicrobial resistance (AMR) genes in two different colocalized hosts, humans and gray mouse lemurs. First, I critically evaluate the use of single nucleotide polymorphism (SNP) thresholds in hospital-associated spread of *S. aureus* and *Pseudomonas aeruginosa*. I argue that a one-size-fits-all approach for SNP difference is insufficient due to evolutionary and ecological differences influencing genomic variability, even within the same epidemiological setting. I next investigate whether patients experiencing MRSA bacteremia exist in genomic clusters with epidemiological links based on SNP distance. I identified that genomic alignment strategy, and the genetic background of strains affect the detection of SNP differences and that bacteremia patients in clusters have common healthcare exposures long before illness onset. I then examine risk factors for MRSA bacteremia recurrence and whether recurrent strains share convergent adaptive traits. I show that in our study set most recurrent infections are relapses from previous strains. These relapse lineages exhibit signatures of positive selection, particularly in genes associated with antibiotic resistance and virulence. Finally, I characterized the AMR resistomes between human and lemur gut microbiomes using metagenomics. The study identified distinct bacterial species profiles but shared antimicrobial resistance genes between hosts and suggests that AMR gene spread is diffuse in this system. This research demonstrates how genomics offers more precise and predictive public health interventions by refining our understanding of pathogen transmission and recurrence and emphasizing that evolutionary dynamics beyond neutrally evolving genes demark epidemiological linkages.

Genomic epidemiology of bacterial pathogen transmission, persistence, and resistance

By

Brooke Morgan Talbot

M.Sc., The George Washington University, 2018

Advisor: Timothy D. Read

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Graduate Division of Biological and Biomedical Sciences
Population Biology, Ecology, and Evolution

2024

Acknowledgements

I have a tremendous amount of love and gratitude for the communities that have helped me through this scientific journey. How lucky am I to have a home in so many places! Thanks foremost to Dr. Timothy Read for his fervent mentorship. He has been generous in his instruction and encouragement, and I am grateful for the welcoming environment that he has cultivated in our lab. Thank you for helping me see new scientific paths when I struggled with my own vision and for helping me gradually grow from a trainee to a collaborator. Thank you also to my dissertation committee, Dr. Sarah Bowden, Dr. Katia Koelle, and Dr. Anne Piantadosi for providing critical feedback and partnership in this thesis. Each of you took individual time to meet with me and review my work and questions, and I am thankful for your considerate efforts. The work in this dissertation would not have been possible without considerable collaborations, specifically Dr. Michael David and his research group at The University of Pennsylvania, and Dr. Thomas Gillespie and his collaborations with Centre ValBio, MICET, and PIVOT. This completed work centers on understanding bacterial biology in the context of community health, and both Michael's and Tom's expertise and longstanding relationships with the communities that this research serves have been so important for learning to how to translate this work accordingly.

I am deeply grateful to a wonderful group of mentors who believed in my career development as a scientist. Thank you to Dr. Hartmut Doebel for instilling an early sense of independence and excitement to chase big scientific questions. Thank you to Dr. Joseph Scafidi for being vigorously supportive of me and helping me develop strong research practices at the bench. Thank you to Dr. Aileen Chang, Dr. Jeff Bethaney, Dr. Jeanne Jordan, Dr. Mimi Ghosh, Kerri

Dorsey, and Jin Peng for your advisory during my master's degree and for teaching me about the vital roles of public health departments in disease prevention and detection. Thank you also to Vasudha Reddy, HaeNa Waechter, Dr. Bruce Gutelius, the "Foodies", and the amazing community of scientists and public health practitioners at NYC DOHMH. I am indebted to you for substantiating in me a passion for public health practice and giving a deep and meaningful opportunity to support community wellbeing. Thank you to Dr. Kathleen Campbell for showing me how to be a compassionate and organized undergraduate instructor. Thank you to Dr. Christin Hanigan, Dr. Matthew Doucette, Dr. Andrew Lang, the "µbioinfo" virtual community, and the Pathogen Genomics Centers of Excellence community for cheering me on and sharing your knowledge of public health genomics.

Thank you to the PBEE community for guidance and kindness. Thank you to Rosalyn Lightfoot for your work as the PBEE Program Administrator and for helping to run the PBEE Seminar Series. Thank you to Emily Neutens, Maureen Thomas, Kathy Smith, Mikko Lindberg, and Dr. Nicole Gerardo for all you do to support students throughout their graduate career. To all my co-representatives in the GDBBS Divisional Student Advisory Council: you were some of my first friends at Emory even when we could not meet in person and for that I am so thankful. Thank you to my cohort mates Carmen Alvarez, Mitchell Kendzel, Marissa Duckett, and honorarily Yeongseon Park. I have never met more stick-to-it scientists like you all especially as we navigated what starting a PhD looked like in the height of a pandemic.

To my lab family past and present – thank you. The EMERGENT group is special with no shortage of intelligent and open scientists, all of whom have shared feedback and support in excess, as well as their comradery. Special thanks to: Dr. Robert Petit, thank you for taking generous time to help me become a better bioinformatician and scientist. Dr. Kim Hoang, thank you for your guidance in assembling this document and for your kind and thoughtful feedback always. Dr. Bri Bixler, Ana Ramos Facio, and Dr. Emily Wissel, I am so thankful for every piece of technical or scientific advice paired with sweet cat photos. Thank you, Dr. Ashley Alexander for being a mentor and friend since the moment I met you as my PBEE recruitment buddy, and to Jacoby Robinson for your openness, honesty, and humor that has helped me find my own peace. Megan Phillips, thank you for sharing in the responsibilities, joys, and stressors of being a program leader, and for the many wonderful escapes in and outside of Atlanta. You make the science journey fun and have changed the way I think and approach evolutionary theory. Dr. Vishnu Raghuram, thank you for endless patience, good humor, and quality frog content. To have you to count on in times of success and struggle was and remains invaluable. Dr. Katrina Hofstetter, thank you for being a rubber duck and a rock. I look up to you in the lab and am glad to collaborate on all kinds of projects, including Jello ones. Thanks also to Alex Banul and Anoch Mohan for trusting me as your mentor. I know you will do great work wherever your careers take you. Many other amazing scientists who call HSRB their lab home have contributed to my wellbeing as a scientist. They have made the space warm and welcoming and were always ready to offer guidance and resources at the bench and otherwise. Thank you! Thanks especially to Dr. Brenda Antezana, Alexa Avecilla, and Emilio Rodriguez for your friendship and bright energy.

I am grateful to the friends outside of my research space that have offered me many words of encouragement and were sources of inspiration during this dissertation. Thanks to the folks at Highland Runners and Atlanta Running Meetup for offering a friendly outlet to reset and refocus my mind and helped me literally and figuratively chase my goals. I am especially grateful to Dr.

Marly van Assen, Dr. Liselotte de Wit, and Matheus Meneghel, who have been sources of wisdom for enduring the PhD process and the best hype people for science and running alike. To the friends who have had my back through the years as I have bounced from place to place, thank you for sticking by my side. I especially thank Tyler Wolanin, my faithful writing partner who never lets me forget to do good work; Rachel Donato and Jen Urbanowski, who keep me imaginative; Esther Schenau, who always helps me see the world in new lights; and Dominique and Mitchell Scarbrough who always makes me feel like I am home. I must acknowledge, too, "Dr." Scully, the most credentialed cat in my household (And thanks to Bri who is responsible for finding her!). Scully has contributed much code and many lines of text to this dissertation (unfortunately lost to revisions) and has brought me joy on days where joy felt impossible.

It is because of the support of my family that I have achieved any success. This journey has been a lifetime, and through it all I have been able to lean on the love of my siblings Brittaney, Amanda, Sarah, and Ben, and my siblings-in-law Brian Martone, Ben Williams, and Chad Wyszynski. Thank you for housing me when I visit from wherever I am coming from and for sharing in all the gritty, silly, glamorous, and mundane days along the way. You are my world and always a beacon back to what matters. Most especially thank you to my parents, Gary and Susan Talbot, who have supported me to become the person and scientist I am today. Your own adventurous nature and independence inspire me. You have listened through practice talks with excitement and curiosity, invigorated my confidence, and invested time and resources so that I and the rest of my siblings could chase down our dreams, career and otherwise. In completing this work, I honor your sacrifices and love.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Overview of Pathogen Genomic Epidemiology

Genomic epidemiology of infectious diseases seeks to use the detection of nucleic acids, such as DNA and RNA, to measure disease distributions within a population and identify biological mechanisms that underlie disease (1–3). The discipline draws from, but is distinctly separate from, genetics, population biology, taxonomy and molecular biology, and is shaped by the availability of technology, especially sequencing and bioinformatics (2). It is therefore vital that public health scientists gain an educational grounding in bioinformatics, surveillance, and population genetics together (4).

In recent years nucleic acid sequencing technology has become highly sensitive and affordable, making genomic surveillance of infectious disease more accessible to public health and clinical laboratories (5–7). The focus on nucleic acids allows for methods to be adapted to different taxonomic groups, including bacteria, fungi, eukaryotic parasites, and DNA and RNA viruses causing human disease. Previously, pulsed-field gel electrophoresis (PFGE), which fragments DNA into specific patterns or "fingerprints," was widely implemented for bacterial disease surveillance and was the standard for matching infection isolates together to identify a possible transmission event for nearly three decades (6). Though PFGE's pattern matching is simple to interpret, it only captures a small portion of genetic information from the bacterial genome. For bacterial isolates, multilocus sequence typing (MLST) further increased the resolution of genetic similarity between isolates and accounted for horizontally obtained DNA by comparing housekeeping alleles across bacterial isolates (8). Both PFGE and MLST require up to date records of identified types, though, and cannot readily provide information about drug resistance or virulence. To address the limitations of earlier typing methods, scientists investigated the utility of whole genome sequencing (WGS) for surveillance. Today WGS has superseded previous technologies as the primary analysis for pathogen genomic epidemiology. The resolution of WGS supports the creation of highly resolved phylogenies and detection of many genes of interest, including drug resistance genes, all from a single sequence run. When analyzed isolates are then paired with high quality metadata, including exposure histories, clinical history of disease, and other ecological factors, transmission sources can be more accurately deduced.

Suites of bioinformatic tools and workflows exist to support full integration of genomics into routine surveillance today, each of which can be tailored to the pathogen taxonomy of interest. WGS of individual bacterial isolates (e.g. from DNA prepared from a single bacterial colony) is normally paired with detailed comparative analyses within a single species for relating evolutionary history with epidemiological causality. Metagenomics, where nucleic acids are isolated, amplified, and sequenced from a raw sample rather than a pure culture, is used largely to characterize diversity of microbial communities and for the unbiased detection of suspected disease-causing agents. Larger systems which have already adopted WGS for routine genomic surveillance include PulseNet in the U.S. to nationally reduce the burden of bacterial foodborne disease (5), and the global coalition PHA4GE supporting transmission tracking and contact tracing of SARS-CoV-2 infections (9). Developments regarding the sequencing and surveillance of pathogens with more complex life histories and genetics, such as C*ryptosporidium* species, are also underway (10).

The definition of "outbreak," while understood to be some level of increase in detected cases of illness over time, is dependent on the biology of the infectious agent, the environment in which it is detected, the primary mode of transmission, and the prevalence of the disease caused by the organism (11). For public health practice, scientists attempt to use comparative genomics as central evidence for characterizing the etiology, pathogenesis, source/reservoir identity, circulation, and transmission of a disease, as well as the development of vaccines and drug therapy (3). The challenge of defining bacterial outbreaks with WGS is that the molecular evidence is not static and exists on a continuum. That is, bacterial genomes are subject to many simultaneous pressures and evolutionary forces that can alter how much genetic similarity will ultimately be detectable between recently diverged isolates from a common source.

To understand how to detect and combat disease, it is imperative to understand the typical ways these organisms reproduce, compete with other species, move through their environment, and ultimately evolve in response to unique environmental pressures. For outbreak management, comparing genomic markers to the known life history of pathogens helps resolve the likelihood of transmission events when pathogens are endemic to an area, have high prevalence or are highly clonal species. Infectious bacteria have unique evolutionary histories, reproduction mechanisms, and prevalence from other pathogens. The diverse bacterial species that cause infection are of great interest among basic researchers and public health practitioners. Some of the major contemporary questions related to bacterial infections include how to relate mutation rate to epidemiological timelines, persistence of infections, and the evolution of drug resistance.

Antibiotic Resistance as a Public Health Target

Monitoring the prevalence and emergence of drug-resistant pathogens is a major goal of public health surveillance. Antibiotic resistance (AMR) in bacteria is defined as the ability of a cell to survive and/or grow in the presence of normally toxic antimicrobials. AMR phenotypes are common among all species of bacteria regardless of pathogen state. Antimicrobials are produced naturally by fungi and other bacteria as a defense against other invading bacteria, and they are also artificially synthesized for clinical use (12). Unique antibiotic molecules target essential physiology of the bacteria, targeting cell wall and cell membrane growths, DNA replication, and protein synthesis (13). Antibiotics can have a bactericidal effect on cells, i.e. causing rapid cell death, and they can also be bacteriostatic, i.e. limiting growth and cell density in the population (13). Naturally occurring antimicrobials may also act as signaling molecules for physiological changes and gene expression, leaving much still to be explored about the anthropogenic impact on increased resistance in natural populations (12).

Consequently, bacterial populations evolve traits that directly and indirectly resist antibiotic chemicals. AMR can be intrinsic where an entire species is naturally resistant to specific classes of antibiotics. Certain classes of antibiotics affect bacteria differently depending on the contents of their cell walls and cell membranes (14). Efflux pumps protect bacteria by eliminating foreign or toxic substrates that enter the cell and are often associated with broad spectrum resistance (14). Bacteria can also acquire resistance relative to the adaptive landscape that certain members of the species experience. Acquired resistance occurs because of de novo mutations on the chromosome that promote a fitness advantage in the presence of an antimicrobial chemical, which is then passed on to daughter cells vertically. These mutations remain within the cell lineage and are a concern for chronically infected individuals (15,16). Bacteria also acquire antimicrobial resistance traits through horizontal gene transfer (HGT), which is a major mechanism of broad resistance between strains and different species of bacteria (17). Genes can move through transduction conferred by a viral phage, transformation, or uptake of free DNA in the environment, and conjugation, which requires cell-cell contact. Mobile resistance genes conferred by HGT are determined by the type of transfer mechanism and cellular restrictions that allows a non-susceptible bacterium to incorporate or expel foreign DNA (18), which in turn is affected by the spatial structure that allows for cellular interaction. Therefore, not all bacteria will be able to acquire resistance in the presence of a resistant cell. However, it is these mobile genetic elements that raise the most concerns for the widespread treatment failure in currently treatable infections (17,19). The combination of intrinsic and extrinsic factors leading to drug resistance has led to the emergence of multidrug resistance further exacerbating the difficulty treating these infections (20–22).

Evidence shows that genes that confer AMR have long evolutionary histories independent of clinical usage (14, 23). However, the massive application of antimicrobial agents since the early 20th century in medicine and agriculture has resulted in a huge increase in resistant bacterial pathogens of humans and animals alike (24). Bacteria must pass through several gauntlets to pass on genetically acquired drug resistance: the volume of antibiotic exposure, survival in the infection environment, susceptibility to HGT, and ease of transmission to another host. Overwhelmingly, human activity determines how likely and how far drug resistance will spread, and this is largely in the context of clinical and agricultural practices (12, 25). Poor penetration of an antibiotic into infected tissue or incompletion of therapy allows for small subsets infecting bacteria to survive and become cryptic reservoirs for possible onward transmission. Antibiotic tolerance, in which the time it takes to kill bacteria at the same dosage of treatment, increases the likelihood of mutations occurring that result in resistant subpopulations (26). Variation in the fitness effects of AMR genes can also change their prevalence, especially if genes are co-selected with other traits. For example, application of metals to agricultural soils can select for genetic elements that confer resistance to metal exposure and antibiotics simultaneously (27). Antimicrobial exposure is critical for genetic emergence of resistance in bacteria, but this selective pressure is not the only force that maintains resistance genes in a bacterial population. Systems that are more open may cause the migration of horizontally transferred genetic elements to populations with susceptible bacteria despite low prevalence of antibiotics, such as the spread of genes from human-generated sewage into natural water systems or soil (28–30). Differently structured healthcare systems can also vary in their openness between the community and the healthcare setting, meaning that there is no one clinical protocol that can be executed to reduce the exchange or the risk of drug resistant bacterial infections (31). Onward transmission is then possible when known routes of infection are not monitored or controlled. Contaminated healthcare equipment and personal protective equipment in healthcare environments is an ongoing challenge in healthcare outbreaks (32–35), and poor management of sewage leads to agricultural contamination and outbreaks related to fecal-oral transmission (28,36,37).

For nearly a decade, The Global Action Plan on AMR has acted as an important framework for global stewardship of antimicrobial use and the prevention of drug-resistant infections. It especially calls for a One Health approach, a movement and programmatic framework which encourages the cooperation between many disciplines and sectors of expertise including and not limited to veterinary and human medicine, public health agencies, agricultural and environmental professionals, and the political and financial sectors (38). Individual efforts to combat resistance include drug stewardship programs (39), development and implementation of new antimicrobial therapies, and modernization and increased efforts in the surveillance of drug-resistant bacteria (40, 41). Given the high genetic association with phenotypic AMR, metagenomics and genomics can be used to assess the strength of contribution from individual drivers of resistance. Detection of the type of gene, its prevalence, and which species it is associated with can be connected to data related to its ecological context, and interventions that prevent onward transmission can be reassessed and re-evaluated. However, the pipeline to create new long lasting and effective antimicrobial therapies is slow and countries and communities alike still experience major gaps in access to systems and equipment that can help diagnose and monitor drug-resistant bacteria. Consequently, global prevalence of many resistant infections is still largely unknown.

Staphylococcus aureus: A Pathogenesis Model

In this thesis I use *Staphylococcus aureus* as a model for identifying epidemiological linkages from genomic changes. I selected *S. aureus* because it possesses a mix of stable, clonal lineages with strong geographic and epidemiological linkages, while also exhibiting patterns of convergent traits across highly divergent members of the species. This allows for exploration of the speed and scale of evolution during transmission and results in the interesting duality of *S. aureus* as a "asymptomatic" colonizer that can frequently cause serious infections. Furthermore, *S. aureus* has well-defined sub-species groupings (aka "strains") with distinct phenotypes, diseases and hosts associated with them. The clinical and public health significance, unique human hosts-adapted virulence traits, and genomic architecture make *S. aureus* a well-suited organism to serve as a model for understanding within-host evolution and pathogen transmission from a genomic perspective.

S. aureus is a pathogen of global concern. It is a gram-positive coccus species that forms "grapelike" clusters of cells. It was first described by Alexander Ogston in 1882 from surgical wounds (42) and further characterized as its own species in 1884 by Frederich Rosenbach based on the distinct yellow coloration of colonies (43). *S. aureus* colonizes humans and domesticated animals as well as some wild animals (44). It can cause a range of infection types including skin and soft tissue infections (45), toxic-shock syndrome (46), and deeply invasive infections associated with bone and joint infections (47, 48), pneumonia (49), and cardiovascular disease (50). It has also been implicated in outbreaks of illness associated with food and agriculture (44, 51) and healthcare exposure (52, 53).

S. aureus has acquired drug resistance to multiple classes of antibiotics, including methicillin. Methicillin resistant *S. aureus* (MRSA) infections cause over 300,000 infections in hospitalized patients, 10,000 deaths, and a burden of $1.7 billion dollars in healthcare costs in the US alone (54). Although MRSA lives commensally on the skin and in the nose for nearly two percent of adults in the US (55), it can be life-threatening when spread in healthcare settings (56,57). Therefore, community-associated spread leading to healthcare introductions is also an increasing infection control challenge (31, 58). Human population structure can also impact the expansion of *S. aureus* lineages (59). The most common MRSA lineages in the US are the USA300 strain (60, 61), a lineage with community-associated (CA) spread, and Clonal Complex 5 (CC5)/USA100 (61), which has been implicated in a variety of healthcare setting transmissions (62). Prior to the 1990s, MRSA rarely spread outside of healthcare settings (63). Rapid emergence of USA300 in the US in community and eventually healthcare settings demonstrated completely unique evolutionary, epidemiological and molecular patterns (31,64). Although CA and healthcare-associated (HA) lineages distinctly differ in clinical, demographic, and microbiological characteristics (63,65), CA strains add to the burden of HA infections (52, 66) and HA strains have transmitted between patients with no known healthcare exposure (67). The decreased incidence of MRSA in healthcare settings is largely due to improvements in healthcare infection control practices. However, the impact of these prevention efforts has slowed in the last decade (54).

S. aureus is notoriously equipped to attack the host and escape the innate and adaptive immune response. *S. aureus* produces a range of toxins that allow it to invade many different tissues. Toxins can damage host cells directly, including leukotoxins which target host immune cells and alpha-toxin which target red blood cells (68). Toxins produced by *S. aureus* can also alter receptor functions and lead to pathological disease for the host. Enterotoxins and toxic shock syndrome toxin are most notable for their association with severe disease states and outbreaks (46, 51, 68). *S. aureus* also produces enzymes to assist in invasion by breaking down host tissue proteins, and degrading or encouraging blood clots for evasion (69). Further *S. aureus* can evade the human immune system by producing superantigens and chemical blockers that prevent the recruitment of neutrophils and their ability to interact with host receptors (70). *S. aureus* can also evade clearance by complement through expression of staphylococcal protein A (70). If bacteria are phagocytized, they can withstand the release of reactive oxygen species using a suite of detoxifying enzymes, antioxidant pigments, and alteration of their cell wall (70–73). Altogether, *S. aureus* has an arsenal of virulence traits that contribute to the variety of diseases that it can cause.

Virulence (49, 65, 74, 75) and antibiotic resistance (76) can differ across *S. aureus* lineages. It has been documented that specific mutational signatures occur in pathogenesis-associated genes from infecting *S. aureus* that differed from commensal *S. aureus* in the same patient (77). Since different body sites can harbor *S. aureus* persistent infections, there are likely different selective pressures for within-host evolution, and consequently differential disease presentations depending on the site of invasion or colonization. Since *S. aureus* can have longevity on its host, long-term screening of *S. aureus* from the same host helps us understand the expectations of the differences in populations that likely arose from the same recent common ancestor relative to host colonization (78, 79).

Bacteria primarily reproduce through binary fission, producing identical daughter cells with vertically acquired copies of chromosomes. Small errors in the replication process introduce mutations into the daughter cells, and these mutations result in neutral effects on proteins or alterations to the proteins. Mutations that are not severely detrimental can then be passed along to other daughter cells and persist and become fixed in the population through random and non-random processes (i.e. selection, genetic drift, and migration). Over time, the genetic record generated through vertical transmission and mutational fixation is used to distinguish and trace subspecies lineages and can provide information about the likelihood of spread between hosts and within the environment. One limitation, however, is that mutation rates, usually defined as the average number of mutations per site in the genome per year during asexual growth, vary at species and subspecies levels due to differences in generation time, DNA replication proteins, and possible environmental mutagens (80). New genetic differences can also be introduced through HGT. Identifying instances of HGT can provide a lot of important information about potential environmental pressures that bacteria experience and possible epidemiological inference about antibiotic exposures. HGT, as well as homologous recombination, are non-descent associated genetic changes, and therefore they make phylogenetic estimations challenging. There are also barriers to genetic exchange in *S. aureus* that shape genetic diversity and species structure. For example, plasmid incompatibility determines whether multiple plasmids may be maintained within the same cell line through interference of the replication process (18). Natural restriction modification systems in bacteria, which act as a defense against the invasion of foreign DNA, can prevent the incorporation of mobile genetic elements introduced into a bacterial cell (81). Therefore, to epidemiologically define strains, it is necessary to account for vertically and horizontally acquired genetic changes in final phylogenetic estimations.

To track drug resistance and virulence markers associated with disease, it is important to find clear and consistent methods for defining common strains within a single species. Through WGS, we can compare entire genomes between bacterial isolates and identify when virulence and resistance likely arose. However, there is no singular definition of a "strain" recognized by microbiologists, and the term is often used interchangeably with other terms such as "lineage" and "clone." Some definitions rely on known ecological and epidemiological association with specific genetic markers, while others are agnostic to the epidemiology and utilize qualitative cutoffs of genetic similarity based on the whole genome or on marker genes. As a species, *S. aureus* forms larger clonal clades with distinct evolutionary histories and similar allelic profiles of core genes, or genes shared by at least 95% of representative isolates (81, 82). It also has a repertoire of non-core genes, some of which are associated with horizontal gene transfer, that are important for antimicrobial resistance and for some of the virulence traits (82, 83). This divide between core and non-core genes contributes to the derivation of subspecies groups (strains) based on various levels of genetic similarity and the presence or absence of genes. Common groupings that exist include clonal complex (CC) and sequence type (ST) which use a gene-by-gene approach to differentiate lineages and group into non-overlapping categories based on central genotypes (84). Other typing mechanisms include categorizing at the individual trait level with structural stability and epidemiological linkage, such as *agr* type (85) SCC*mec* type (for *mecA* positive *S. aureus*) (86), and staphylococcus protein A (*spa*) (87). With the advent of whole-genome sequencing, there is an increasing opportunity to use the entire set of genomic information, including core genes and non-core, or "accessory" genes, to group individual isolates into unique lineages.

Summary of Chapters

The goal of this thesis is to broaden the scope of evolutionary concepts utilized for applied epidemiological practice and identify the strengths and limitations of genomic approaches when paired with clinical and epidemiological metadata. In applied epidemiology, practitioners still face challenges in justifying the use of genomics for improved outbreak prevention or clinical care, developing systems for processing large amounts of data, ensuring effective communication between laboratory staff and epidemiologists regarding genomics, and discerning transmission based on genomic thresholds. In this thesis I aim to tackle some of these challenges directly: first, by testing the value of specific genetic distance thresholds as markers for lineage relatedness within and between individuals. Second, by describing and testing techniques in both metagenomics and genomics that can lead to improved detection of genetic relatedness.

In Chapter 2, I review the use of single nucleotide polymorphism (SNP) thresholds in hospital-associated spread of two major pathogens, *S. aureus* and *Pseudomonas aeruginosa*. I describe important evolutionary and ecological contributors that can lead to genomic differences present when two or more isolates are compared. Simultaneously, I highlight how these forces may differ in their effect on genetic difference depending on the species being investigated in an outbreak, even when the epidemiological setting remains the same. I argue that SNP thresholds should be species-specific and refined into SNP threshold ranges to better guide outbreak investigations. I plan to modify this chapter and submit it for peer review.

In Chapter 3, I investigate the presence of close genomic relationships between patients experiencing infections caused by methicillin-resistant *S. aureus*. I evaluated the impact of genomic alignment tools and genetic background of the infection strains on the detection of SNP differences and subsequent putative transmission clusters, followed by classification of those clusters at different thresholds. I identify potential risk factors for clustering, including recent hospital overlaps, and offer a logistic analysis to relate SNP distance and likelihood of detecting a hospital overlap as a tool for future cluster investigations of MRSA in hospital settings. This chapter was published in the *Journal of Clinical Infectious Diseases* in 2022, entitled “Unsuspected Clonal Spread of Methicillin-

Resistant Staphylococcus aureus Causing Bloodstream Infections in Hospitalized Adults Detected Using Whole Genome Sequencing."

In Chapter 4, I expand upon my research from Chapter 3 by investigating risk factors for recurrence of MRSA bacteremia as well as investigate whether recurrent strains share convergent adaptive traits. I describe the phylogenetic and clinical diversity between strains that do and do not cause subsequent infections in patients with bacteremia. I further use SNP distance and phylogenetic topology to differentiate persistent lineages associated with a host from genetically new infections for the same individual. I show that most recurrent infections are from relapsing strains, and that these strains share demographic, molecular, and clinical characteristics associated with recurrence as seen in the overall body of work. I demonstrate that relapse isolates have a signature of positive selection compared to the overall population of MRSA isolated from bloodstream infections, and that common genes among these relapse lineages occur in antibiotic resistance and virulence-associated genes. This work will be submitted as a unique publishable unit upon further revision and review.

In Chapter 5, I examine how antimicrobial resistance profiles can be compared in a larger ecological network using metagenomics. I first characterize the bacterial species abundance and antimicrobial gene abundance profiles between gut microbiomes of human residents and sympatric gray mouse lemurs living near Ranomafana National Park in Madagascar. I show that human communities have indistinguishable abundance profiles but are significantly distinct from lemur gut microbiomes. I then identify gene presence overlaps and compare the nucleotide sequence similarities between genes detected in both groups and their surrounding genetic context. I identify shared antimicrobial resistance genes with highly conserved nucleotide sequences, several of which were evidently a part of a larger cassette that is likely associated with horizontal gene transfer. This chapter was published in *PeerJ* in 2024, entitled "Metagenome-wide characterization of shared antimicrobial resistance genes in sympatric people and lemurs in rural Madagascar."

In Chapter 6, I conclude this thesis with a summary of the work thus far and suggest future directions for study. I outline some of the benefits and ongoing challenges in using sequencing to understand transmission risk factors and within-host evolution. I further suggest potential analyses to explore within-host adaptations of staphylococcus aureus and how that can contribute to different disease states in colonized patients. For example, building upon a hypothesis that within-host adaptation may lead to a virulence versus transmission trade off, I suggest looking more closely at phenotypic changes relative to virulence profiles in addition to antimicrobial resistance profiles. I also plan to conduct additional analyses using boosted regression trees to assist with predicting relapse as an outcome.

Chapter 2: What's in a SNP?: Deducing transmission events of bacterial infections using genetic thresholds of relatedness

Brooke M. Talbot and Timothy D. Read

Abstract

Whole genome sequencing (WGS) is now the preferred molecular typing method for applied epidemiological surveillance of bacteria causing infectious diseases due to the technology's ability to highly resolve pathogen genomes. Outbreak and transmission investigations using WGS in the last decade have successfully detected clusters of related illnesses, ruled out unrelated cases from investigations, and identified important risk factors causing disease spread. For these investigations, practitioners typically use a "SNP threshold," a measure of single nucleic acid base pair differences between infection isolates, as a tool for deciding if two infections are closely related enough to signify a recent transmission event. However, the justification for these thresholds varies across investigations as well as the considered ecological and evolutionary processes acting uniquely on bacteria for determining the relationship between genetic change and transmission source. By comparing how these processes are understood and handled across outbreak investigations for two model organisms that plague healthcare settings, *Staphylococcus aureus* and *Pseudomonas aeruginosa,* I demonstrate that to best define SNP thresholds that are most useful for solving bacterial outbreaks, investigators should ensure that their WGS investigation workflows account for the contribution of the biological effects from cellular processes and population among sampled strains of bacteria.

Introduction

Of increasing interest is the use of WGS **surveillance** in hospital and healthcare settings, where the introduction and spread of any disease into these spaces has serious consequences for patients receiving healthcare. Hospital-associated pathogens make for a great case of how an epidemiologist can think through the different biological components of bacterial investigations using WGS because 1. Hospital environments are known to have outbreaks with multiple modes of transmission, 2. A high-risk population benefits from time sensitive and highly granular investigations, 3. Bacterial pathogens in these environments have high morbidity for patients, and 4. There is currently no widespread surveillance system using WGS uniformly across all hospital and healthcare settings.

There are inherent challenges for fully implementing WGS into routine surveillance, with special concern for tracking bacterial infections. For bacterial species, genetics change through both **horizontal** and **vertical gene transfer** and variation in the environment can trigger different stress responses and adaptation among different populations of the same species, the impact of which muddies the generalizability of too fine grain a genetic match. One of the most executed strategies for bacterial disease transmission is reporting single nucleotide polymorphisms (SNPs), where there is a single base pair change at specific **loci**. This metric is primarily modelled on the idea that, under a neutrally evolving population, more SNPs are equivalent to more time passing since the common ancestor of two isolates. Consequently, epidemiologists have taken to reporting some "SNP threshold" when using WGS as a part of an outbreak investigation. A SNP threshold is the measured number of base pair differences between two isolates. The terms single nucleotide variant (SNV) threshold and SNP threshold have been used interchangeably throughout bacterial outbreak reports, though SNV more accurately encompasses general variation of a single base of a nucleic acid, while SNPs refer to specific single base changes at a locus. Most reports do not disentangle these definitions, and largely instead report SNP differences as the total number of base pair variations across all loci between two isolates. For this review I will use SNP threshold as it is the more commonly expressed term across investigation reports.

To establish these thresholds, it is critical that infectious disease scientists and epidemiologists invoke basic understanding of the evolutionary context of the molecular markers that they are relying on for disease prevention. In this review, I will discuss the important genomic, evolutionary, and epidemiological factors that impact the detection, comparison, and epidemiological patterns of present single nucleotide polymorphisms (SNPs) in bacterial infections which have been addressed to varying degrees across different bacterial outbreak reports. The ecological and evolutionary concerns of investigators can broadly be broken into those at the individual bacterial cell level, random mutations, homologous recombination, and adaptation to the environment; and at the bacterial population level, intrahost demography and interhost and environmental **richness**. To exemplify how these concepts impact **phylogenetically** distinct pathogenic bacteria that are operationally considered in the same way for surveillance, I compared and characterized SNP ranges documented and summarized in outbreak investigations utilizing WGS for two model bacterial pathogens, *Staphylococcus aureus* and *Pseudomonas aeruginosa*, associated with healthcare-associated outbreaks.

Ecology, Epidemiology, Evolution, oh my!

Though WGS has often been used to confirm outbreaks initially detected by another sentinel event (i.e. a sudden increase in culture positive cases of illness in a short span of time), WGS and metagenomic sequencing will inevitably become the standard of practice in prospective surveillance of bacterial outbreaks. It is therefore important to explore the current state of outbreak investigations that have used WGS and help define SNP threshold ranges that can be used in a variety of outbreak settings and transmission events. The sheer volume of outbreak investigations using WGS for various bacterial pathogens in numerous settings provides an excellent opportunity to identify empirical evidence of current pairwise SNP differences and define the criteria for effective SNP thresholds or threshold ranges. Early implementation of WGS into hospital surveillance was reactive to ongoing outbreaks and used to confirm or rule out cases (88). These early investigations were important for defining early ranges of pairwise SNP differences between isolates when there was a clear epidemiological link, which was the case with tracking cases related to a single-introduction of a MRSA infection into a neonatal intensive care unit (89). However, investigators moved quickly from this reactive and confirmatory practice with WGS toward testing the utility SNP thresholds to detect outbreaks prospectively in the hospital and provide swift intervention. This was demonstrated in 2012 with investigations of *Clostridium difficile* and methicillin-resistant *S. aureus*, where, the combination of prospective sequencing and regular infection control identification of outbreaks was able to rapidly ruled in or out patients from of a cluster on the basis of pairwise SNP differences between patient isolates (90). Contemporarily, investigations now tend toward using WGS in a prospective manner, where pairwise SNP differences are the principal sentinel event for cluster detection. The sensitivity of a SNP threshold is quite important for this wave toward prospective surveillance of clusters. Some investigators have implemented predictive models to come up with a species-specific SNP threshold. For example, Coll et al. suggests *S. aureus* isolates be included in a cluster at a threshold of <=15 SNPs within six months of isolation (79). Other Investigators take a broad approach and define a single threshold value for multiple hospital pathogens from the available case reports, such as Sundermann et al.'s definition of 15 SNPs as a cut-off for 14 different species of bacteria (7).

Justifications for SNP thresholds between bacterial infections

Suggestions for optimal workflow strategies that integrate WGS and bioinformatics into outbreak investigations have been reviewed elsewhere(91–93). These reviews highlight that the choice of a SNP threshold must be contextualized in the availability and expertise of the investigation team in terms of bioinformatics, epidemiology, microbiological technique, and evolutionary biology. Briefly, the first component to determine an appropriate SNP threshold is to ensure that workflows include suitable sampling to ensure infecting isolates are from pure colonies, adequate sequencing depth and quality, and appropriate variant calling software. For this review, I will not focus on these technical bioinformatic considerations but instead focus on underlying biological processes fundamental to pathogen evolution and ecology that drive the presence of SNPs in the bacterial genome. Ultimately, SNP thresholds demonstrate some expected difference that investigators have about the relationship between two or more sampled isolates in the temporal period of interest. Investigations discuss important genetic and ecological principles in pieces that influenced the choice of the threshold. The most common patterns that emerge for justifying SNP thresholds occur at multiple stages related both to natural processes that result in an accumulation of SNPs and the population structure and **niche** of a pathogen. At the individual bacterial cell level mutations accumulate over time through random processes at a given rate. Selection on these mutations can remove deleterious mutation from the population, but less deleterious or neutral mutations may remain or be removed more gradually over time. This accumulation of differences can be profoundly adjusted by homologous recombination and cell adaptation to the environment via **positive** and **purifying selection**. At the broader population level, intrahost demography and interhost and environmental richness also influence the number of SNPs ultimately detected once isolates are compared to one another (Fig. 2.1). Comparing how these processes are understood and handled across outbreak investigations will help better define their relative contribution to SNP thresholds, and which are most useful, stable, and for solving outbreaks and assist in the adaptation of future workflows as more is understood about pathogen biology.

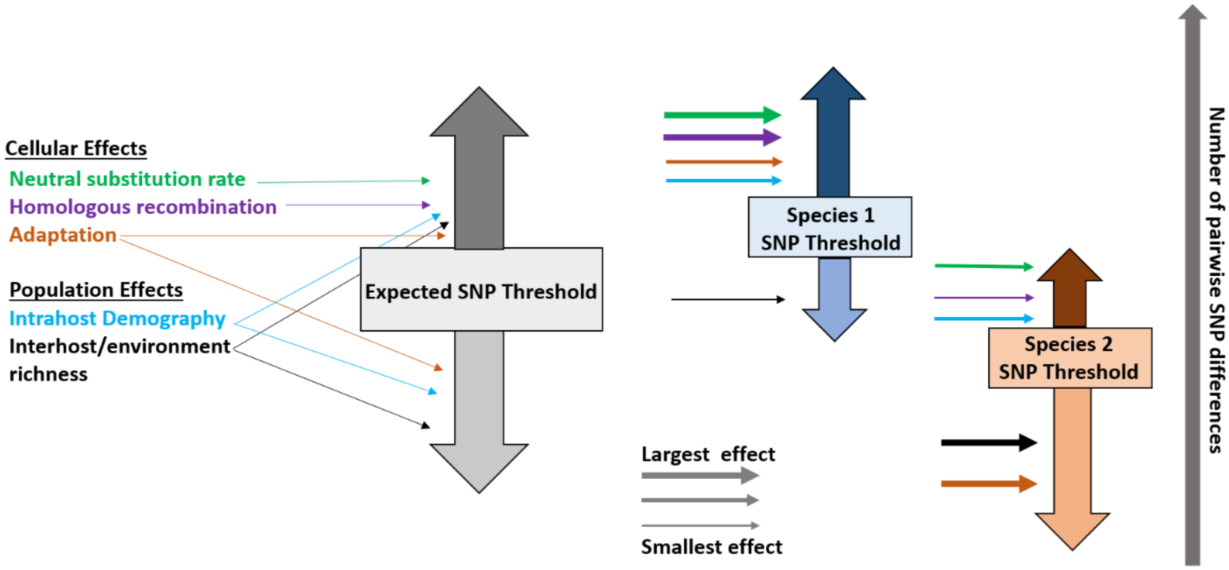

Figure 2.1. Evolutionary and ecological processes leading to genetic difference accumulation between bacterial core genomes. After a transmission event occurs, genetic differences accumulate in the core genome over time causing divergence between two related isolates. The expected SNP threshold between related bacterial isolates in a given time frame can increase or decrease depending on the weighted effect of cellular processes and population diversity in the sample. An example between two hypothetical bacterial species shows that the expected SNP threshold can differ from one another depending on the impact of each biological effect within the system, indicated by line weight and color.

The traditional characteristic relationships investigated for transmission events and outbreak relationships are those among the person, environment, and infecting agent in a given course of time(11). Describing these relationships with a simple metric like a SNP threshold can help decide if that detected diversity relates to an appreciable time and place. Investigators debate whether SNP thresholds should be used at all for inferring transmission events and outbreaks. Arguments against a single threshold include that their context varies between studies, that they may not necessarily imply transmission probability, and that different genetic lineages within the same species vary in **substitution rate** (94,95). The advantage of thresholds, however, is that they allow investigators to make quick, simple, and consistent decisions to prioritize which patients to investigate for possible exposures and transmission events such as in hospital infection prevention or contact tracing. Consequently, thresholds are readily documented in the literature for both prospective and retrospective outbreak analyses and have been used successfully to identify transmission.

Use Cases: Interpreting thresholds across healthcare-associated pathogens

Important considerations to define for any disease under surveillance are the expected modes of transmission (eg direct or indirect) and the expected source or environment that resulted in the disease. Healthcare settings are a high priority for preventing infection transmission, as these pose a major threat to patient safety and are worldwide the most common adverse event associated with healthcare (96). These infections can be prevented with good surveillance and infection control practices. WGS surveillance in these settings offers high returns for understanding disease spread and promoting patient welfare. From a clinical and public health perspective, molecular surveillance stands to massively benefit local communities served by a single healthcare system, as well as the wider community.

Globally, two pathogens of great concern in healthcare settings are Gram positive *S. aureus* and Gram negative *P. aeruginosa* (97) because of their ability to harbor antibiotic resistant phenotypes and cause persistent and recurrent infections in hospitalized patients and in patients with comorbidities. Healthcare settings utilizing SNP-based thresholds for outbreak investigations demonstrate that infections of *S. aureus* and *P. aeruginosa* with epidemiological links to other patients or environmental samples cluster tightly together under 20 SNPs difference (Fig. 2.2).

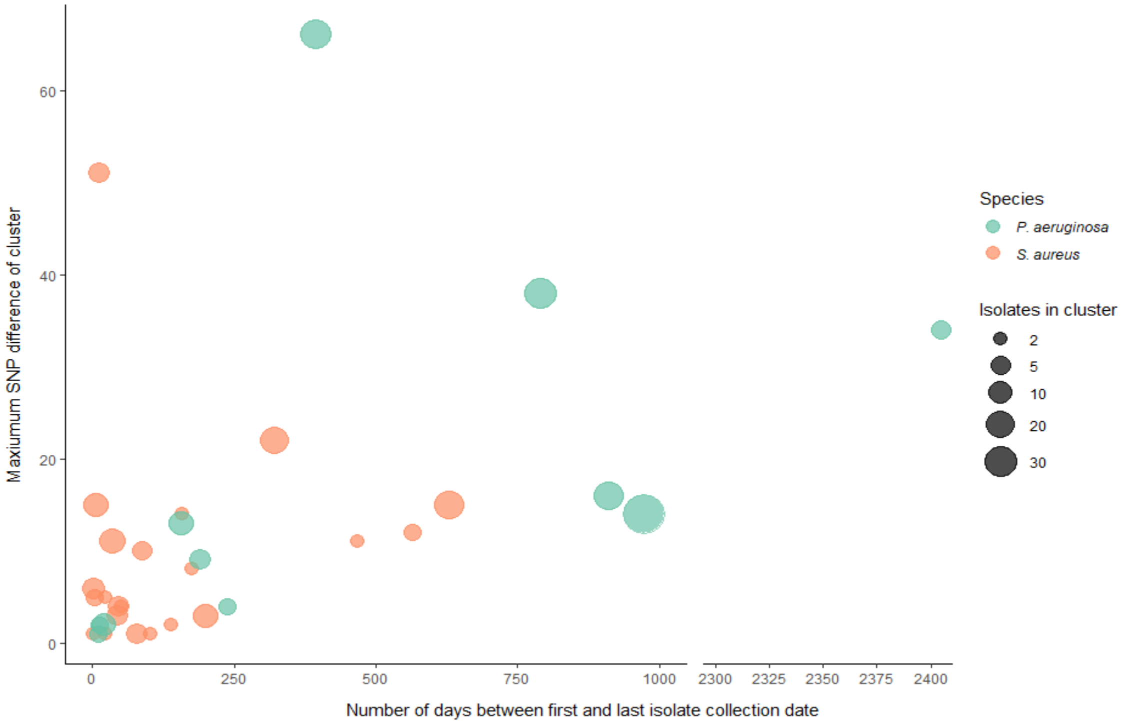

Figure 2.2. Epidemiologically investigated clusters of healthcare-associated *S. aureus* (n=22) and *P. aeruginosa* (n=14) investigated with whole genome sequencing. Cluster isolates constitute clinical isolates and/or environmental isolates. Maximum pairwise single nucleotide polymorphism (SNP) difference, number of isolates in a cluster, duration between collection dates were gathered from main and supplementary figures, text, and tables from 14 outbreak investigations (Table S1). Clusters were excluded from visualization if there was no defined time period or if pairwise SNP distances could not be identified. Epidemiological linkages were defined in the reports and include hospital overlaps between patients, common exposures to equipment or care personnel (cross-transmission), and exposure to contaminated environments.

Most practitioners applying thresholds must account for all pathogens of interest during prospective surveillance, including *S. aureus* and *P. aeruginosa*, and therefore additional simplicity has been used for surveillance by setting the same threshold for multiple species (7). This should be met with some caution and considerable review given that evolutionary and ecological processes can affect bacterial species differently. Particularly, *P. aeruginosa* and *S. aureus* have very distinct genetic characteristics than result in different estimations of the relationship between the pathogen's genetics and the time between onward transmission and human detection from a recipient patient (Table 2.1).

Table 2.1. Taxonomic, genomic, and ecological characteristics of *Staphylococcus aureus* and *Pseudomonas aeruginosa* causing human infections

	Staphylococcus aureus	***Pseudomonas aeruginosa***
Taxonomic group	Gram positive cocci	Gram negative bacilli
Genome size (Mbp)	Average: ~2.8 (~2.69 to ~2.96) (98)	6.34 to 7.15 (99)
Estimated substitution rate (SNPs/year)	*Non-hypermutator* 3.5 (89) 5.8 (53,100) 8.7 (101)	*Non-hypermutator* 1.0 (102) ~1.2 (103) 2.5 (15) 5.5 (104) *Hypermutator* 50 [60]
Isolate sources/environments	Human nose (77), skin and soft tissue infections (105), hospital equipment (106–108)	hospital water sources (103,109,110), human lung (15,104,111)

The unique niches of *S. aureus* and *P. aeruginosa* contribute to the investigation priorities of documented outbreaks and likely contribute to differences in the genomic characteristics of individual clusters. *S. aureus* typically colonizes human hosts and becomes an **opportunistic infection**, with sequelae including skin and soft tissue infections, bloodstream infections, pneumonia, and bone infections (63). *P. aeruginosa* is a commensal organism and can cause acute (112,113) and chronic opportunistic infections (111), but it is also a known contaminant of water sources in healthcare settings (34,109,113). Previous outbreaks demonstrate years-long persistence of a single *P. aeruginosa* clone can persist and spread in a healthcare setting due to the environment (16,34,112–115), though acute are also detectable with WGS (106,112,116). In contrast, *S. aureus* is frequently reported in person to person spread, or shared equipment or healthcare workers, where the epidemiological link is overlapping hospital stays between patients (88,89,106,107,117). However, *S. aureus* can also contaminate the environment and result in short-term outbreaks with few (≤3) involved patients (52,108)[31,32], and some larger and more prolonged outbreaks (90,108). These distinct transmission situations are ultimately reflected in the SNP difference profile of reported outbreaks, where smaller, more rapid, and very genetically similar isolates are seen for *S. aureus*, and larger and more genetically diffuse clusters are documented for these long-term *P. aeruginosa* outbreaks (Fig. 2). In order to assess how expected SNP thresholds can be readily tailored to each species, investigators must therefore gain a good understanding of the evolution and ecology of each pathogen of interest. This starts with considering how much cellular processes and population sampling processes influence the expected detectable SNP range.

Random mutations and substitution rate

The accumulation of random mutations at a predictable rate over time is foundational to the rationale of SNP thresholds. Binary fission, the asexual reproduction mechanism of bacteria, produces two identical daughter clones, and it is this clonality which underpins the detection of related infections. However, the clock starts ticking upon this split. Under the neutral mutation theory, synonymous SNPs have no effect on organismal fitness and accumulate in a pathogen population through random processes (118). Though this idea is now more often regarded as a null hypothesis in most evolutionary explorations, it is nevertheless an important assumption for modeling relatedness over time between prokaryotic isolates. Using these assumptions, investigators can build a molecular clock that demonstrates a linear relationship between base substitutions and time, in order to predict the time at which last common ancestor emerged between two or more sampled organisms. In the context of an outbreak, by following this molecular clock, investigators hope to find epidemiological links that can occur within the predicted time frame.

A priori knowledge of the per site and per genome substitution rate for some pathogens is available in literature, which investigators have made use of during outbreak investigations in order to identify clusters of significance (110). However, even for the same pathogen, estimates of the substitution rate relative to the reference genome are vary between investigations, leading to different expectations of the SNP differences over a time period of interest. For example, *S. aureus* substitution rate estimates include one SNP every 15 weeks (89), one SNP every nine weeks (53,100), and one SNP every six weeks (101). The diversity in values may arise from improved estimation techniques over time, size of the dataset used to calculate the estimates, and the size of the reference core genome. However, these estimates could also reflect some critical biological differences, including strain-specific substitution rates or the specific ecology of the transmission event, such as the source coming from long-term host carriage. *P. aeruginosa*, in comparison, has a lower reported mutation rate than *S. aureus*, which ranges from 1 SNP a year to 1 SNP every nine weeks (Table 1). When considering a common threshold for multiple pathogens, it is important to factor in how divergent the expected substitution rates will be between species in the time period of interest, as drastically different rates could result in very different estimates the longer the investigation period proceeds.

Though not frequently documented as a concern in transmission cluster investigations, hypermutation in some strains can nevertheless contribute to irregularities in the estimated substitution rate that would determine an expected SNP threshold. In one documented methicillin-resistant *S. aureus* (MRSA) outbreak, one isolate with a confirmed mutation in *mutS,* a gene regulating DNA repair mechanisms, was identified as connected to a cluster of patients and had a substitution rate 6.9 times greater than the next-related isolate and many unique SNPs (88). Had this isolate been a part of prospective surveillance where the substitution rate was assumed to be the same across isolates, this isolate likely would have been excluded despite ultimately having an epidemiological connection to other isolates with a tighter connection. Comparatively, *P. aeruginosa* has known documentation of hypermutator phenotypes, particularly among chronically infected cystic fibrosis patients and less commonly among acutely infected patients (119). However, *P. aeruginosa* isolates have large variation in reported mutation phenotype when derived from clinical isolates of cystic fibrosis patients or from the environment (120). Future investigations should consider iterative WGS analyses as epidemiological links are identified between closely related cases as a method of exhaustive case finding to account for the possibility that other patients could be involved. Curating and tracking mutations associated with hypermutation and screening isolates for these changes as a part of surveillance practices would be a wise early step in determining how to best use substitution rate to estimate a SNP cutoff.

Homologous recombination

The main goal of phylogenetics in transmission and outbreak investigations is to identify precise relationships between closely related isolates through vertical gene transmission. Novel genetic elements are also introduced into bacteria populations through horizontal gene transfer, which includes the processes of conjugation, transduction, and transformation. These introduced genes, such as plasmids, transposons, and phage DNA are mobile genetic elements and form the accessory genome. Using the whole genome, and therefore all available genes have been demonstrated to lead to the same epidemiological conclusions. Comparison of core- and whole-genome MLST and SNP-based analyses for the same set of *P. aeruginosa* outbreak isolates showed that the methods are comparable (115). Limitations to the comparison arise, however. First, the **alleles** in the accessory genome undergo evolutionary pressures in the same way as alleles in the core, but not necessarily in concordance with the number or rate of changes among core genes. Different bacterial species also have different sized accessory genomes and different compositions of mobile genetic elements. Predictability of vertical transmission becomes difficult when the whole genome is considered in an investigation because of the possible different evolutionary trajectories of core and accessory genes and because the contribution of the accessory is not standard across species [51]. Second, the rate of genetic recombination is variable between species, which can influence gene-by-gene comparisons if some genes undergo a more rapid rate of change over time than others. Therefore, most transmission investigations focus on core genome analysis, which identifies conserved genes across a species without the contribution of elements of the accessory genome. This ensures that the observed genetic SNPs are detected among the most conserved alleles within the population.

Variation may also be introduced into components of the whole genome through homologous recombination, in bacteria a process of gene conversion, where donor DNA sequence replaces recipient sequences, both in core and accessory genes (121). However, homologous recombination also affects bacterial species differentially. The percentage of recombined genome in species significantly differs across bacteria, with intracellular pathogens exhibiting the least variation and opportunistic pathogens exhibiting the most. Considerable differences can be observed between species even occupying similar niches though (e.g., *S. aureus* having comparably low recombination to the intracellular pathogen *Neisseria meningitidis*) (122). Accounting for recombination is not uniform across outbreak investigations. Access and understanding of appropriate bioinformatic tools to handle this concern is still beginning to enter into common use among investigators of bacterial infections, but workflows can account for recombination during the variant identification phase to filter out regions of concern (112). Freely available tools to identify and mask recombinant regions, such as Gubbins (123), ClonalFrameML (124), and fastgear (125) can also be incorporated into investigations (107,113,117). Even with available tools, the practice of excluding recombinant regions is still not adopted in all workflows, and this could contribute to the variation seen in different thresholds for similar investigation periods.

Regardless of tool access, though, the choice not to measure SNPs in recombinant regions ultimately removes genetic diversity from the sequence comparisons, and this choice may come down to the species of bacteria under consideration. For example, Eyre et al. chose to remove known mobile genetic elements from their investigation of *Clostridium difficile* clusters because 11% of the genome is affected by these elements. In their same prospective analysis, which included MRSA isolates, mobile genetic elements were not removed, though SNPs were identified only in non-repetitive sites on the genome for both pathogens (90). However, in a *P. aeruginosa* hospital outbreak, removal of these regions from the analysis still allowed investigators to differentiate clades for distinct transmission events despite a decrease in comparable core genes (113). Since the effects of recombination can affect detectable variation at the level of a SNP, it is important to at least document consideration of these methods in a public health workflow for the sake of comparability across studies.

Adaptation to the host and environment

Neutral mutations cannot explain all accumulated SNPs. Rare adaptive non-synonymous mutations, and alleles undergoing genetic **hitchhiking** associated with these selected SNPs, change the genetic diversity of bacterial populations as they adapt to changes in their environment (126). Adaptation to the host or to an environment can have variable effects on the overall number of differences detected. Adaptations reflect the function of random mutations to the genome and their presence and persistence in a population. Cells carrying an advantageous adaptation may increase in numbers because of their increased fitness compared to cells without the adaptation. This is dependent on the diversifying effects that various niches in the healthcare environment have on pathogen adaptation, and consequently the persistence of genetic lineages associated with different niches. These niches exist within the host where bacteria interact with immune cells and also exhibit variable tropisms to host tissues. They are also part of the external environment, such as the water systems, machinery and equipment, and surfaces subjected to regular antibiotic treatment. Depending on the organism, a pathogen may exhibit adaptations that allow it to move between a host and the environment, or its niche range may be narrow. When a pathogen infects a host, it is faced with an array of complex selective pressures, particularly adapting to the host immune response (77). An infecting strain of bacteria must also contend with other present bacteria commensal to the host in order to establish infections. Together these components of the host environment act as different selective pressures on individual cells which contribute to within-host evolution. Particularly important among hospitalized patients and those with chronic infections, the presence of antibiotics can also act as a selective pressure as well as cause genetic bottlenecks that drastically decrease the population size and available genetic variation (111).

S. aureus causing infections tend to occupy host-associated niches, though the bacteria can be isolated from a variety of infection-relevant tissues like whole blood, stool, lung aspirate, and wounds. Adaptation and long-term residency on a host result in detectable variation between bacteria. However, Young et al. showed that intrahost nasal and bloodstream *S. aureus* isolates differed by eight SNPs (127), and another investigation, observed no more than 10 SNPs between isolate pairs from the same patient, leading to a defined threshold of ≤15 SNPs for identifying transmission pathways over a seven-month investigation period of a neonatal intensive care unit (108,128). In contrast, evidence in the literature also shows that the diversity and within-host evolution of *S. aureus* samples even from the same body site could be as high as 40 SNPs (129). Assessment of transmission may also be clouded by patient exposure to multiple phylogenetic lineages whose most recent common ancestor would be outside of the relevant time scale for transmission investigation. Two scenarios emerge: Few SNPs occur even with distinct within-host environmental pressures, or many SNPs occur when the environment stays the same. For both scenarios, different evolutionary forces may be acting with different magnitude: in the former example, adaptation and selection may be decreasing genetic diversity, and in the latter these differences may be accumulating because of random processes resulting in nonsynonymous SNPs.

In comparison, *P. aeruginosa* is known to be both a free-living organism, as well as live commensally within a human host (Table 1). Therefore, there is potential for *P. aeruginosa* to transmit both from person to person or from a contaminated environment shared by multiple patients. This is of particular concern for patients who maintain long residency in a healthcare setting and repeatedly change between care units, which was observed in a patient infected with two lineages of *P. aeruginosa* that shared the same strain type (103). In this case, the genetic differences seen between multiple samples could reflect multiple transmission events.

Microevolution of *P. aeruginosa* is well documented among chronically infected patients, and it leads to specific adaptations for pathogenesis. Even on a time scale as short as a year, *P. aeruginosa* infections gave rise to multiple clonal emergences with specific adaptations that are unique to different parts of the patient's lungs (111). Interestingly, evidence of transmission from hospital water sources to a patient with cystic fibrosis showed that over a 60 day period isolates of *P. aeruginosa* between the two sources differed by about two SNPs (110). This observation is interesting given the expected substitution rate of *P. aeruginosa* to be about one SNP per year. From the perspective of a practitioner, though, this contrast can cast some insight on the utility of a threshold for *P. aeruginosa*. If separate clusters of *P. aeruginosa* differ by two SNPs in a two-year period and a two-month period, Likelihood of successfully linking other cases decreases if the threshold is generalized and does not consider different sources for each cluster. By not accounting for this adaptive behavior, longer-term investigations might detect more SNPs than would fall within a SNP threshold, and therefore miss transmission events.

Intrahost demography of bacterial infections

Considering neutral mutation, a colonizing bacterial population that started from the introduction of a single cell may acquire increasing intra-population diversity, (known as a cloud of diversity). This demography becomes important when tracing transmission events as the genetics of the truly transmitted cell may differ by a certain number of SNPs from both the sampled isolate of the donor host (or donor environment) and the recipient host. Thus, the difference between historical infection isolate, infection host, and recipient host all have the potential to start with a non-zero value of SNP differences from one another depending on how close in time these samples are collected to the time of infection. The consequence of this "non-zero" value means that we might under- or over-estimate the expected number of SNPs different between two infections, especially if the source has been infected for a long enough period of time that there is already significant genetic variation within the donor pathogen population. Similarly, the genetic diversity ultimately observed in transmission investigations depends on the number of samples collected and assessed from a patient, any of which could harbor mutations reflective of its response to the host and the structure of the population from which it is derived. For routine surveillance and diagnostics in a hospital, it is most common to observe in outbreak reports that a single, representative index isolate per patient is assessed in the final data set, thus all available diversity is not captured.

Modeling approaches can help define a range of diversity to expect in an outbreak, which can account for the accumulation of genetic differences over time by combining data regarding interhost genetic diversity, intrahost genetic diversity, and epidemiological links to predict a reasonable threshold. Coll et al. demonstrated that a core genome SNP threshold for *S. aureus* that predicts transmission events among asymptomatic and symptomatic patients within a six-month period is between 11-13 SNPs by testing multiple different models to predict *S. aureus* genomic differences over time, which included empirical data from infections between different hosts and isolates over time isolated from the same host (79). Their model was generalizable to unrelated populations of isolates than the original test cohorts and showed a similar result of 14 core genome SNPs across all isolates and 11 core genome SNPs from infecting isolates. However, this modeling approach was based on the assumption of person-to-person transmission events, and therefore it may not necessarily apply to other pathogens with a different mode of transmissions, such as through water. Development of different models and evaluation of Coll's model and others in the literature to new scenarios would be useful nonetheless to assess their generalizability.

Interhost and environmental richness

To set up prospective genomic surveillance of a hospital, many isolates will have to be sequenced, and the majority of these isolates will be genetically distinct from one another and not related to a cluster. How then should investigators compare the highly genetically diverse pool of samples in real time and still capture all of the true SNPs? Investigators are usually interested in establishing the phylogenetic relationships of all isolates in a period of interest and test hypotheses for the relationships between genetic distance and isolation source or environment. Investigators create core genome alignments that compare common alleles between isolates and a selected representative reference, or in absence of a reference between all sequences in a sample set (7). Core genome analysis may slightly reduce accuracy and may possibly reduce resolution of lineage specific events (130). To better detect these more nuanced genetic relationships, investigators have first identified isolates that form broad but distinct phylogenetic clades (110) or that belong to the same known MLST identified in silico (108), and then take a stepwise approach to build phylogenies among more similar isolates. For example, in an attempt to understand *S. aureus* transmission between healthcare workers, patients, and the hospital environment, Popovich et al. chose first to align all investigated isolates to the same USA300 reference strain, the most prevalent background of methicillin-resistant *S. aureus* in the U.S., and then separately build USA300 or USA100 phylogenies with references or closed genome derived from the study set (52). In this way, investigators were able to maximize conserved genes within vertical lineages and consequently the number of observable variants. However, core genome analyses without further sub-setting of clade-specific references can still resolve epidemiological linkages. Harris et al. conducted an investigation of MRSA in a hospital special care baby unit where clustering isolates were on average over 500 SNPs different from the reference strain but still epidemiologically resolved, even revealing important connections between hospital and community transmission (89).

Strategies for *P. aeruginosa* are similar, but the overall values of SNPs are distinct from the previous examples of *S. aureus* transmission. In a prolonged outbreak of *P. aeruginosa* in a hospital ICU, Buhl et al. produced a polyclonal phylogenetic tree with the use of a single reference strain, PAO1, meaning that their detected cases were among distinct evolutionary clades rather than one lineage transmitting through the space. Even if this phylogenetic diversity reduced the detectable genetic differences between more closely related individuals, the investigators still identified unique epidemiological linkages within the two distinct detected clades (16). Other investigators chose to artificially diversify their phylogenetic investigations in order to create more evolutionary context for the outbreak. For an investigation of an outbreak of *P. aeruginosa* compared to other patient and environment samples in the same hospital, Parcell et al. included 15 additional reference samples in their phylogenetic reconstruction in order to resolve the evolutionary relationships (113). Addressing or adjusting diversity may change the absolute number of detectable differences, but it is apparent from these investigations that it might not always be necessary to find cases in a single cluster.

Perhaps the biggest concern instead becomes deciding what to do with cases at the edge of a threshold, especially if there is a precedent for gathering additional helpful exposure information to augment an investigation or to maximize detection of individuals still able to transmit disease.

For investigations using WGS to prospectively find clusters, investigators should be concerned about the contribution of genetic diversity in the pool of samples on the recreation of a core gene and core gene comparisons. As genetic diversity increases, the core genome shared among bacteria decreases, which in turn decreases the number of comparable loci. Given that prospective surveillance will sample from whatever is present, it is likely that most phenotypically similar samples will not be related and that datasets within a period of surveillance will be genetically diverse, exemplified by the diverse collection of bacterial isolates gathered under hospital surveillance by Ward et al. (106). To account for this inevitable diversity, some investigations implement sequence-by-sequence comparisons that estimate sequence similarity and group more highly similar alleles into gene families. This can be done with k-mer matching, followed by reference-based alignment (108). Other options include creating a core pangenome by identifying unique gene families among available sequences with Markov Cluster Algorithm (MCL) to find sequence distances (117,131). A gene-family approach may still decrease the detectable variation, but these variations remain more stable in alignments necessary for building phylogenies even with increased diversity compared to reference-based approaches (117). Using gene-family approaches are therefore advantageous in long-term surveillance, as it allows for flexibility in the sample pool over time and does not rely on a static reference that may become less relevant over time.

Since many isolates sampled from a setting will likely be unrelated, it is also important to consider a threshold that can successfully rule-out cases from investigation. Phenotypically similar isolates that are unrelated tend to have fairly distinct genetic profiles from unrelated isolates. For example, in a two-year *P. aeruginosa*, outbreak associated isolates differed from one another by 0-14 SNPs while a patient with no epidemiological link to the outbreak differed from the group of isolates by more than 100 SNPs (112). This distinct genomic distance hints that very distance genomic relationships can often be quick instances to separate cases from one another, and that energy and time should largely be spent on discerning edge cases that are closer to a cut off.

Conclusion

Transmission investigations have demonstrated that WGS is useful for forensic and public health purposes and can be done in an actionable time span for health interventions. These investigations highlight a transition from reactive and confirmatory use of WGS for outbreaks detected through other measures toward being the primary prospective surveillance tool. Public health practitioners now must go beyond the acceptability and feasibility of WGS. Incorporating a critical assessment of evolutionary and ecological processes within the context of each species is necessary to discern appropriate epidemiological links. From the observed biology of the richly documented outbreak reports, there are pathogen-specific patterns relating SNP differences and epidemiological linkage. Therefore, picking a target SNP threshold within the observed realm of SNP differences across investigations is a good starting strategy for building and refining outbreak detection systems. Additionally, anomalies that do not quite fit into a threshold can cue investigators to identify new and interesting patterns in disease, such as the role and emergence of hypermutators in disease spread or antibiotic resistance. However, SNP thresholds are tools that require understanding of underlying evolutionary biology and the pathogen's relationship to the surrounding environment. For investigators already implementing a standardized threshold, the best next steps would be to build in regular evaluation of the system to monitor the stability and sensitivity of the threshold, especially as disease prevalence in the community changes. Most fundamentally and importantly, thoughtful refinement of these thresholds using empirical data can guide us toward more rapid and complete prevention of further morbidity and mortality.

Though these investigations have demonstrated that we are likely to find epidemiological links from molecular surveillance, it is also likely that we will continue to find "closely related" clusters without detectable epidemiological or environmental relationships. New priorities and areas for further research must then become: 1. Sequencing and sampling equity (What are other epidemiological or environmental factors not currently measured within a study that should be considered? Do networks become better resolved when we widen the frequency of sampling? What institutional settings still lack access to sequencing, bioinformatic, and epidemiologic workforces?), 2. Ecological discovery (Are these results indicative of a new/unknown route of transmission?), 3. Evolutionary discovery (Do these pathogenic strains have new or unknown adaptations that could guide epidemiological surveillance? What are the phenotypic consequences of detected SNPs?), 4. Technological development (e.g., Does long-read sequencing provide additional/better resolution to WGS practices), 4. Refinement in the context of evolutionary principles (How do substitution rates compare between pathogen-only samples, commensal/asymptomatic, and environmental sampling?), 5. Non-outbreak health outcomes (How do disease prevalence and health outcomes change in a community when WGS is utilized?). WGS is simply the newest chapter in the history of molecular epidemiology, but by contextualizing the technology in evolutionary concepts we can understand disease transmission processes and behavior of pathogenic bacteria in exciting and unexplored ways.

Supplementary Tables

Table S2.1. Summary and citation sources of epidemiologically investigated clusters of *S. aureus* and *P. aeruginosa*

Cluster Citation	**Species**	**Days**	**Isolates**	**Detection Method**	**Source Environment**	**Suspected/Identified Source**	**Transmission Type**	Min SNP	Median SNP	Max SNP
Ward 2019 (106)	*P. aeruginosa*	12	3	Prospective	Healthcare	None	Unknown	1	1	2
Davis 2015 (116)	*P. aeruginosa*	21	11	Retrospective	Healthcare	Water source	Indirect: Common vehicle	0		2
Parcell 2017 (113)	*P. aeruginosa*	237	3	Prospective	Healthcare	Room environment	Indirect: Common vehicle	0	1.33	4
Buhl 2019 (16)	*P. aeruginosa*	394	27	Retrospective	Healthcare	Hospital environment	Indirect: Common vehicle	1	15	66
Sundermann 2021 (35)	*P. aeruginosa*	191	7	Prospective	Healthcare	Healthcare equipment	Indirect: Common vehicle	0	2	9
Magalhães 2020 (112)	*P. aeruginosa*	10	3	Prospective	Healthcare	sink trap	Indirect: Common vehicle	0	1	1
Magalhães 2020 (112)	*P. aeruginosa*	157	13	Prospective	Healthcare	ICU	Indirect: Common vehicle	0	–	13
Blanc 2020 (115)	*P. aeruginosa*	911	23	Retrospective	Healthcare	Burn unit environment	Indirect: Common vehicle	0	–	16
Snyder 2013 (114)	*P. aeruginosa*	2405	5	Retrospective	Healthcare	handwashing bin	Indirect: Common vehicle	5	11	34

Moloney 2020 (34)	*P. aeruginosa*	Unk	25	Prospective	Healthcare	washbasin u bend	Indirect: Common vehicle	0	–	8
Moloney 2020 (34)	*P. aeruginosa*	Unk	2	Prospective	Healthcare	washbasin u bend	Indirect: Common vehicle	5	5	5
Moloney 2020 (34)	*P. aeruginosa*	789	31	Prospective	Healthcare	washbasin u bend	Indirect: Common vehicle	0	–	38
Ward 2019 (106)	*P. aeruginosa*	6	13	Prospective	Healthcare	shared inpatient environment	Indirect: Common vehicle	2	7.5	15
Ward 2019 (106)	*S. aureus*	321	21	Prospective	Mixed	Intravenous drug use, unknown	Direct: Person to person	0	13	22
Ward 2019 (106)	*S. aureus*	4	3	Prospective	Healthcare	shared inpatient environment	Indirect: Common vehicle	4	5	5
Ward 2019 (106)	*S. aureus*	139	2	Prospective	Community	Intravenous drug use	Direct: Person to person	2	2	2
Ward 2019 (106)	*S. aureus*	47	6	Prospective	Community	Intravenous drug use	Direct: Person to person	0	3	4
Ward 2019 (106)	*S. aureus*	50	2	Prospective	Healthcare	shared inpatient environment	Indirect: Common vehicle	4	4	4
Ward 2019 (106)	*S. aureus*	23	2	Prospective	Healthcare	shared inpatient environment and hospital stay	Direct or indirect	5	5	5
Harris 2013 (89)	*S. aureus*	200	12	Retrospective	Healthcare	shared unit	Indirect: Common vehicle	1	–	3
Harris 2013 (89)	*S. aureus*	175	2	Prospective	Mixed	Mother to child	Direct: Person to person	8	8	8
Kristinsdottir 2019 (107)	*S. aureus*	35	16	Retrospective	Mixed	Parent and healthcare worker spread	Direct: Person to person	0	–	11

Kristinsdottir 2019 (107)	*S. aureus*	2	9	Retrospective	Community	Parent	Direct: Person to person	0	–	6
Berbel Caban 2020 (108)	*S. aureus*	159	2	Prospective	Unknown	Unknown	Unknown	14	14	14
Berbel Caban 2020 (108)	*S. aureus*	102	2	Prospective	Unknown	Unknown	Unknown	1	1	1
Berbel Caban 2020 (108)	*S. aureus*	2	2	Prospective	Healthcare	Shared ward overlap	Unknown	1	1	1
Berbel Caban 2020 (108)	*S. aureus*	24	2	Prospective	Healthcare	Shared ward overlap	Unknown	1	1	1
Berbel Caban 2020 (108)	*S. aureus*	467	2	Prospective	Unknown	Unknown	Unknown	11	11	11
Berbel Caban 2020 (108)	*S. aureus*	565	3	Prospective	Mixed	Unknown	Unknown	3	–	12
Berbel Caban 2020 (108)	*S. aureus*	628	24	Prospective	Healthcare	Shared Overlapping ward transmissions	Direct or indirect	1	–	15
Berbel Caban 2020 (108)	*S. aureus*	88	5	Prospective	Healthcare	Shared Overlapping ward transmissions	Direct or indirect	5	–	10
Köser 2012 (88)	*S. aureus*	12	7	Prospective	Healthcare	NICU carriage	Indirect: Common vehicle	1	–	51
Eyre 2012 (90)	*S. aureus*	45	7	Retrospective	Healthcare	Ward overlaps	Direct or indirect	1	–	3
Eyre 2012 (90)	*S. aureus*	79	6	Retrospective	Healthcare	Ward overlaps	Direct or indirect	0	–	1

Table S2.2. Glossary of terms

Term	**Definition**
Surveillance	Systematic collection of health-related data for the purpose of research, evaluation, and planning
Horizontal gene transfer	Movement of genetic material between organisms other than direct transfer from a parent to offspring
Vertical gene transfer	Movement of genetic material from a parent to offspring
Loci	Specific positions of genes within a genome
Richness	A measure of the number of different taxonomic units (eg species) within an ecological community
Alleles	Forms of a gene found at the same locus of a genome
Phylogenetics	The study of the evolutionary history between organisms
Niche	The impact of biotic and abiotic factors in a specific environment on an organism and that organism's interaction with those factors
Positive selection	Process of advantageous mutations arising and increasing in frequency in a population
Purifying selection	Process of the removal of disadvantageous mutations through selection
Substitution rate	The number of new mutations that arise in each generation multiplied by the probability that these mutations become fixed in a population
Opportunistic infection	An infection caused by an organism that non-pathogenic when under typical host-organism interactions

Hitchhiking	Mutations or genes which are not directly under selection but fix in a population due to close proximity to a gene undergoing selection

Chapter 3: Unsuspected Clonal Spread of Methicillin-Resistant *Staphylococcus aureus* Causing Bloodstream Infections in Hospitalized Adults Detected Using Whole Genome Sequencing

Reprinted material from: Talbot, B. M., Jacko, N. F., Petit, R. A., Pegues, D. A., Shumaker, M. J., Read, T. D., & David, M. Z. (2022). Unsuspected Clonal Spread of Methicillin-Resistant Staphylococcus aureus Causing Bloodstream Infections in Hospitalized Adults Detected Using Whole Genome Sequencing. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, *75*(12), 2104–2112. https://doi.org/10.1093/cid/ciac339

Abstract

Background: Though detection of transmission clusters of methicillin-resistant *Staphylococcus aureus* (MRSA) infections is a priority for infection control personnel in hospitals, the transmission dynamics of MRSA among hospitalized patients with bloodstream infections (BSIs) has not been thoroughly studied. Whole genome sequencing (WGS) of MRSA isolates for surveillance is valuable for detecting outbreaks in hospitals, but the bioinformatic approaches used are diverse and difficult to compare.

Methods: We combined short-read WGS with genotypic, phenotypic, and epidemiological characteristics of 106 MRSA BSI isolates collected for routine microbiological diagnosis from inpatients in two hospitals over 12 months. Clinical data and hospitalization history were abstracted from electronic medical records. We compared three genome sequence alignment strategies to assess similarity in cluster ascertainment. We conducted logistic regression to measure the probability of predicting prior hospital overlap between clustered patient isolates by the genetic distance of their isolates.

Results: While the three alignment approaches detected similar results, they showed some variation. A Gene-family-based alignment pipeline was most consistent across MRSA clonal complexes. We identified nine unique clusters of closely related BSI isolates. Most BSI were healthcare-associated and community-onset. Our logistic model showed that with 13 single nucleotide polymorphisms the likelihood that any two patients in a cluster had overlapped in a hospital was 50 percent.

Conclusions: Multiple clusters of closely related MRSA isolates can be identified using WGS among strains cultured from BSI in two hospitals. Genomic clustering of these infections suggests that transmission resulted from a mix of community spread and healthcare exposures long before BSI diagnosis.

Introduction

S. aureus caused nearly 119,000 bloodstream infections (BSIs) and 20,000 associated deaths in 2019 (132). These infections are exacerbated by the emergence of methicillin-resistant *S. aureus* (MRSA) strains which are resistant to treatment with conventional ß-lactam antibiotics.

Concerted national infection control efforts have decreased MRSA healthcare-associated infections (HAIs) in the United States (U.S.), particularly BSIs caused by MRSA strains historically associated with HAIs. However, the decrease in MRSA BSIs in the U.S. has slowed since 2013, and community-onset infections have recently made up the largest proportion of cases (132).

Onset of a clinically significant infection is influenced by bacterial virulence, human host factors, and triggers such as skin trauma or underlying illnesses that predispose patients to opportunistic infections (133). Asymptomatic *S. aureus* carriage is a risk factor for infection, and can be harbored in sites across the body (134), complicating elimination since detecting carriage or transmission can occur long after exposure. Consequently, hospital (95) and community (135,136) outbreaks of *S. aureus* result from direct or indirect contact with colonized individuals, contamination of an intermediate person such as a healthcare worker (32), or through environmental reservoirs. Though detecting transmission clusters of MRSA is an infection control priority in hospital settings, the transmission dynamics of MRSA among hospitalized patients with BSIs has not been thoroughly studied.

Whole genome sequencing (WGS) of bacterial genomes provides high resolution of genetic relationships between MRSA isolates and possible recent transmission. Improved access and ease of use of open-source bioinformatic resources, lower costs, and expansion of publicly available DNA sequences increases the feasibility of routine genomic analysis for cluster detection (137,138). Of great importance for detection is maximizing gene homology through genome alignments. Alignment creation includes reference-free or reference-dependent methods, which have unique trade-offs for sensitivity, specificity, and completeness of genetic data (93).

Epidemiological investigations use genetic thresholds between *S. aureus* isolates to identify or rule out clusters of related infections (79,90,95). Commonly, single nucleotide polymorphisms (SNPs) are quantified to compare isolate sequences, create multisequence alignments for phylogenetic reconstruction, and estimate the likelihood of a recent common ancestor and possible transmission given a SNP threshold (94,130). Reference choice, sample genetic diversity, and bioinformatic tools all impact which and how many SNPs are detected in a sample set and necessitate exploration of the consistency genomic alignments used to infer transmission clusters.

To elucidate transmission of MRSA BSI, we conducted a retrospective analysis of MRSA BSI at two hospitals in one university system over 12 months. We compared core-genome sequences from the isolates to detect putative transmission events between BSI patients and examined epidemiological and molecular traits of isolates shared between cluster patients. We also tested the consistency of detectable SNP differences between isolates using different sequence alignment pipelines.

Methods

Patient cohort

We identified all patients diagnosed with a MRSA BSI between July 2018 and June 2019, admitted to either of two hospitals of the University of Pennsylvania hospital system. The Hospital of the University of Pennsylvania (HUP) is a 625-bed academic tertiary and quaternary care medical center in West Philadelphia with approximately 32,000 patient admissions, 633,000 outpatient visits, and 40,000 Emergency Department visits annually. The Penn Presbyterian Medical Center (PMC) is a 324-bed urban community hospital in West Philadelphia with 12,000 admissions, 130,000 outpatient visits, and 26,000 Emergency Department visits annually. A single case of MRSA BSI was defined as a MRSA isolate collected from blood of any patient at HUP or PMC during the study period. Each subject was only included once. The study was approved by the University of Pennsylvania Institutional Review Board and given a waiver of consent, as the study was retrospective, and no data or samples were collected specifically for research purposes.

Isolate selection and DNA sequencing

Isolates were obtained from a biobank of clinical MRSA isolates cultured for routine diagnosis in the HUP Clinical Microbiology Laboratory during the study period. Isolates were screened for phenotypic antibiotic resistance using the Vitek 2 automated system, and assigned susceptibility/resistance in accordance with CLSI protocols (139). A 1µL loopful of frozen isolate was streaked onto blood agar, incubated overnight at 37℃, and a single representative colony was grown under the same conditions on a new plate. A 10µL loopful of each isolate was then frozen in a bead beating tube and underwent WGS using an Illumina MiSeq at the Penn/Children’s Hospital of Philadelphia Microbiome Center. Sequencing libraries were prepared using the Illumina Nextera library preparation kit. Sequences were made publicly available through the Sequence Read Archive (Bioproject PRJNA751847).

Bioinformatic pipelines

Paired-end 150 bp FASTQ files were passed through Bactopia workflow to assess data quality, assemble contigs, and call MLST, SCC*mec* type, antibiotic resistance and virulence genes (140). To compare SNP-based core-genome multiple-sequence alignments, the total number of assembled contigs or subsets grouped by clonal complex (CC) were passed through three pipelines: (1) randomly fragmenting assembled genomes to create “pseudoreads” and mapping these to a reference genome using Snippy (v.4.6.0) (141) (“Pseudoread pipeline”); (2) mapping assembled genomes to a reference using Parsnp (v.1.5.6) (142) (“Assembly pipeline”); and (3) evaluating reads with Markov Cluster Analysis, identifying overlapping gene clusters, and aligning core genes using the Bactopia Tools pangenome workflow (“Gene-family pipeline”). The Gene-family pipeline included PIRATE (131), ClonalFrameML (124) and maskrc-svg (v0.5) (https://github.com/kwongj/maskrc-svg) to identify and mask possible recombinant regions within the core-genome alignment. For the two reference-based pipelines, we used strain N315 (GCF_000009645.1) as reference for non-CC specific alignments (all 104 available sequences regardless of CC) and CC5-specific alignments (N=40). For the CC8-specific alignments (N=55) we used NCTC 8325 as reference (GCF_000013425.1). Pairwise SNP distances of the core-genomes were calculated using snp-dists (143). Maximum likelihood trees were created with IQ-Tree (v2.1.2) (144) using a general time reversible model allowing for invariant sites and unequal base frequencies and midpoint-rooted and visualized using ggTree (145). Bootstrap values were calculated for 1000 repetitions. Phylogenetic similarity across pipelines was measured by calculating cophenetic correlation (146) between SNP distance matrices and estimated phylogeny tip distance, and assessing Robinson-Foulds distances (147) between different alignment trees and randomly generated trees using ape (v5.5) (148).

Epidemiological investigation of clustered isolates

A transmission cluster was defined as two or more subjects whose isolates’ core-genomes differed from one another by 35 or fewer SNPs, based on the approximate cutoff for within-patient versus between-patient BSI lineages in a hospital setting (89,95). We also examined a threshold of 15 SNPs, a proposed threshold for recent inter-patient MRSA transmission (79).

Demographic data, comorbidities, Pitt Bacteremia Score, source of BSI, and in-patient mortality were abstracted from the electronic medical record (EMR) summarized and assessed for association with CC using Fisher's exact test or Student's t-test. BSIs were considered healthcare-associated (HA) if the index blood culture was drawn >48 hours after hospital admission; healthcare-associated, community-onset (HACO) if the index culture was obtained <48 hours after admission or in the community setting, and if the subject had one or more previous healthcare risk factors (hospitalization, surgery, hemodialysis, or nursing home/residential medical facility stay in the previous year; or presence of an indwelling intravascular catheter at time of culture); and community-associated (CA) if the index culture was obtained <48 hours after admission or in the community setting and the subject lacked these healthcare risk factors. The EMR was examined for evidence of overlap or sequential hospital/unit stays among cluster-subjects and visualized using vistime (https://github.com/shosaco/vistime). Admission and discharge dates were recorded for each cluster-subject for all hospital stays at any of four networked hospitals within one year before the first collected BSI isolate in a cluster and one year after the last collected BSI isolate in the cluster. These included HUP, PMC, Pennsylvania Hospital (PH), and a single, University of Pennsylvania long-term acute care hospital in Philadelphia. PH is a 481-bed urban community hospital located in the Society Hill district of Philadelphia with >27,000 hospital admissions, >24,000 Emergency Department visits, and 201,000 outpatient visits annually.

Logistic regression assessed the predictive power of SNP distances and likelihood of patient hospitalization overlaps. Goodness of fit was assessed using a receiver operating characteristic (ROC) curve and measuring the area under the curve. All analyses were conducted in R studio (v1.4.1106) (149) run with R version 4.0.4, and final figures labelled in InkScape (v0.92.5) (150). Analysis code is available at https://github.com/Read-Lab-Confederation/MRSA_bloodstream_clusters.

Results

Patient demographics and isolate characteristics

We screened all patients diagnosed with a MRSA BSI at two academic hospitals between July 2018 and June 2019, identifying 106 qualifying subjects. Of the BSI source sites that could be identified from EMR, skin site infections made up 19% and central venous catheter infections made up 14% (Table 3.1). Among included subjects, 17% died while hospitalized. From each individual, single MRSA isolates were sequenced, of which 105 had sufficient coverage for further analysis and 104 isolates were *S. aureus*. One isolate was identified by WGS as *Staphylococcus argenteus* and was excluded. Among the 104 genomes, 55 were assigned to CC8, 49 of which were USA300 strains; 40 were assigned to CC5; and the remaining nine were assigned CC30, CC72, and CC78. No significant association emerged between the two most common CCs (CC5 and 8) and sex, age group, race, ethnicity, BSI source site, hospital death, Pitt bacteremia score, or hospital of diagnosis (Supplementary Table S3.1 and S3.2).

Table 3.1. Demographics and clinical outcomes of subjects with MRSA bloodstream infection (n=106)

Demographic Characteristic	Number (%) Patients
Total	106
Age Group	
20-29	12 (11%)
30-39	13 (12%)
40-49	16 (15%)
50-59	18 (17%)
60-69	32 (30%)
70+	15 (14%)
Sex	
Female	51 (48%)
Male	55 (52%)
Race	
Asian	1 (1%)
White	50 (48%)
Black	49 (46%)
Other/Unknown	6 (6%)

Clinical Characteristic	Number (%) Patients
Total	106
Source of BSI	
Arteriovenous graft	4 (4%)
Central venous catheter infection	15 (14%)
Device infection	4 (4%)
Respiratory source	2 (2%)
Skin site	20 (19%)
Surgical site	4 (4%)
Other	3 (3%)
Unknown	52 (49%)
Hospital of BSI diagnosis	
Hospital A	65 (61%)
Hospital B	41 (39%)
Infection setting	
HA	22 (21%)
CA	11 (10%)

Ethnicity		HACO	71 (68%)
Hispanic/Latino	2 (2%)	**In-hospital death**[a]	
Non-Hispanic/Latino	99 (93%)	No	88 (83%)
Unknown	5 (5%)	Yes	18 (17%)
		Pitt Bacteremia Score	
		Mean (SD)	2.1 (2.6)
		Median (Range)	1.00 (0, 10.0)

[a] Indicates death prior to discharge during the index MRSA BSI hospitalization. Abbreviations: BSI: bloodstream infection; CA: community-associated; HA: healthcare-associated; HACO: healthcare-associated, community-onset; SD: standard deviation.

Assessment of sequence alignment pipelines

We generated multiple alignments of all isolate sequences using three approaches to determine their effect on pairwise SNP distances. Alignments generated with all 104 isolates had lower distances compared to CC-specific alignments. SNP distances produced by the Gene-family-pipeline were consistent between CC groups and whole species alignments (Fig. 3.1A-B), whereas the SNP distances produced by the Pseudoread- and Assembly-pipelines were greater when isolates of the same CC were the input (Fig. 3.1 C-F). Pipeline choice on phylogenetic structure was assessed by comparing tree topology and SNP matrices across pipelines and sequence input groupings (Table 3.2). The cophenetic correlation showed the highest correlation for alignments produced from CC-specific inputs, though all alignment pipelines and inputs produced a value greater than 0.90. Tree topology across pipelines suggested that trees are highly similar to one another compared to a random tree.

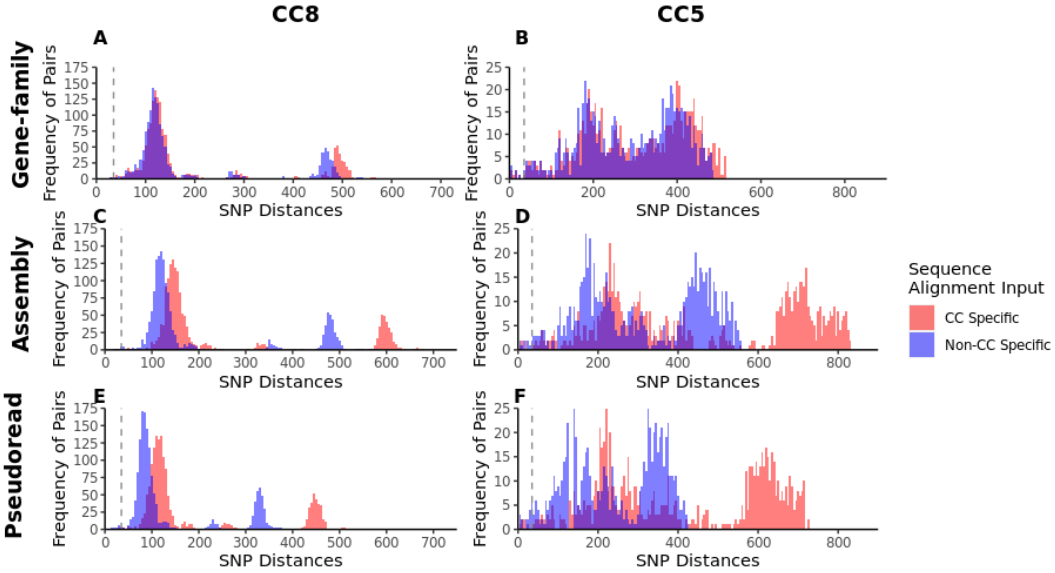

Figure 3.1. Frequencies and distribution of single nucleotide polymorphism distances between isolates vary by alignment tool. The frequency of pairwise distances between isolates from clonal complexes (CC) 8 and 5 were quantified from distance matrices derived from alignments generated from two groupings of isolate input: The total number of isolates in the investigation (blue) or CC-specific isolates only (red). Isolate inputs were aligned using each of the three alignment pipelines, the Gene-family pipeline (A,B), Assembly pipeline (C,D), and Pseudoread pipeline (E,F).

Table 3.2. Comparability phylogenetic fit of alignment pipelines using Cophenetic correlation (R^2), alignment size, and Robinson-Foulds (RF) comparison[a] by alignment pipeline

Pipeline	Total isolates (N=104)			CC5-specific isolates (N =40)			CC8-specific isolates (N = 55)		
	R^2	Alignment Size (bp)	RF-values	R^2	Alignment Size (bp)	RF-value	R^2	Alignment Size (bp)	RF-values
Gene-family-pipeline	0.984	2,141,357	56,52,200	0.984	2,182,742	10,8,72	0.987	2,176,046	40,36,104
Pseudoread-pipeline	0.983	2,839,469	46,56,202	0.993	2,839,469	10,10,72	0.999	2,821,361	34,40,104
Assembly-pipeline	0.929	2,163,693	52,46,202	0.995	2,497,454	8,10,72	0.999	2,482,874	34,36,104

[a]Row alignment pipeline compared to each other alignment pipeline and a random tree of the same number of phylogenetic tips

Identification of suspected transmission clusters

Using alignments from each pipeline containing 104 isolates, we identified nine clusters (C1-C9) among 29 isolates that differed by 35 SNPs or fewer from at least one other subject isolate (Table 3). The Pseudoread-pipeline clustered 29 isolates, the Assembly-pipeline clustered 21, and the Gene-family-pipeline clustered 19. Five clusters contained CC5 isolates, three clusters were CC8, and one cluster was CC30. The median cluster size was three isolates (range

2-6). The longest collection date difference between clustering isolates was 265 days (C1), and the shortest 12 days (C6). Median SNP differences were variable across clusters, and smaller differences did not correlate with shorter collection date differences.

Table 3.3. Summary of suspected MRSA transmission clusters identified through Pseudoread-, Assembly-, and Gene-family-alignment pipelines among 104 sequential MRSA bloodstream infection patients at 2 hospitals

Transmission Cluster	MRSA Isolate	Clonal Cluster	Number of isolates	Median Pairwise SNP Difference (Range)			Median Collection Date Difference, Days (Range)
				Pseudoread-pipeline	Assembly-pipeline	Gene-family-pipeline	
C1	SAMN20960259, SAMN20960281, SAMN20960331	CC5	3	11 (3-12)	16 (4-16)	14 (5-16)	177 (110 - 265)
C2	SAMN20960260, SAMN20960274	CC5	2	6	7	7	61
C3a	SAMN20960263, SAMN20960326, SAMN20960280, SAMN20960314, SAMN20960328	CC5	5	35 (20 - 46)	44 (26-62)	50 (36 - 62)[a]	128 (7 - 241)
C3b	SAMN20960280, SAMN20960314, SAMN20960328	CC5	3	25 (20 - 25)	32 (26-36)	39 (36 - 39)[a]	113 (37 - 150)
C4	SAMN20960270, SAMN20960325	CC5	2	20	26	24	189
C5a	SAMN20960271, SAMN20960343	CC5	4	29 (6 - 35)	44 (10-53)[a]	41 (33 - 46)[a]	119 (56 - 237)

C5b	SAMN20960271, SAMN20960343	CC5	2	29	42[a]	33	237
C5c	SAMN20960298, SAMN20960324	CC5	2	6	10	7	78
C6a	SAMN20960276, SAMN20960282, SAMN20960287, SAMN20960293, SAMN20960301, SAMN20960306	CC8	6	29 (15 - 42)	39 (21-62)	38 (26 - 53)[a]	54 (12 - 121)
C6b	SAMN20960276, SAMN20960282, SAMN20960293, SAMN20960301, SAMN20960306	CC8	5	26 (15 - 31)	35 (21-40)	37 (26 - 43)	66 (12 - 121)
C6c	SAMN20960276, SAMN20960282, SAMN20960293, SAMN20960306	CC8	4	23 (15 - 30)	32 (21 - 36)	34 (26-38)	63 (32 - 121)
C7	SAMN20960299, SAMN20960305, SAMN20960334	CC8	3	34 (30 -34)	39 (36 - 41)	45 (41 - 50)	80 (24 -104)
C8	SAMN20960313, SAMN20960323	CC8	2	1	1	1	28
C9	SAMN20960316, SAMN20960337	CC30	2	23	25	22	67

[a]Partial or no detection of isolates as part of the cluster

Phylogenetic analysis of isolates

To assess phylogenetic relationships within clusters, we created a representative tree using the Gene-family-pipeline of the 104 isolates. This tree was selected because it had the strongest cophenetic correlation, tree structure similarity, and conservation of SNP distances between pipelines for the 104 isolates together (Table 3.2; Fig 3.1A-B). BSI isolates occupied significantly divergent clades of CCs (Shimodaira–Hasegawa – approximate likelihood ratio test and ultrafast bootstrap values >70) (Fig 3.2A). Candidate transmission clusters arose from distinct sub-lineages (Fig 3.2B). The largest cluster, C5, diverged significantly from other CC8 isolates, and isolates were identified as part of the CC8c lineages (151). Cluster and non-cluster isolates had varied distributions for infection setting, with most BSIs categorized as HACO (68%). At a 15-SNP threshold, only isolates in clusters C1, C2, C5(a,c), and C8 remained clustered. All isolates were susceptible to vancomycin and daptomycin but isolates in both the CC5 and CC8 clades showed resistance to multiple ß-lactams and quinolones. Thus, multiple lineages of MRSA associated with BSI could transmit multiclass-resistant strains between patients.

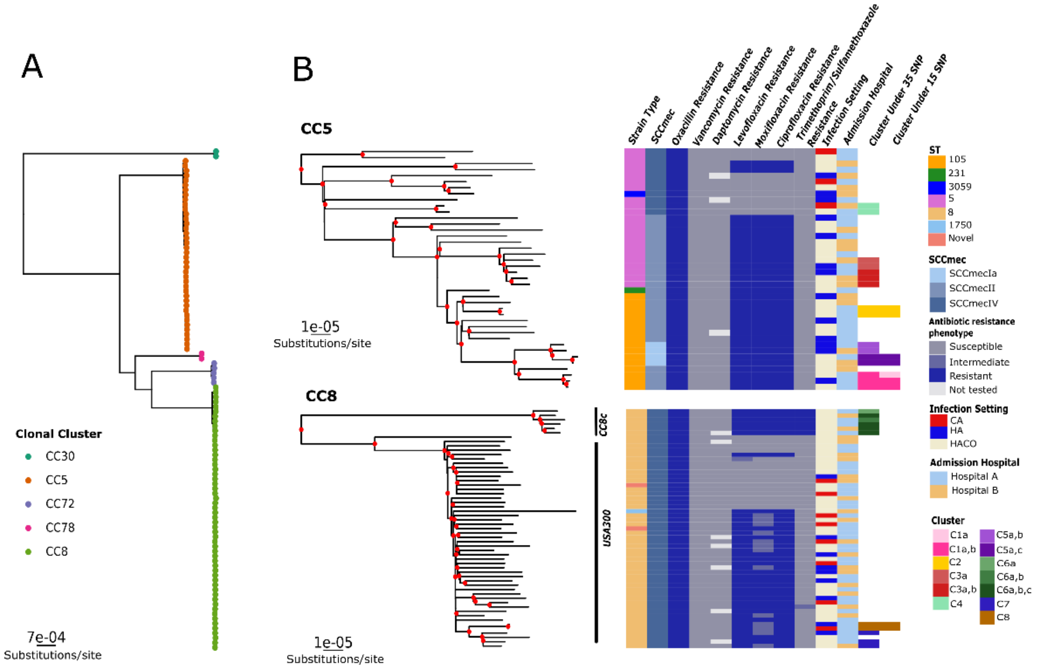

Figure 3.2. Suspected transmission clusters fall into distinct clonal groups. Maximum likelihood trees were generated from the PIRATE alignment of 104 isolates and visualized using ggtree. (A) Tree indicating clades containing individual clonal complexes (CCs). (B) Subtrees from the complete maximum likelihood trees for the two most abundant CCs. Nodes with bootstrap values >= 70 are marked in red. Heat maps show strain type, SCC*mec* element type, and resistance phenotype for indicated antibiotics per sequence, infection setting (Healthcare-associated [HA], Community associated [CA] and Healthcare-associated community-onset [HACO]), admission hospital, and transmission cluster at a threshold of 35 SNPs or 15 SNPs.

Genomic similarity predicts overlapping hospital stay in transmission clusters

For every cluster-subject we examined hospitalization history at four networked hospitals in the University of Pennsylvania system one year before the first index BSI isolate and one year after the last patient index isolate per cluster. Six clusters included subjects with overlapping hospital stays, of which three had median SNP distances between 1-16 with corresponding hospital unit overlaps (Table 3.3; Fig. 3.3). Cluster C5c had a median SNP difference of seven (range 6-10 SNPs across pipelines) with no common hospital overlap. In comparison, cluster C4 had no subjects with overlapping hospital admissions prior to their index BSI, but a median SNP distance range of 20 - 26 SNPs across pipelines.

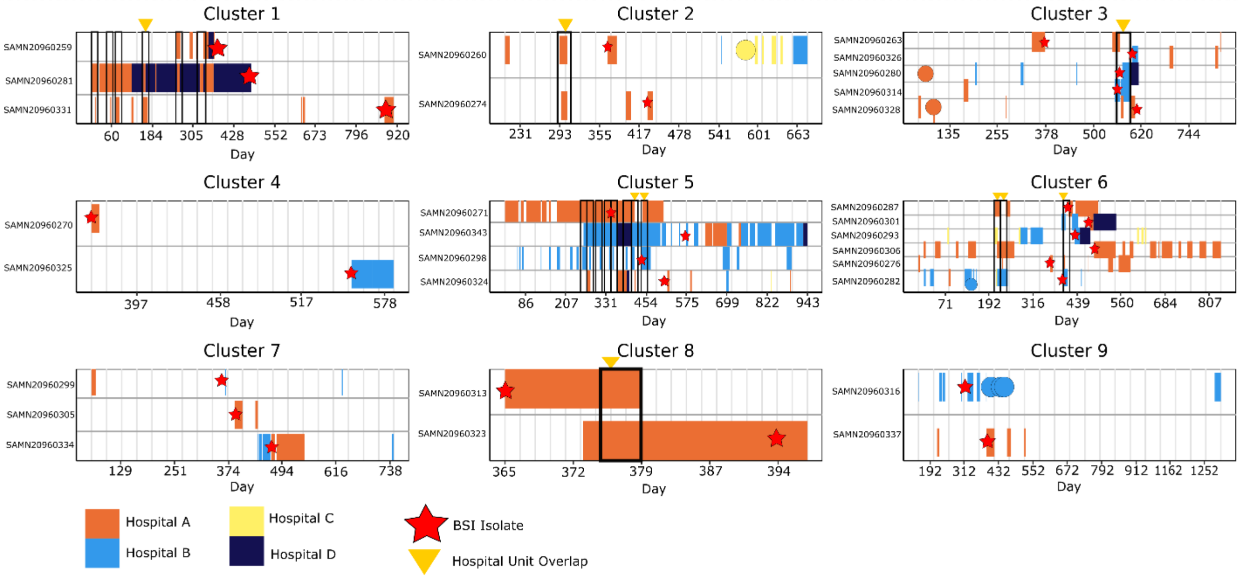

Figure 3.3. Hospitalization history among patients in genomic BSI clusters. Hospitalization history at 4 study hospitals (A, B, C, and D) up to 365 days before the date of the earliest MRSA bloodstream isolate culture in each cluster (relative Day 0) and up to 365 days after the latest MRSA bloodstream isolate in the cluster. Note that bloodstream infections were only included at hospitals

A and B. Rows represent the hospitalization history of each patient associated with a sequenced cluster isolate. Colored rectangles and circular marks represent individual hospitalization durations (rectangles) or one-day admissions (circles); the color indicates Hospital A, B, C, or D. Black outlined boxes represent areas where two or more patients overlapped in the same hospital at the same time. Red stars indicate the date of collection of the sequenced BSI isolate for each patient. Yellow triangles indicate a hospitalization where two or more patients overlapped in the same hospital unit.

We performed a logistic regression to measure the association between likely hospital exposure and SNP difference assessing a SNP threshold range (Fig. 3.4). The log odds of clustered patient pairs overlapping in the same hospital decreases by 0.065 with every increase of one SNP (p=0.05), and showed that with 13 SNPs the likelihood that any two patients in a cluster overlapped in a hospital was 50 percent, with a trend toward no overlap at higher SNP differences (Fig. 3.4A). The ROC area under the curve classified known prior overlapping hospitalizations 66% of the time from the SNP difference (Fig. 3.4B).

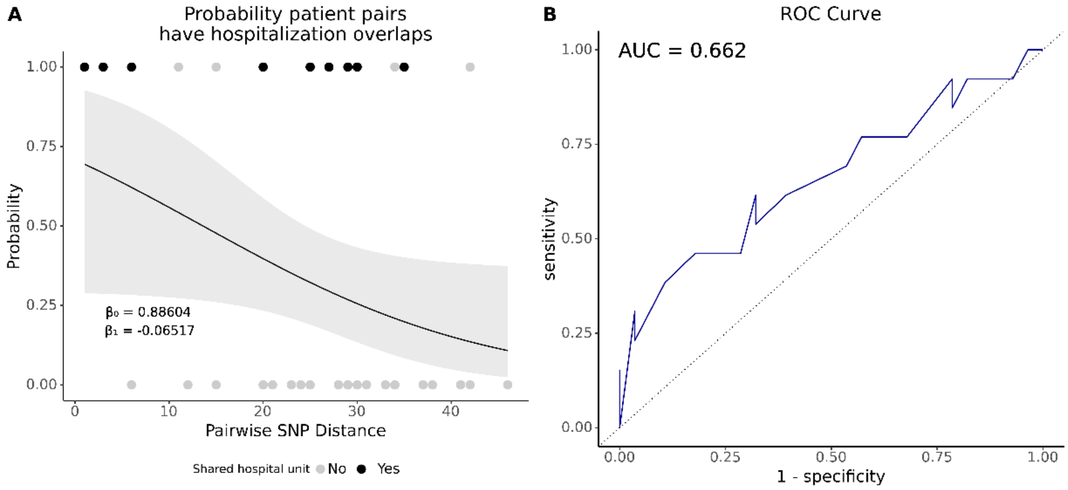

Figure 3.4. Higher SNP distances trend toward ruling out hospital overlaps between clustering patients. (A) Logistic regression model indicating the relationship between patient pairs overlapping in the same hospital at the same time (prior to the diagnosis of an index MRSA bloodstream infection) and the pairwise SNP distance. Points indicate the true result for each pair as overlapping (1.0) or not overlapping (0). The color of the points indicates whether hospital overlap patient pairs also overlapped (black) or did not overlap (gray) in the same hospital unit. Gray ribbon indicates the 95% confidence interval. (B) Receiver operating characteristic (ROC) curve of the logistic model in A. Area under the curve (AUC) = 0.662.

Discussion

We combined clinical and genome data to describe a cohort of 104 U.S. MRSA BSI patients. The predominant genetic backgrounds of MRSA isolates in this study is consistent with known prevalence of CC8 and CC5 MRSA strains causing healthcare- and community-associated infections in the U.S (152). The resolution of WGS was critical for identifying clusters of BSIs that would have otherwise gone unnoticed in the hospital setting. It is well characterized that WGS is useful for *S. aureus* outbreaks in hospitals (88–90,95,137,153), ,though many reports focus on its use in emergent, point-source outbreaks, such as those occurring in neonatal intensive care unit with an identifiable index case (88,89,153). In other instances, WGS confirmed related cases of MRSA infection only after initial outbreak detection by other means, including an unusual antibiograms (88) or uncommon strain types (137). Collectively, these investigations identified an epidemiologically significant core-genome SNP difference as small as 13 SNPs (79) to as large as 40 SNPs (52) among outbreak isolates.

A SNP threshold under 35 was effective for cluster detection with evidence of prior hospital overlaps among adult patients in a population where transmission pathways are difficult to identify. Four clusters showed pairwise differences between 1-25 SNPs and patients with diagnosis date within three months. Considering estimates of *S. aureus* neutral mutation of approximately 5-6 SNPs per genome per year (100), a likely scenario is a recent common exposure in a healthcare setting several weeks to months prior to BSI onset for clustered subjects. However, clusters lacking evidence of a hospital overlap also had small SNP difference ranges, suggesting alternative routes of MRSA transmission among BSI patients, such as hospital environmental reservoirs like equipment (32,154) or a community reservoir of patients carrying MRSA (155), possibly reintroducing bacteria to the hospital. We demonstrated that it is reasonable to investigate healthcare histories for patients at or below 13 SNPs to find sources of transmission associated with hospital settings.

Most U.S. hospitals have not yet implemented a WGS surveillance system for infection control. Hospitals can approach bioinformatic surveillance using commercial workflows with integrated processes (153) or open source options (108), or create robust in-house surveillance methods (7). We demonstrated that different approaches to sequence alignment detect similar

SNP differences and phylogenies. However, alignment sizes and the number of clusters at the threshold of interest did differ. Choosing the most appropriate tool ideally optimizes sensitivity and comparability across investigations. The Gene-family approach consistently detects similar SNP differences among alignments of mixed clonal clusters and is suited to studies comparing diverse sample sets. However, higher sensitivity can be achieved using an Assembly- or Pseudoread-pipeline because they also compare a larger portion of the genome where SNPs can accumulate. We suggest future studies use both approaches, first for general detection of clusters with highly sensitive approaches, followed by a Gene-family approach to compare clusters across a broader context of transmission cluster history in a specific environment. A sliding scale (156) or a threshold range (79) could also offer a more flexible alternative for including patients in transmission investigations.

Reference-based alignments and phylogenetic reconstruction is advantageous for identifying transmission events in healthcare settings, particularly where MRSA infections are rare (130,156). However, *S. aureus* transmission from healthcare facilities into community settings and back suggest that hospitals and the surrounding community are a single reservoir of transmission (89). Our investigation also points to the importance of long-term MRSA carriage prior to diagnosis of a BSI. Overlapping hospitalization may provide an opportunity for MRSA transmission and subsequent asymptomatic colonization in a recipient patient but BSI symptoms may occur weeks or months later. Consequently, clusters are not identified after the critical moment of transmission when infection control interventions could be implemented. As WGS surveillance becomes prospectively implemented, Gene-family alignments are advantageous for assessing increasingly diverse collections of isolates in a hospital or single healthcare system.

In our analysis, the long span of time between BSI onset among cluster patients and lack of an obvious transmission pathway suggests possible intermediate patients without a BSI but still carriers of the infecting MRSA strains. We did not collect isolates from the hospital environment or from healthcare workers directly, so we cannot discern the role of these intermediaries for transmission in the clusters.

We revealed MRSA BSI clusters among adults with various prior healthcare exposures in a setting with relatively high incidence of MRSA infections. We identified genetically similar clusters while routine epidemiological signal was weak, but with further investigation suggested healthcare exposures well before BSI presentation. Including WGS as a part of current routine colonization screenings for MRSA in high-risk clinical settings could identify and prevent transmission events in areas of hospitals not regularly scrutinized by infection control staff. With robust and consistent cluster detection pipelines and the prospective collection of detailed exposure histories, with a focus on identifying exposures during hospitalization to specific healthcare workers, fomites, and medical procedures, outbreak sources can be better resolved before the onset of a BSI event.

Additional Information and Declarations

Funding

This work was supported by supported by grant # AI139188 from the National Institutes of Health (NIH) to BMT, MZD, and TDR.

Conflicts of Interest

DAP is an Associate Editor for *Clinical Infectious Diseases*.

Acknowledgements

Thank you to Laurel Glaser for assistance with the biobanking of isolates used in this study, and to Katrina Hofstetter for support in troubleshooting R code. We also thank the Penn/Children's Hospital of Philadelphia (CHOP) Microbiome Center for performing the isolate sequencing used in this analysis.

Supplementary Information

Table S3.1. Distribution of patient demographics and clinical outcomes across isolate clonal clusters

	CC30 (N=2)	CC5 (N=38)	CC72 (N=5)	CC78 (N=2)	CC8 (N=52)
Age Group					
20-29	0 (0%)	3 (7.9%)	2 (40.0%)	0 (0%)	7 (13.5%)
30-39	0 (0%)	2 (5.3%)	1 (20.0%)	1 (50.0%)	8 (15.4%)
40-49	0 (0%)	9 (23.7%)	0 (0%)	0 (0%)	7 (13.5%)
50-59	1 (50.0%)	4 (10.5%)	1 (20.0%)	1 (50.0%)	8 (15.4%)
60-69	0 (0%)	15 (39.5%)	0 (0%)	0 (0%)	14 (26.9%)

70-79	1 (50.0%)	4 (10.5%)	1 (20.0%)	0 (0%)	6 (11.5%)
80-89	0 (0%)	1 (2.6%)	0 (0%)	0 (0%)	1 (1.9%)
90-99	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
100-110	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (1.9%)
Sex					
Male	2 (100%)	19 (50.0%)	3 (60.0%)	1 (50.0%)	25 (48.1%)
Female	0 (0%)	19 (50.0%)	2 (40.0%)	1 (50.0%)	27 (51.9%)
Race					
Black	1 (50.0%)	20 (52.6%)	3 (60.0%)	2 (100%)	20 (38.5%)
White	1 (50.0%)	16 (42.1%)	2 (40.0%)	0 (0%)	28 (53.8%)
Asian	0 (0%)	1 (2.6%)	0 (0%)	0 (0%)	0 (0%)
Other	0 (0%)	1 (2.6%)	0 (0%)	0 (0%)	2 (3.8%)
Don't Know	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (3.8%)
Ethnicity					
Non-Hispanic/Latino	2 (100%)	36 (94.7%)	5 (100%)	2 (100%)	48 (92.3%)
Hispanic/Latino	0 (0%)	1 (2.6%)	0 (0%)	0 (0%)	1 (1.9%)

Refused	0 (0%)	1 (2.6%)	0 (0%)	0 (0%)	3 (5.8%)
Source Site of BSI					
Skin site	2 (100%)	6 (15.8%)	2 (40.0%)	0 (0%)	10 (19.2%)
Arteriovenous Graft	0 (0%)	1 (2.6%)	0 (0%)	1 (50.0%)	1 (1.9%)
Central venous catheter infection	0 (0%)	7 (18.4%)	1 (20.0%)	0 (0%)	6 (11.5%)
Device infection	0 (0%)	1 (2.6%)	1 (20.0%)	0 (0%)	2 (3.8%)
Other	0 (0%)	1 (2.6%)	0 (0%)	0 (0%)	2 (3.8%)
Respiratory source	0 (0%)	1 (2.6%)	0 (0%)	0 (0%)	1 (1.9%)
Surgical site	0 (0%)	1 (2.6%)	0 (0%)	0 (0%)	1 (1.9%)
Unknown	0 (0%)	19 (50.0%)	1 (20.0%)	1 (50.0%)	28 (53.8%)
Urinary source	0 (0%)	1 (2.6%)	0 (0%)	0 (0%)	1 (1.9%)
Hospital					
Hospital A	1 (50.0%)	23 (60.5%)	1 (20.0%)	2 (100%)	33 (63.5%)
Hospital B	1 (50.0%)	15 (39.5%)	4 (80.0%)	0 (0%)	19 (36.5%)
Pitt Bacteremia Scale					

Mean (SD)	0.500 (0.707)	2.21 (2.78)	2.80 (3.83)	0 (0)	2.25 (2.60)
Median [Min, Max]	0.500 [0, 1.00]	1.00 [0, 9.00]	1.00 [0, 9.00]	0 [0, 0]	2.00 [0, 10.0]
In-hospital Death					
No	2 (100%)	30 (78.9%)	5 (100%)	2 (100%)	43 (82.7%)
Yes	0 (0%)	8 (21.1%)	0 (0%)	0 (0%)	9 (17.3%)

Table S3.2. Association of Patient demographics and clinical outcomes with clonal clusters CC8 and CC5

	CC5 **(N=38)**	**CC8** **(N=52)**	**P-value***
Age Group			
20-29	3 (7.9%)	7 (13.5%)	0.519
30-39	2 (5.3%)	8 (15.4%)	
40-49	9 (23.7%)	7 (13.5%)	
50-59	4 (10.5%)	8 (15.4%)	
60-69	15 (39.5%)	14 (26.9%)	
70-79	4 (10.5%)	6 (11.5%)	
80-89	1 (2.6%)	1 (1.9%)	
90-99	0 (0%)	0 (0%)	

100-110	0 (0%)	1 (1.9%)	
Sex			
Female	19 (50.0%)	27 (51.9%)	1
Male	19 (50.0%)	25 (48.1%)	
Race			
Asian	1 (2.6%)	0 (0%)	0.369
Black	20 (52.6%)	20 (38.5%)	
Other	1 (2.6%)	2 (3.8%)	
White	16 (42.1%)	28 (53.8%)	
Don't Know	0 (0%)	2 (3.8%)	
Ethnicity			
Hispanic/Latino	1 (2.6%)	1 (1.9%)	0.82
Non-Hispanic/Latino	36 (94.7%)	48 (92.3%)	
Refused	1 (2.6%)	3 (5.8%)	
Source Site of BSI			
Arteriovenous Graft	1 (2.6%)	1 (1.9%)	0.995
Central venous catheter infection	7 (18.4%)	6 (11.5%)	
Device infection	1 (2.6%)	2 (3.8%)	
Other	1 (2.6%)	2 (3.8%)	
Respiratory source	1 (2.6%)	1 (1.9%)	
Skin site	6 (15.8%)	10 (19.2%)	
Surgical site	1 (2.6%)	1 (1.9%)	

Unknown	19 (50.0%)	28 (53.8%)	
Urinary source	1 (2.6%)	1 (1.9%)	
Hospital			
Hospital A	23 (60.5%)	33 (63.5%)	0.828
Hospital B	15 (39.5%)	19 (36.5%)	
Pitt Bacteremia Scale			
Mean (SD)	2.21 (2.78)	2.25 (2.60)	0.946
Median [Min, Max]	1.00 [0, 9.00]	2.00 [0, 10.0]	
In-hospital Death			
No	30 (78.9%)	43 (82.7%)	0.786
Yes	8 (21.1%)	9 (17.3%)	

*P-values reflect results of two-tailed T-test or Fisher's Exact test

Chapter 4: Genomic investigation of MRSA bacteremia relapse reveals diverse genomic profiles but convergence in bacteremia-associated genes

Brooke M. Talbot, Natasia F. Jacko, Katrina Hofstetter, Timothy D. Read, Michael Z. David

Abstract

Background. Recurrence of Methicillin-resistant *Staphylococcus aureus* (MRSA) bacteremia is a high-risk complication for patients. Characterizing the patterns of risk relative to the initial infections is complex.

Methods. We investigated clinical and bacterial factors contributing to recurrence of MRSA bacteremia among a cohort of patients in Philadelphia, Pennsylvania. Patient demographics, clinical history, and suspected sources of BSI were collected. Infection isolates were short read whole-genome sequenced and de novo assembled. All BSI isolates were core genome-aligned to assess pairwise single nucleotide polymorphism (SNP) distances, and to create a maximum likelihood tree to infer phylogenetic relationships. Recurrence was defined as MRSA bacteremia occurring 30 days or more from previous MRSA bacteremia experienced by the same person. Infections were relapses if isolates from the same patient were less than 25 SNPs different or if the genomic distance was smaller between isolate from the same patient than the next closest isolate from a different patient. CC and time between infections were compared between relapse and non-relapse recurrences. Convergent genetic traits were assessed by quantifying unique SNPs per gene emerging in relapse-infection lineages.

Results. Among 411 BSI subjects, 32 had at least one repeated MRSAbacteremia event. There were 26 subjects with relapse infections and 8 with infections from a new strain, with two patients with both relapse and distinct recurrences. CC distribution was similar between recurrence and non-recurrence isolates (p=0.6132). Relapses occurred sooner after the prior infection (Median 155 days, interquartile range (IQR) 88-269 days) compared to new strain recurrences (Median 248 days, IQR 105-599 days), though this was not statistically significant. Recurrences sharing the same CC as their paired previous infection were not distinct from chance and occurred 55% of the time. Genes with SNPs occurring in multiple relapse lineages have roles in antibiotic resistance and virulence, including 5 lineages with mutations in *mprF* and 3 lineages with mutations in *rpoB*.

Conclusions. Recurrent MRSA infections have a diverse strain background, but relapses can be readily distinguished from newly acquired infections. Continued genomic analysis can reveal the roles of unique mutations in pathoadaptation.

Introduction

Staphylococcus aureus bacteremia (SAB) is a complex clinical syndrome which often leads to severe patient outcomes, including endocarditis and other metastatic infections (157). SAB is associated with high mortality and strains with increased antibiotic resistance (158–160). In the healthcare setting, where it is most intensively studied, asymptomatic colonization (161), intravenous drug use (162,163), and involvement of central lines in clinical care (33) all increase the risk of SAB.

Recurrence of SAB, where patients experience SAB after assumed resolution of a previous infection, is an ongoing clinical challenge. Global records demonstrate that among five and 15 percent of patients with a Methicillin-resistant SAB episode experience a recurrence, and the risks associated with recurrence in the blood are similarly heterogenous to bloodstream infections overall

(33,159,163–168). Known risk factors for recurrence of SAB in adults include younger patient age, presence of a foreign body, hemodialysis dependence, valvular heart disease, liver cirrhosis, and endocarditis (165,168). Drug resistance can also emerge from mutations that confer cross resistance of first-line treatments or sequential mutations that lead to multidrug resistance, further complicating prevention of more difficult to treat recurrences among patients with persistent infections or patients with deep-seated foci (169). Therefore, understanding the nature of the recurrence helps better understand a clinical course of action and future prevention of ongoing infection.

One clinical challenge is disentangling new infections after true clearance from cryptically persisting bacteria within the host. With primary bacteremia, where the focus of infection is unknown, only general measures can be taken to prevent a future infection and recurrence becomes more difficult to predict. After diagnosis of SAB, follow-up blood cultures are collected typically 2-4 days after the beginning of antibiotic treatment and through to the first negative blood culture, though intermittent negative cultures are known to occur in some persistent *S. aureus* infections. Genetic typing has been used to distinguish persistent populations of bacteria that remain from a previous infection in the same person, known as relapse of infection (165,170). However, there is no definitive time interval that is used to determine within-host persistence from successfully circulating clones that may cause re-infections. Therefore, recurrence of infections currently are heterogeneously defined by a combination of genetic pattern, time interval between negative cultures and a new onset, and suspicion of a previous source of infection as the cause of a subsequent infection (33,165–168). Further, the overall prevalence of *S. aureus* colonization, and the known asymptomatic spread of *S. aureus* from person to person make it challenging to determine if a recurrence emerges because of exposure to anew *S. aureus* strain, or if a previous strain was cryptically persistent on a person even after treatment. The consequence of misidentifying recurrence could result in failures to prevent future infections; If an infected person is indicated as having a new infection when in fact there is persistent colonization or infection of a foci, future recurrence may be possible; In contrast, new infections thought to be associated with a previous infection could make it difficult to clearly identify patterns of *S. aureus* introduction in a healthcare setting or community setting, slowing down efforts to identify areas for infection prevention.

Certain *S. aureus* lineages have been implicated as more likely to develop SAB, such as clonal complexes (CC) 5 (65), 30 (65), and 8 (159) and these are also strains more likely to be encountered in healthcare settings in the US. Individual mutations and genes associated with bacteremia alter traits related to virulence regulation (22,72,73,158,160,170–177), antibiotic targets (158,160,170,178,179) and the development of small colony variants (180,181). Both methicillin-resistant (MRSA) and susceptible strains (MSSA) can cause SAB (182,183). *S. aureus* causes chronic invasive infections across different body niches, and the transition from colonizing to invasive results in relatively quick host adaptation (129,169,184). However, adaptations to invasion are generally considered to exist only in an extent colony present in the invasion (i.e. the bloodstream), and therefore SABs are thought to be an evolutionary dead-end for the clone (129,184). Several genomics studies have revealed that mutations in a limited number of loci appear to increase the risk for metastatic and persistent infections, and possible recurrence. Arguably, however, SAB episodes occupy a continuum between bacterial populations chronically invading tissue and acutely invading the cardiovascular system. Therefore, the quality and quantity of genetic change present in a set of subsequent SAB episodes could help researchers understand the collective set of drivers in the host environment that lead to recurrences as well as help differentiate new infections from cryptic and persistent infections.

To define genetic differences and risks for recurrence of SAB, we examined a cohort of patients in Philadelphia, Pennsylvania experiencing SAB across five years who received care at a single hospital system. We performed whole-genome sequencing on single isolates from episodes of SAB from individuals, including isolates from subsequent episodes in individuals experiencing a recurrence. We examined the strength association in host clinical factors and bacterial genetics to help further define risks for recurrent SAB, as well as distinguish the transmission and adaptive history of infections that result in a relapse or a new infection.

Methods

Subject cohort and isolate collection

This study was considered exempt by the University of Pennsylvania Institutional Review Board. Subject isolates were included from a cohort of adult patients admitted to at least one of two hospitals in Philadelphia, Pennsylvania and diagnosed with Methicillin-resistant *S. aureus* (MRSA) bacteremia between July 2018 and February 2022. A single colony representative isolate was taken for each bacteremia episode. Additional clinical and demographic information was collected through medical chart reviews for all infections, including age at time of isolate collection, race, ethnicity, sex at birth, death within 30 days of infection, comorbidities and chronic conditions, antibiotic treatment, suspected source site of infection, and healthcare-associated acquisition of infection. Antibiotic resistance phenotypes were collected from clinical microbiology records associated with the unique bacterial isolates. Antibiotic resistance was assessed using the Vitek 2 automated system, and assigned susceptibility/resistance in accordance with Clinical and Laboratory Standards Institute protocols (185).

Clinical distinction of relapse and new infections

For subjects with a recorded subsequent SAB event (a "recurrence") in the study period each event was classified as a "relapse" or "new infection" according to the following criteria as outlined in Figure S4.1: MRSA bacteremia episodes among subjects include all isolates collected in a 30-day period from the first episode isolate. At each discrete medical encounter where MRSA bacteremia was detected, subjects with one or more MRSA isolate were assessed for any previous MRSA bacteremia. If patients had a record of MRSA bacteremia, then the time interval between the collection date of the isolate at the encounter and the last known index isolate collection date was checked for whether it was greater than 30 days. All MRSA bacteremia isolates collected outside of the 30-day period but with record of a previous bacteremia episode during the study period were considered "recurrences." Recurrent infections were then categorized into "new infections" and "relapse" infections according to either clinical criteria or genomic criteria. For clinical criteria, new infections had to fulfill any of the following: The episode was 30 days or more since the last positive blood culture, all symptoms at the source site and metastatic sites of the previous infection were resolved, no new antibiotics were prescribed after completion of definitive therapy, the site of the infection was different from the previous infection and the subject was on suppressive antibiotics, if a central venous catheter was changed then it was changed over a wire (186), or the source of infection was clinically ruled to be different from the previous infection. Otherwise, the bacteremia episode was considered a relapse.

Sequencing quality and phylogenetics

Genomic DNA was extracted from *S. aureus* isolates and sequenced at the Children's Hospital of Pennsylvania SEQ Center, using a paired-end short read whole-genome shotgun strategy as previously described (187). Reads were processed using Bactopia (v 3.0.0) (140). Briefly, in the Bactopia pipeline adaptor sequences were removed, and reads were de novo assembled using Shovill (v.1.1.0). Sequences were used for further investigation if reads had at least an average per-read quality score of Q12 and a mean read length of 49bp, and the genomic assembly had at least 20x coverage and no more than 500 contigs. Multilocus sequence type (ST) and CC were assigned by calling Ariba (v 2.14.6) in the Bactopia workflow, which utilized the *S. aureus* specific ST scheme from PubMLST (84) . For novel STs or STs without a defined CC, the CC was manually defined based on the clade position of the isolate within a maximum likelihood tree and its closest relative with a defined CC. In order to differentiate distinct infection lineages, a core genome alignment was created among all MRSA BSI isolates and methicillin-susceptible *S. aureus* Strain Newman (GCF_020985245.1) as an outgroup using the Bactopia subworkflow "pangenome" using PIRATE (131). Areas of likely recombination were identified and masked using ClonalFrame ML (v.1.12) (124). A maximum likelihood tree of all isolates was created from the masked alignment with IQTree (2.1.2) (144) and ModelFinder (188), which determined the best fit model to be a generalized time reversible model with Empirical base frequencies plus the FreeRate model. Raw reads for this study are publicly available in the Sequence Read Archive under the project ID PRJNA751847.

Genomic definition of recurrences and cluster identification

Snp-dists (0.8.2) was used to calculate the pairwise distance of single nucleotide polymorphisms (SNPs) between aligned isolates. Subjects with recurrent episodes of bacteremia were categorized as having a relapsed infection or new infection based on the genomic difference from the isolate of the most recent previous episode relative to the episode of comparison. Relapsed infections were defined as episodes where the isolate genomes were <= 25 SNPs different from one another, or isolates whose most-recent common ancestor to another isolate in the overall population was from an infection within the same human subject and which shared the same ST. Cohen's kappa (189,190) was calculated between the clinical and genomic definitions of relapse to assess agreement between the methods, as well as sensitivity and specificity of the clinical definition in comparison to a genomic approach.

Association of clusters with demographic and clinical data

The distribution of demographic, clinical, and genomic characteristics was compared between genomically defined relapses and new infections using Peason's Chi-square test or Fisher's exact test. Isolates between subjects were also compared to identify potential clusters of highly related infections. For all relapse-associated isolates that clustered with isolates from a separate subject, subtrees of the larger ML tree were created and examined for branch structure to identify the role of relapse isolates in putative onward transmission. The time in days between collection date of the earliest isolate in the cluster (designated as day 0) and subsequent isolates was annotated for each cluster.

Variant calling and identifying mutations in common genes in relapse lineages

The Snippy Bactopia subworkflow, which utilizes Snippy v4.6.0 (141), was used for variant calling of all isolates against a previously generated ancestrally reconstructed *S. aureus* genome (191). Variant calling was also conducted for each set of identified relapse infections by using the first known isolate (index) in each lineage as the reference. The index isolates were annotated using Bakta (v1.9.3)(192) in order to generate the reference sequence for variant analysis. Small insertions and deletions (indels) and single nucleotide polymorphisms (SNPs) were identified. The –snippy-core option was run on all isolates to identify variants in core coding regions across all genomes in the set. The snippy-core alignment produced an output with the predicted effects of variants on amino acid structure and potential functional changes. Any change that altered the amino acid compared to the reference was designated as non-synonymous, and if there was no change to the amino acid composition, the prediction was a synonymous change. Additionally, to identify commonly occurring non-synonymous changes within each recurrence lineage, the –snippy-core option was run to create a SNP alignment of the set of isolates from the same subject and identify the predicted effect of variants on amino acid structure and functional change. SNPs in coding regions were concatenated and summarized for all lineages.

Creating a database of bacteremia-associated mutations

A PubMed literature review was conducted to create a database of previously identified genes and/or mutations in *S. aureus* genomes that were associated with host bloodstream infections. Pubmed was searched in December 2023 using query (((staphylococcus aureus)) AND ((bacteremia) OR (bloodstream infection))) AND (genetic mutation). Only peer-reviewed articles (no reviews or preprints) were scanned for evidence. Genes or mutations in genes were considered if the study reported that they occurred in the *S. aureus* genome, samples were derived from bloodstream infections, were associated with changes in virulence expression or host survival or were associated with phenotypic changes related to blood cells, including immune cells. A collection of sources where these genes were identified can be found in Table S4.1.

Calculation of the index of neutrality

A McDonald-Kreitman (MK) test (193) was performed across all isolates associated with a relapse to assess the impact of selection on the whole genome or on known bacteremia-associated genes. A core genome alignment was created using snippy-core for all isolates and for all isolates that were associated with a relapse lineage. The ancestral reference served as the outgroup. Fixed sites were determined by counting the nucleotide sites that were universally conserved from the reference for all sequences when all sequences were included, and per relapse lineage. Polymorphic sites were determined if there were nucleotide sites that differed between individuals within a lineage. A G-test was used to evaluate the significance of the neutrality index.

Relationship between phylogenetic background and relapse/new infection

In cases where a patient suffered a recurrent new infection, we tested whether the second strain was likely to be of the same genetic background (CC or ST) as the first infection. We used a permutation approach, where we drew sequential isolates at random with the same frequency as seen among the patients with new infections. Random isolates were drawn from all the isolates within the sample population. The proportion of pairs that did not share a CC or ST compared to all pairs was calculated. We selected pairs for a total of 1000 drawings and the distribution of the non-clade sharing pairing proportions was compared to the observed number of recurrent infections that did not result in a subsequent relapse event. We also conducted a permutation test on whether the difference in collection date of the bacteremia isolate between recurrent and non-relapse pairs of infections was associated with sharing of clade background between infections. For all pairs of unrelated recurrences, the date difference between the first and next infection were calculated, and whether the two isolates shared the same ST or CC was noted. The mean date difference was calculated for pairs with shared and nonshared clade backgrounds. Subsequently, a permutation test was conducted on the mean date difference for 10,000 permutations.

Results

Recurrent bloodstream infections are from common clinical and phylogenetic backgrounds as other circulating strains.

We sequenced 456 *S. aureus* isolates from episodes of bacteremia and included 411 subjects, of which 32 subjects had at least one recurrent MRSA bacteremia episode (77 isolates total). Over the study period the number of new MRSA bacteremia episodes per month was consistent. (Figure 1). Recurrence-associated isolates occurred across phylogenetic clades including CC5 (N = 29), CC8 (N = 41), CC30 (N = 3), CC78 (N = 3) and CC72 (N = 1) (Figure 2). CC distribution was similar between recurrence and non-recurrence isolates (p=0.91), and there was no difference in the number of recurrences over time.

Demographic characteristics, healthcare exposures, comorbidities, and course of treatment of the first bacteremia episode of the patient were assessed to identify differences in the clinical risk factors that could contribute to the emergence of a recurrent strain of MRSA causing bacteremia. Only the length of time daptomycin was administered for the initial infection was significantly associated with a patient experiencing a recurrence compared to non-recurrence subjects, with a median time of 39 days compared to 21 days of therapy respectively (p = 0.05). Subjects with younger age, more cardiovascular disease and kidney disease, and healthcare-acquired community onset (HACO) acquisition were more frequent among recurrent subjects compared to non-recurrence patients, even if these factors did not attain p values < 0.05. Additionally, while there was no difference in the presence of a foreign body involved in the infection, removal of the foreign body was much more commonly occurring among non-recurrence subjects (37/80 subjects, 46%) compared to recurrence subjects (2/7 subjects, 29%). Taken together, recurrent bacteremia episodes share clinical and genetic characteristics with the overall population of infectious MRSA causing non-recurrent bloodstream infections and that larger population sizes will be needed to identify any variables with small effect.

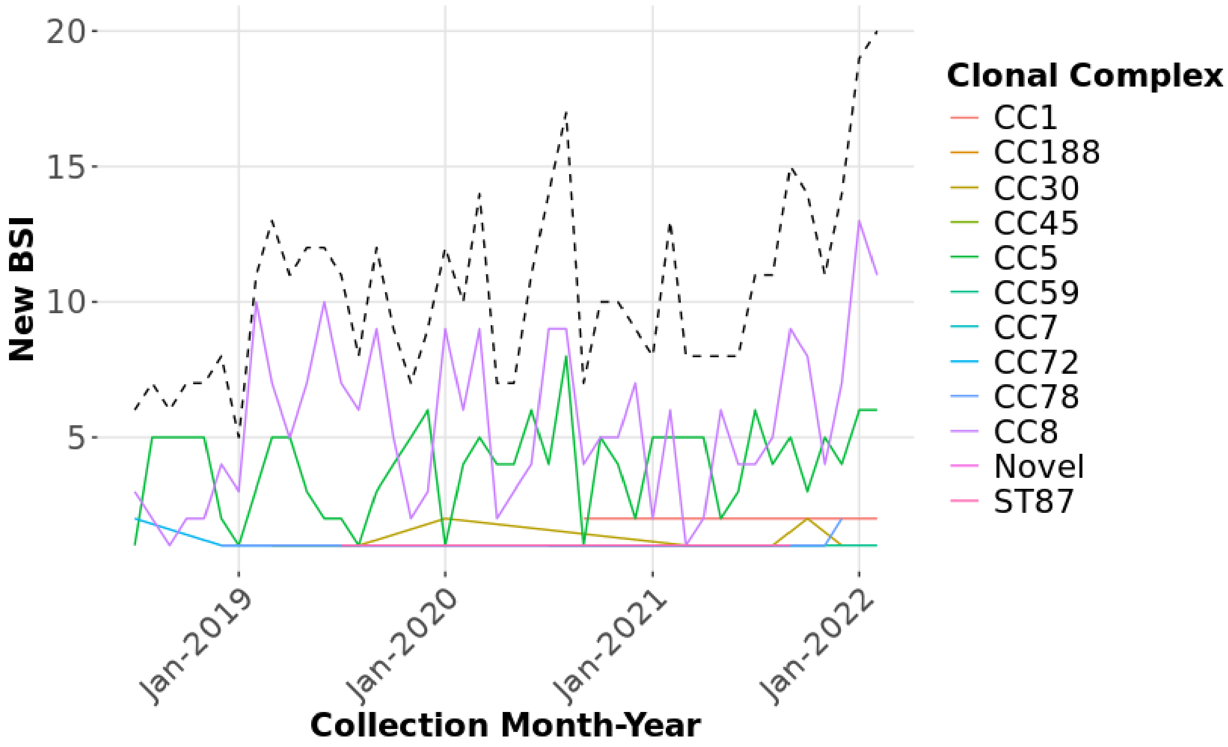

Figure 4.1. Case count of bacteremia over time by phylogenetic category. The total number of SAB episodes was counted from July 2018 to February 2022 and aggregated per month. Individual colored lines represent the clonal complex background of the isolate associated with each episode. The dotted line represents the total number of episodes over time during the study period (N=456).

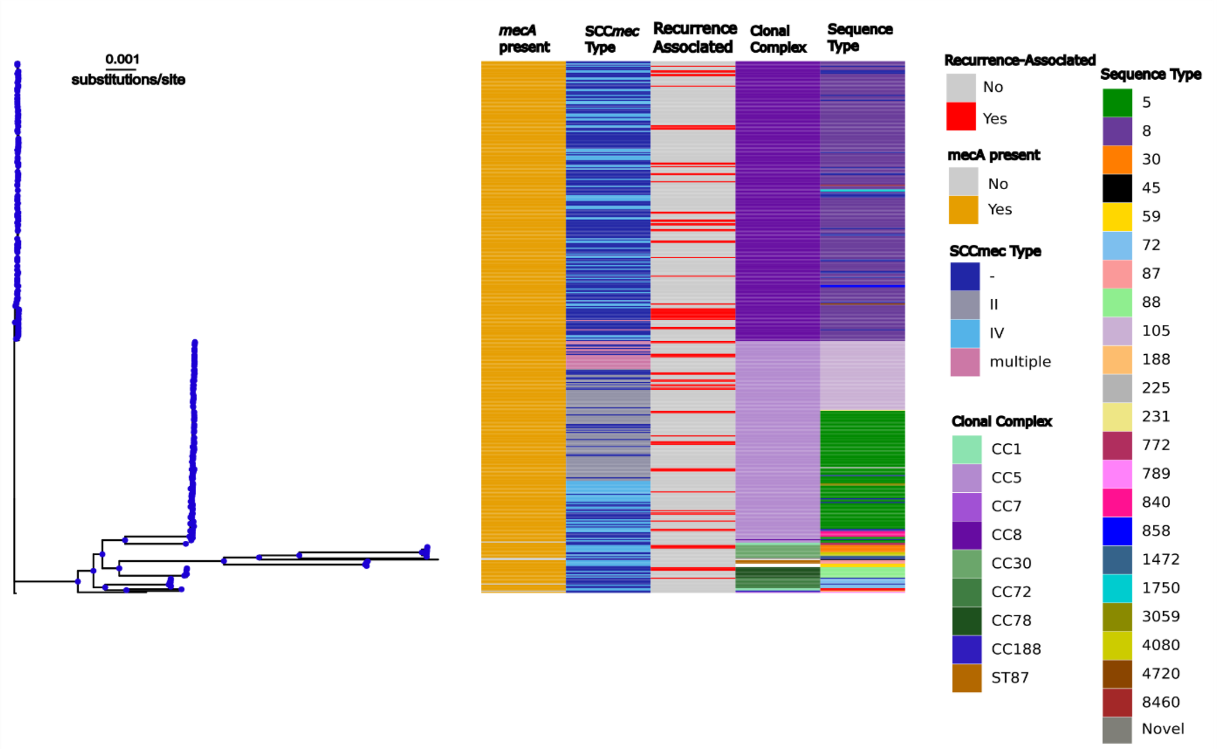

Figure 4.2. Recurrent SAB isolates share a similar phylogenetic distribution to non-recurrent SAB episodes. A core pangenome tree was constructed and rooted using MSSA strain Newman (GCF_020985245.1). Whether or not an isolate was a recurrence in a patient with a history of a previous infection is indicated in red in the first heat map column. Additional molecular characteristics included in the heat map are, from left to right, the presence or absence of *mecA*, *mecA* type, clonal complex, and sequence type. Blue dots indicated nodes where Shimodaira-Hasegawa approximate likelihood ratio test and ultrafast bootstrap values were >70).

Table 4.1. Clinical and demographic characteristics at the time of the first bacteremia episode for recurrent- and non-recurrent-episode subjects

Patient Attribute	Overall (N = 411)	Non-recurrent (N = 379)	Recurrent (N = 32)	P-Value
Age at Diagnosis (median, IQR)	56 (42.5-68)	57 (43-68)	51 (33-62)	0.06
Sex				
Male	232 (56%)	215 (57%)	17 (52%)	Ref.
Female	179 (44%)	163 (43%)	16 (48%)	0.71
Race and Ethnicity				
White	185 (45%)	168 (44%)	17 (52%)	Ref.
Asian	9 (2%)	8 (2%)	1 (3%)	0.59
Black	171 (42%)	161 (42%)	10 (33%)	0.31
Hispanic or Latino	10 (2%)	9 (2%)	1 (3%)	0.91
More than one race or ethnicity	9 (2%)	7 (2%)	2 (6%)	0.22
Other Race/ethnicity	10 (2%)	9 (2%)	1 (3%)	0.91
Don't know/refused	17 (4%)	17 (4%)	0 (0%)	Inf.
Chronic Skin Disease	36 (9%)	33 (9%)	3 (9%)	0.75
Diabetes	143 (35%)	130 (34%)	13 (39%)	0.56
Cancer	67 (16%)	64 (17%)	3 (9%)	0.33

Respiratory disease	95 (23%)	85 (23%)	8 (27%)	0.82
Cardiovascular Disease	199 (48%)	180 (47%)	19 (61%)	0.20
Infective Endocarditis	71 (17%)	67 (18%)	4 (12%)	
Liver Disease	41 (10%)	37 (10%)	4 (12%)	0.58
Kidney Disease	114 (28%)	102 (27%)	12 (36%)	0.22
Treated with Hemodialysis in the last 12 months?	77 (19%)	68 (18%)	9 (27%)	0.16
Current Intravenous Drug Use	87 (21%)	80 (21%)	7 (21%)	1
Involvement of foreign body in the bacteremia	87 (21%)	80 (21%)	7 (21%)	1
Was the foreign body removed?	39 (45%)	37 (46%)	2 (29%)	0.52
Healthcare Acquisition of index				
Community Acquired	50 (12%)	48 (13%)	2 (6%)	Ref.

Healthcare Acquired	92 (22%)	88 (23%)	4 (12%)	1
Healthcare Acquired-Community Onset	269 (72%)	243 (64%)	26 (82%)	0.28
Death in hospital or 30 days after index infection				
Yes	86 (22%)	86 (22%)	0 (0%)	<0.01
Unknown	14 (3%)	14 (4%)	0 (0%)	1
How many antibiotics was the patient exposed to during their index infection?	2 (1-2)	1 (1-2)	2 (1-3)	0.23
Vancomycin	383 (93%)	352 (94%)	31 (97%)	0.7
Vancomycin Duration (days)	21 (7-42)	20 (7-42)	24 (6 – 42)	0.65
Daptomycin	153 (37%)	136 (36%)	17 (55%)	0.08
Daptomycin Duration (days)	27 (8-42)	25 (7-42)	39 (21 – 42)	**0.05**

Recurrent but new infections are genetically distinct from relapse infections.

Using a genomic definition for relapse and recurrent but new infections, we identified 26 subjects with relapse infections and 8 with infections from a new strain; two subjects experienced both relapse and new infections. Relapses occurred sooner after the prior infection (Median 155 days, interquartile range (IQR) 88-269 days) compared to new strain recurrences (Median 248 days, IQR 105-599 days), though this was not statistically significant (Fig. 4.3A). Most relapse infections fell well below the set SNP threshold of 25 SNPs, with only 3 episodes requiring additional review for phylogenetic clustering, The six recurrent but new infections were distantly related from the subject's previous infection by hundreds to thousands of SNPs, suggestive of an evolutionary common ancestor well outside of a reasonable epidemiological time parameter (Fig. 4.3B). Indeed, when these pairs are compared in their phylogenetic context, they often are derived from wholly separate clades of the pangenome tree (Fig. 4.3C).

We were interested to identify the concordance between clinical definitions of relapse in the absence of genomic information compared to a strictly genomic definition. When pairs of bacteremia episodes were compared, the overall concordance was poor (Cohen's Kappa = 0.18, CI: -0.41), with the genomic definition predicting that 82% of subsequent infections are related to the previous infection, and the clinical definition predicting that 50% are related (Fig. 4.4). When genomics was used as a standard for relapse likelihood, the clinical definition has a sensitivity of 55% and a specificity of 75%. When a device or foreign body was implicated in any infection, however, there was often high concordance in identifying relapsing infections. The genomic definitions of relapse and new infections were used for the remainder of analyses.

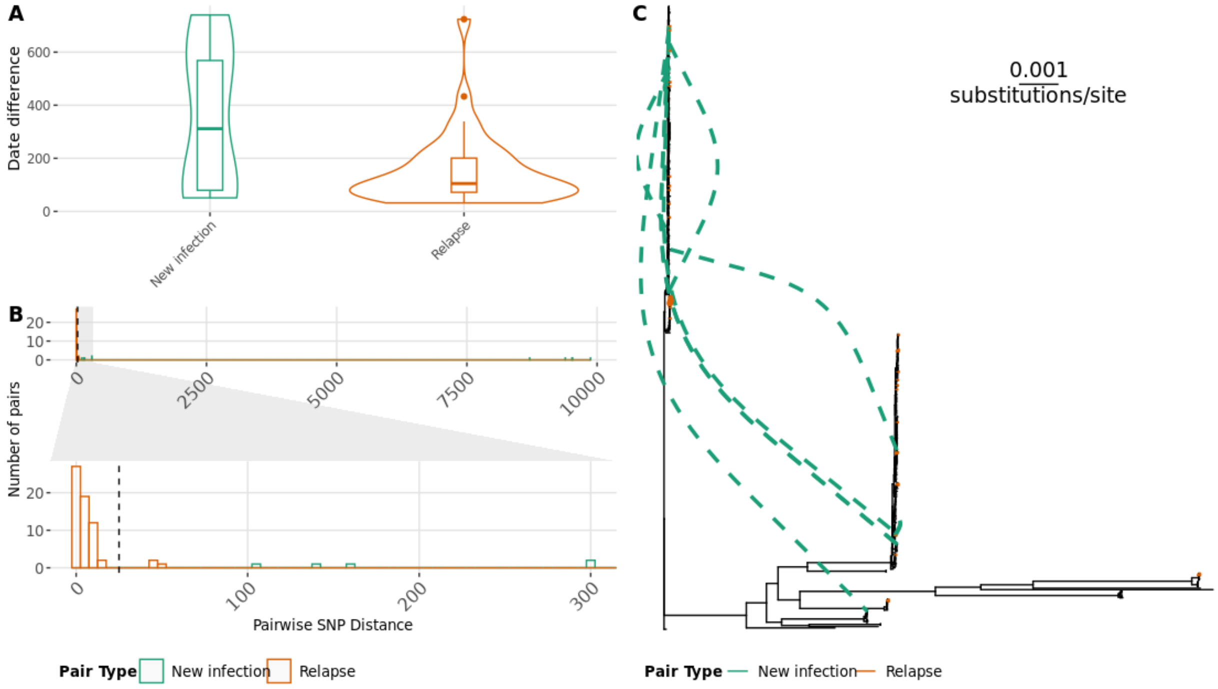

Figure 4.3. Relapsing infections and new infections within a patient are genomically distinguishable. Recurrence-associated isolates were separated into relapse-associated isolates or new infections based on pairwise SNP distance between isolate pairs within the subject. (A) Difference in time between subsequent episodes for genomically new infections and relapse associated infections were compared. (B) Counts of the number of pairs within an individual subject were assessed relative to their pairwise SNP distance. The inset display demonstrates counts where the SNP distance was between 0 and 350 SNPs. The dotted line represents the 25 SNP threshold for categorizing relapse infections. (C) A core-pangenome tree of all bloodstream isolates with recurrence-associated isolates linked by lines, with recurrent but new infections represented by green dotted lines and relapse infections represented by orange links.

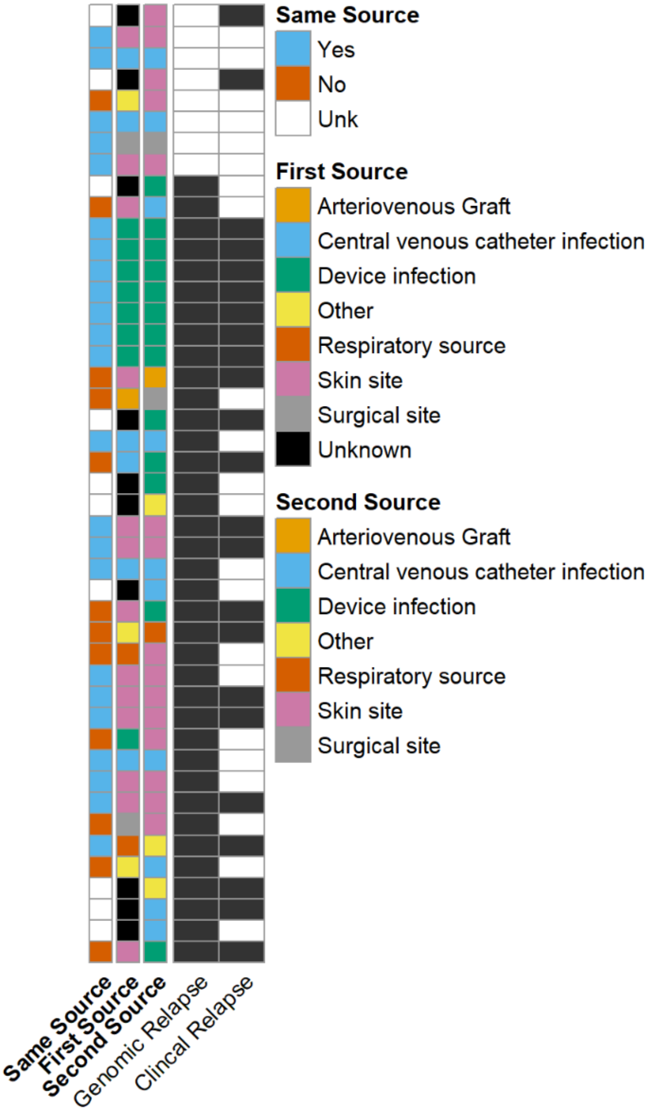

Figure 4.4. Clinical and genomic definitions of relapse are discordant. Pairs of isolates from all recurrent infections were compared and identified as relapse (filled black rectangle) or new infections (white rectangles) based on a genomic definition or clinical definition, for a total of 4X pairs. The suspected source type of each infection within the episode was identified, with the least recent isolate associated with the “First Source” and the most recent isolate associated with the “Second Source.” Clinical source type was additionally assessed to determine if the first and second source were physically the same source.

Relapse infections, but not new recurrent infections, demonstrate distinct adaptation to the host.

Among isolates that were recurrent but ultimately genomically distinct, subsequent re-infection with a strain of a different strain type or clonal complex was not more likely than by chance in the overall population. The difference in days of the subsequent infection did not differ from chance between isolate pairs that shared a phylogenetic background (ST or CC) compared to those with different backgrounds (Fig. 4.5). Taken together the genomic background of a previous MRSA bacteremia episode does not play a greater role than chance in determining strain a person may become infected with in a future new infection.

Relapse infections may be associated with long-term carriage of the strain on the body, which could result in adaptations different from populations that are cleared after the initial infection is treated. We utilized the McDonald-Kreitman neutrality index to examine whether there were significant signatures of adaptation overall and relative to known bacteremia-associated genes (Table 4.2). Across all bloodstream isolates, there is a general trend of neutral evolution across the genome and in the subset of genes specifically associated with bacteremia. Comparatively, relapse-associated lineages show a significant signature of positive selection in the whole genome. Although relatively few bacteremia-associated genes were synonymously mutated and comparable to all strains of relapsing lineages, there were notably no synonymous mutations at polymorphic sites. Together, this suggests that recurrent lineages likely undergo ongoing selection after the index infection and dissemination.

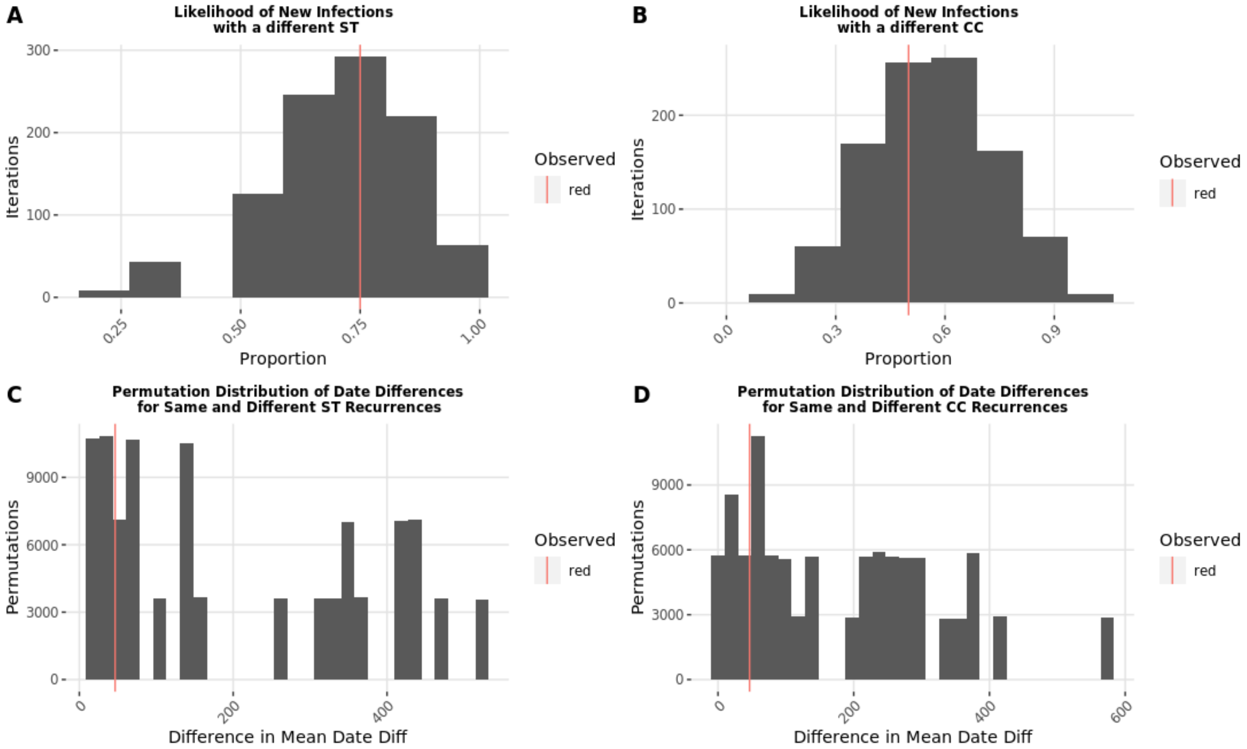

Figure 4.5. Previous strain background and time between infections does not impact the type of strain in recurrent but new infections. Two permutation tests were conducted to assess the likelihood of shared sequence type (ST) or clonal complex (CC) between distantly related recurrent infections. Eight pairs of isolates were sampled from the 456 isolate sample population 1000 times and the proportion of pairs with matching STs (A) or CCs (B) was calculated for each iteration. The observed proportion in the sample population is indicated by a red line. The difference between mean number of days between infections with or without a shared ST (C) or CC (D) were compared over 1 x 105 permutations. The observed difference in mean date duration is plotted with a red line.

Table 4.2. Neutrality index calculations of the whole genome and bacteremia-associated genes for all bacteremia isolates and for relapse-associated lineages.

Group	Fixed, NS	Fixed, S	Polymorphic, NS	Polymorphic, S	MK Value	P-value (G-test)
All Isolates - Whole Genome	17271	24529	1381	1812	0.92	**0.03**
All Isolates - Bacteremia genes	232	330	19	22	0.81	0.53
Relapse - Whole Genome	2689	5326	45	12	0.133	**> 0.01**
Relapse - Bacteremia genes	28	81	5	0	–	–

Antibiotic resistance genotypes and phenotypes in relapses correspond with patient exposures.

To identify potential genes involved in convergent adaptation within the host, genes that contained non-synonymous SNPs in more than one relapse lineage were identified. A total of 11 genes with unique SNPs had mutations in 2 or more relapse lineages (Fig. 4.6A). Mutations in these genes occurred regardless of clonal complex, indicating that clade background alone did not contribute to the presence of mutations in these genes. Genes in which multiple relapse lineages showed non-synonymous mutations were implicated in known virulence traits and antibiotic resistance. The genes most commonly mutated were *mprF*, among five separate subject lineages, and *rpoB,* among four subject lineages. Since changes in both *mprF* and *rpoB* are associated with drug resistance for treatments commonly used for bacteremia, we investigated the specific amino acid change, isolate MIC, and course of treatment for the patient at the time of the bacteremia episode. Multiple different amino acid changes were detected between patients for proteins encoded by both genes. Notably the same amino acid change occurred in two subjects with elevated rifampin resistance, Ala477Asp, though for one subject the resistance phenotype was present in their first infection before relapse (Fig. 4.6B). Only one patient with a *rpoB* mutation had neither a history of rifampin exposure nor the emergence of a resistance phenotype. Three subjects with *mprF* mutations during relapses demonstrated acquired daptomycin resistance alongside previous exposure to daptomycin (Fig. 4.6C). These included amino acid changes Ser337Thr, Ser337Leu, and Leu291Ile. All patients with *mprF* mutations had exposure to daptomycin prior to the emergence of their mutation regardless of the emergence of daptomycin resistance.

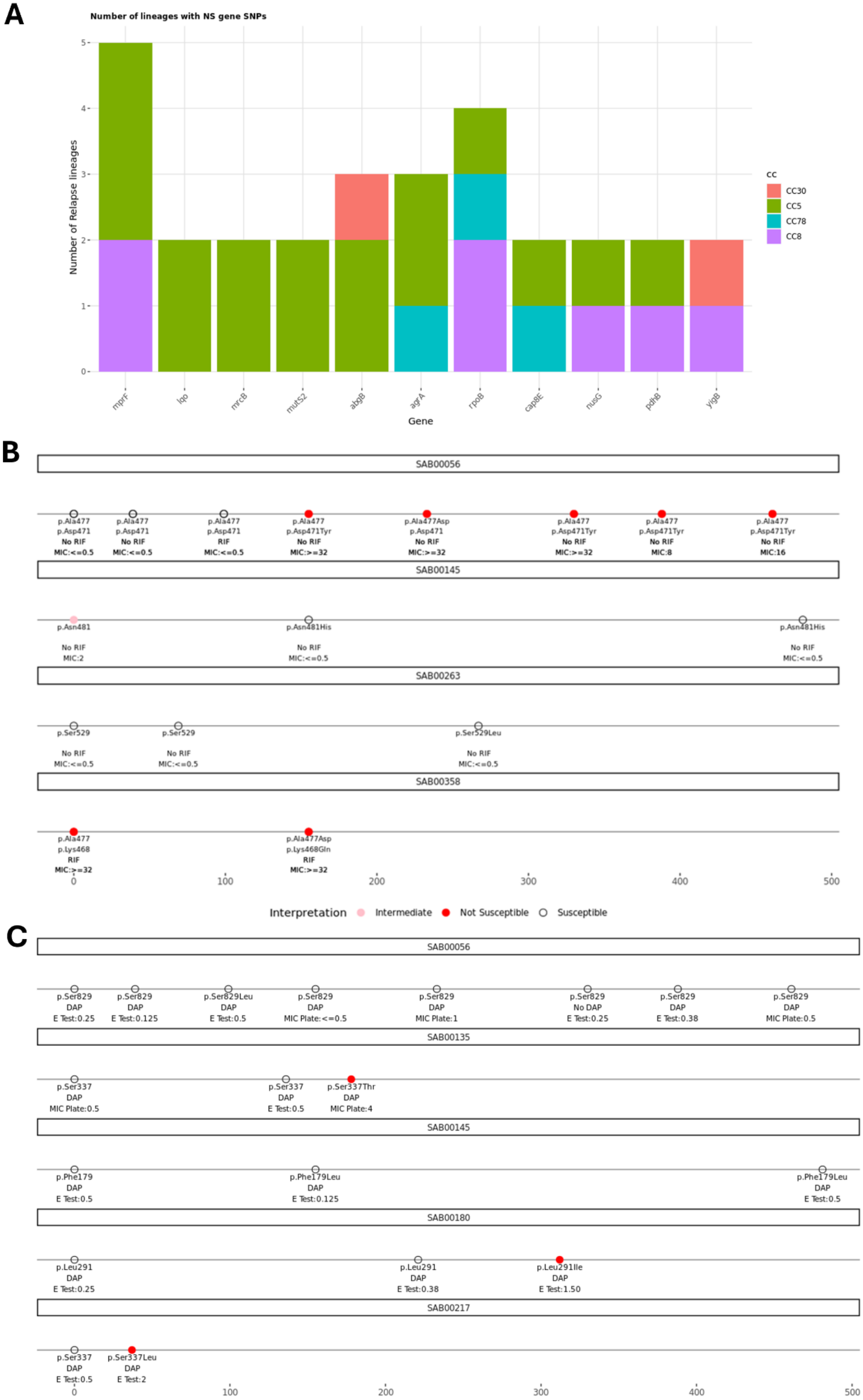
A
Number of lineages with NS gene SNPs
Number of Relapse lineages
0
1
2
3
4
5
mprF
lqo
mrcB
mutS2
abgB
agrA
rpoB
cap8E
nusG
pdhB
yigB
Gene
cc
CC30
CC5
CC78
CC8
B
SAB00056
p.Ala477 p.Asp471 No RIF MIC:<=0.5
p.Ala477 p.Asp471 No RIF MIC:<=0.5
p.Ala477 p.Asp471 RIF MIC:<=0.5
p.Ala477 p.Asp471Tyr No RIF MIC:>=32
p.Ala477Asp p.Asp471 No RIF MIC:>=32
p.Ala477 p.Asp471Tyr No RIF MIC:>=32
p.Ala477 p.Asp471Tyr No RIF MIC:8
p.Ala477 p.Asp471Tyr No RIF MIC:16
SAB00145
p.Asn481 No RIF MIC:2
p.Asn481His No RIF MIC:<=0.5
p.Asn481His No RIF MIC:<=0.5
SAB00263
p.Ser529 No RIF MIC:<=0.5
p.Ser529 No RIF MIC:<=0.5
p.Ser529Leu No RIF MIC:<=0.5
SAB00358
p.Ala477 p.Lys468 RIF MIC:>=32
p.Ala477Asp p.Lys468Gln RIF MIC:>=32
0
100
200
300
400
500
Interpretation
Intermediate
Not Susceptible
Susceptible
C
SAB00056
p.Ser829 DAP E Test:0.25
p.Ser829 DAP E Test:0.125
p.Ser829Leu DAP E Test:0.5
p.Ser829 DAP MIC Plate:<=0.5
p.Ser829 DAP MIC Plate:1
p.Ser829 No DAP E Test:0.25
p.Ser829 DAP E Test:0.38
p.Ser829 DAP MIC Plate:0.5
SAB00135
p.Ser337 DAP MIC Plate:0.5
p.Ser337 DAP E Test:0.5
p.Ser337Thr DAP MIC Plate:4
SAB00145
p.Phe179 DAP E Test:0.5
p.Phe179Leu DAP E Test:0.125
p.Phe179Leu DAP E Test:0.5
SAB00180
p.Leu291 DAP E Test:0.25
p.Leu291 DAP E Test:0.38
p.Leu291Ile DAP E Test:1.50
SAB00217
p.Ser337 DAP E Test:0.5
p.Ser337Leu DAP E Test:2
0
100
200
300
400
500

Figure 4.6. Commonly mutated genes among relapse lineages are associated with antibiotic resistance phenotypes. (A) Unique non-synonymous mutations in coding regions of the genome were quantified by gene from isolates of relapse-associated infections and summarized relative to the subject from which the isolate was collected. Each unique mutation was annotated with the clonal background of the lineage from which the set of relapses were derived. For lineages with *rpoB* mutations (B) and *mprF* mutations (C), a timeline (days since index infection) was created for each set of relapsing infections by the subject experiencing that set of relapses. Individual episodes were annotated with the amino acid changes detected in the respective genes, the clinical assay used to assess minimum inhibitory concentration (MIC) and the corresponding MIC, and whether the patient was exposed to rifampin (RIF)(B) or daptomycin (DAP)(C). Dots are colored based on the clinical assay determination of drug susceptibility to RIF or DAP.

Relapse isolates cluster with other bacteremia isolates, but do not contribute to onward transmission.

To determine the burden of recent transmission between patients with relapse infections and other bacteremia patients we identified genomic clustering of isolates containing fewer than 25 SNPs difference. Nine clusters with at least one subject with a relapse were identified. These clusters occurred in distinct lineages across CC5 and CC8 clades (Fig.4.7). If relapse infections were contributing to onward transmission, or if patients were becoming reinfected with a closely related circulating strain, we might expect that infections between hosts would cluster within the relapse lineage clade. Across all nine clusters, isolates from different subjects clustered significantly outside of the relapse lineage isolates. This suggested that transmission events occurred prior to the onset of relapsing infection, and that the unique lineages within a host were highly specific to the individual.

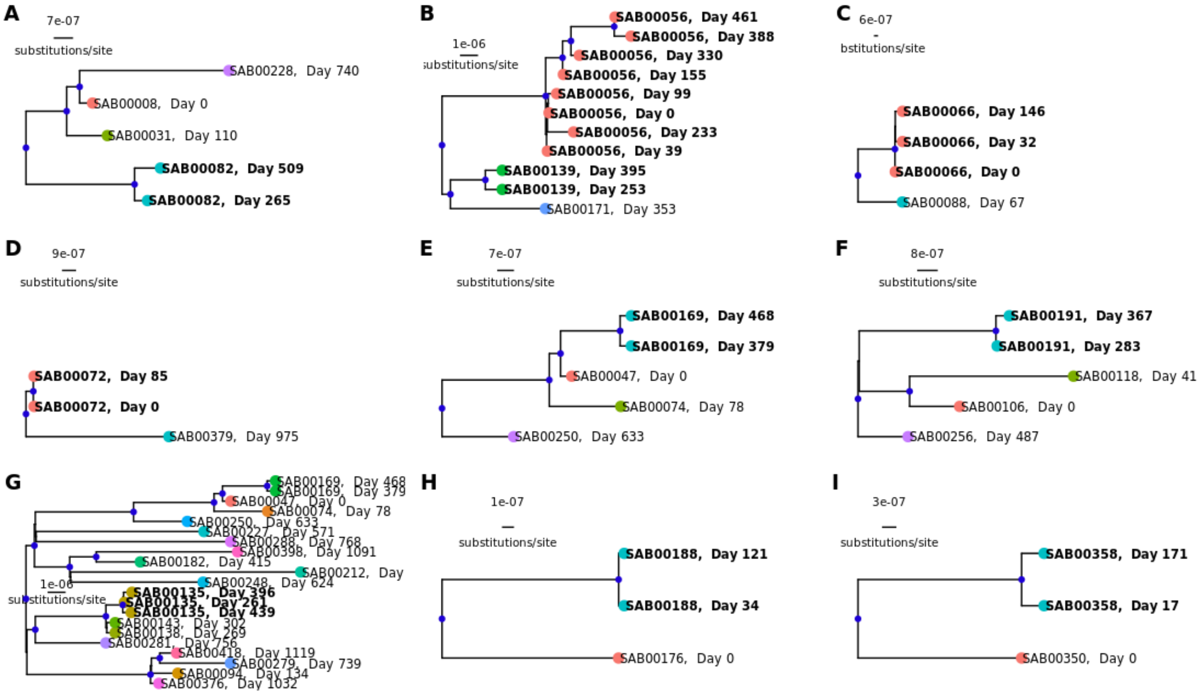

Figure 4.7. Non-relapse associated isolates cluster separately from relapse-associated isolates. A pangenome tree was generated using the unequal transition/transversion rate plus empirical base frequencies model to investigate branching positions. Clusters were investigated when at least one relapse-associated isolate was 25 SNPs or fewer from an isolate from a different subject. Subtrees were extracted based on the most-recent common ancestor shared by relapse subject isolates and the clustered additional subjects. Nodes denoted with a blue dot indicate ultrafast bootstrap values and Shimodaira-Hasegawa approximate likelihood ratio test values that are greater than 70. Bolded tip labels indicate an isolate that is part of a relapse. Tips are annotated with the patient subject IDs and the number of days at which the isolate was collected relative to the earliest isolate in the cluster. Branch lengths are scaled in substitutions per site.

Discussion

This research shows that recurrent SAB is well differentiated into new and relapse infections by genomic analysis and patterns of pathogen evolution. Relapsing strains show ongoing adaptive mutations in bacteremia genes, especially those also associated with antimicrobial resistance. Further, strains of bacteremia in new reinfections are not determined by previous exposure to the same strain, suggesting that host adaptive immunity may not play a strong role in preventing new infections from the same strain, or that individual patients are sensitive to infection with a specific strain. Finally, although relapses of MRSA bacteremia do occur in transmission clusters, they are not directly contributing to ongoing spread in healthcare settings after their index infection.

We demonstrated that relapses and new infections are often distinguishable by their genomic distance, source site of infections, and by detection of adaptive traits, especially those associated with antibiotic resistance. Previous studies have identified risk factors for recurrence from the pathogen level to the type of care received. We identified similarities in our dataset that have been recorded previously, including a high prevalence of SCC*mec*II (167), a high proportion of unremoved foreign objects (165,166), increased cardiovascular disease including endocarditis (168), increased hemodialysis (166). The interaction of the variables of Black race and hemodialysis has also been implicated in a higher incidence of relapse SAB (165). Our study did not show any significant association between race and recurrence; nevertheless, over a third of the subjects in this study experiencing recurrence identified as Black. Considering these previously documented patterns, consideration of the interaction of community demographics and the prevalence of chronic conditions among social groups served by common healthcare groups is critically important to monitor to prevent bacteremia and subsequent relapse. We found a significant association between relapse and longer duration of daptomycin therapy. However, other researchers have reported a negative or no association between antibiotic duration and the risk of relapse (33,194). These discrepancies may be explained by the specificity of the type of antibiotic or by confounding conditions that prolong the use of a certain antibiotic, such as a diagnosis of endocarditis.

We found that most isolates from the same person were separated by fewer than 25 core genome SNPs, and the few that did not fall within our definition of relapse were evolutionarily distinct enough that recent common ancestry within a reasonable infection timescale was highly unlikely. Previous studies have identified a similar pattern in distinguishing relapses from recurrences even using less precise technologies such as pulsed-field gel electrophoresis (168). Choi et al identified a split between 45 and 54% between new infections and recurrence based on PFGE pattern and infections with a difference no greater than 150 days, and they found that WGS aligned with their original molecular definition (165). In our study, we used a SNP based definition for clustering and found a much larger ratio of relapse infections comparatively. We also found that although there was no significant difference in time between infections in relapses compared to recurrent infections from new strains, relapse intervals were still much shorter overall and most date differences under 200 days.

Examination of the SNP in its individual gene context and phylogenetic context can be extremely useful to help discern the likelihood of persistence as a cause of relapse, especially when there may be a concern for infection from external but closely related strains. In this study, we examined the emergence of unique mutations within a lineage of relapses, as well as examined the branching patterns between individuals that shared a common strain. When comparing those relapses that cluster very closely with other hosts, we observe unique traits (i.e., daptomycin resistance), which would suggest that the most parsimonious explanation of their occurrence among within-host isolates is a host-associated lineage rather than a unique clone seeding back into the patient. Though we are still not able to rule out entirely that an external source carrying a common strain could nevertheless cause reinfection in patients who encounter these sources, the nesting of the infections and the genetic changes suggest a higher likelihood of specific host-associated exposures related to treatment and possible persistent colonization.

Because SAB may result from a compounding set of risks and exposures, management of SAB and recurrence typically involves a comprehensive assessment of patient history, physical examination, and source identification (157). Source control is important, as delays in the removal of a contaminated source of infection can increase the risk of persistent bacteremia (157) or metastatic spread to other body sites (195). When central lines or foreign bodies are suspected to be the cause of the SAB, removal of the foreign body is considered, though it is not always possible if it would lead to increased morbidity or mortality for the patient. In this study we found genomically similar episodes of SAB common among patients with foreign body infections. Using genomic analysis as gold standard for relapse increased the association with foreign body exposure. Further, we had previously reported that the odds of MRSA infections associated with a foreign body was nearly five times greater among patients that had a previous MRSA infection within a year compared to those that did not have a previously reported infection (196). Although it may not always be possible to remove a foreign body to eliminate relapse infections, clinicians should maintain high suspicion of devices and implants as a source of recurrent SAB and advise patients accordingly in their post-recovery of a known SAB.

Two genes known from previous studies to be commonly associated with antimicrobial resistance, *mprF* and *rpoB*, gained non-synonymous mutations multiple times across relapses occurring in separate subjects. Changes to *rpoB* and *mpfF* have been implicated in resistance to rifampin and daptomycin, and also potential cross-resistance to vancomycin treatment through multiple multistep evolutionary pathways (160,179,197,198). In our study, nearly all subjects, and all but one subject that experienced relapse, received vancomycin as a part of their course of treatment for their first infection. Multiple different amino acid changes were detected in isolates with a rifampin-resistant phenotype, both in the index infection and in acquired resistance over the course of resistance. Among these isolates, we detected one relapse lineage in which the index isolate carried asparagine at position 481 and which had a rifampin intermediate resistance profile, followed by relapses with a change to histidine. The change of asparagine at this position to histidine and subsequent susceptibility to rifampin is consistent with the opposite change from histidine to several other amino acids leading to phenotypic rifampin resistance (199). In two subjects, at least two different changes at multiple sites of the rpoB amino acid sequence were altered between episodes, with intermittent reappearance of the index allelic profile. One subject's isolates also carried the Ser529Leu mutation, known to be associated with vancomycin intermediate resistance (VISA) and heteroresistance (hVISA), but developed no change in rifampin or vancomycin resistance. We also noted emergence of phenotypic rifampin resistance, notably with one subject carrying a resistant clone in their original infection and no noticeable change to the MIC in rifampin as the alleles changed. Multiple simultaneous mutations in *rpoB* have been associated with an increased resistance to rifampin (199). The amino acid changes present in this study demonstrate the wide diversity in the mutational profile of *rpoB*. Additional phenotypic investigation of phenotypic cross-resistance and combination therapy on the emergence of these mutations is needed.

We also identified two point mutations in *mprF* that resulted in changes at the same amino acid site (Ser337Thr and Ser337Leu) within two patient lineages which corresponded with emergence of phenotypic resistance to daptomycin. For one of these lineages (Ser337Thr), a corresponding intermediate resistance to vancomycin was also recorded without the emergence of any known genetic markers for VISA or hVISA, though we simultaneously detected a stop gained in the protein nusG, a known global transcription regulator (200). In this sequence of isolates, we did not capture possible intermediary isolates that could suggest whether there was a specific sequence or evolutionary pathway that could suggest the order in which resistance and intermediate phenotypes were gained. Since patterns of two-step evolution of resistance phenotypes in *S. aureus*, for example by regulation of *walK* and then gain of function of membrane charge increase for *mprF* (201), a possible hypothesis here could be a similar stepwise change in gene expression elsewhere followed by changes to membrane charge.

This study is subject to several limitations. The incidence of recurrence among bloodstream infection episodes in our study was 9.8%, which is consistent with the rate seen in other studies (159,165,167,168). Nevertheless, the number of recurrent infections we detected is a limited sample for generalization. We also characterized cases from a singular geographic area. The smaller sample size and unique demographic structure of this area could make it difficult to detect the same risk factors associated with demographics or clinical factors.

Whole-genome sequencing was able to provide additional support for case-identification of new and relapsing infections in the context of the clinical history of previous infections. The complexity of host-associated factors, within-host selection pressures, strain background and type of antibiotic treatment all play interacting roles in the emergence of relapse, but specific mutations can help create stronger evidence for how best to identify the patterns of host and pathogen factors that lead to relapse for unique individuals. Relapsing isolates undergo positive adaptation to the host, and convergently mutating genes are consistent with long-term usage or high exposure to antibiotics, but other traits necessary for survival in the cardiovascular system may still play an important role in persistence. Ongoing work is necessary to understand the length of survival of individual strains of

S. aureus at different body sites on individual hosts. Uniting the frequency of genetic mutations, genetic relatedness, and known clinical risk factors for recurrence lay the groundwork for better prediction of future relapse.

Supplemental Material

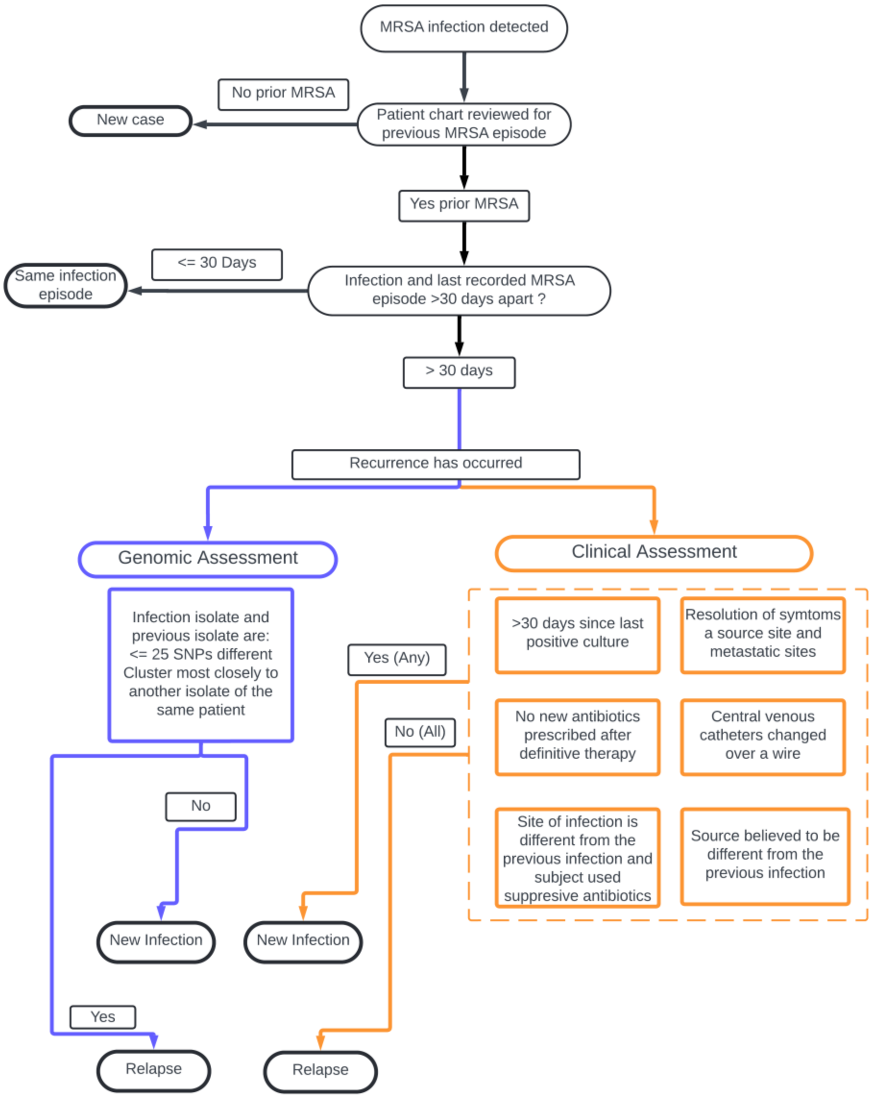
MRSA infection detected
No prior MRSA
New case
Patient chart reviewed for previous MRSA episode
Yes prior MRSA
<= 30 Days
Same infection episode
Infection and last recorded MRSA episode >30 days apart ?
> 30 days
Recurrence has occurred
Genomic Assessment
Clinical Assessment
Infection isolate and previous isolate are: <= 25 SNPs different Cluster most closely to another isolate of the same patient
>30 days since last positive culture
Resolution of symtoms a source site and metastatic sites
Yes (Any)
No (All)
No new antibiotics prescribed after definitive therapy
Central venous catheters changed over a wire
No
Site of infection is different from the previous infection and subject used suppresive antibiotics
Source believed to be different from the previous infection
New Infection
New Infection
Yes
Relapse
Relapse

Figure S4.1: Clinical and Genomic Criteria for defining Recurrent new and relapsing infections. The flowchart is formatted as a decision tree with round ovals indicating terminals for branch points and rectangles indicating options that correspond with the most recent terminal. Black lines indicate criteria assessed regardless of the definition. Blue lines indicate criteria assessed for genomic definitions, and orange lines indicate criteria that fulfill the clinical definitions.

Table S4.1: List of Bacteremia-associated genes and corresponding supporting studies

Gene	**Product**	**Isolate Source**	**Allelic or mutational change detected**	**Phenotype of mutation**	**Literature Support**	**Literature refute/no evidence**
ACME (arc and opp3)	arginine catabolic mobile element	clinical isolates		Presence increased pathogenicity in bacteremia model (rabbits)	Diep 2008a (202)	
Agr	virulence regulator	clinical isolates		Increased genetic diversification compared to agr+ strains and colonizers; dysfunction leads to increased VAN MICs	Altman 2018b(174), Cheung 1994 (176), Tsuji 2009 (175), Chong 2013 (177)	

AgrA		clinical isolates	Non-synonymous, truncation	High-level rifampin resistance	Hachani 2023 (171), Giulieri 2018 (170), Benoit 2018 (172)	Howden 2008, (83)
AraC	AraC family transcriptional regulator	clinical isolates	premature stop	Untested, but present in the BSIs separate from nasal carriage	Young 2012 (203)	
ausA		clinical isolates	Non-synonumous, truncation	Escape from epithelial cell endosomes	Hachani 2023 (171)	
clpX		clinical isolate	Non-synonymous	Reduction of expression of virulence	Baek 2015 (179)	
cna	collagen-binding adhesin	clinical isolates	non-synonymous	Decreased attachment to collagen	Iwata 2020 (204)	
coa	coagulase	laboratory strain	deletion	Loss of coagulase function,	Liu 2021 (73), Altman 2018 (174)	

				decreased virulence		
dfrB		Clinical isolates	Non-synonymous	Trimethoprim resistance	Young 2021 (183)	
edinB	epidermal cell differentiation inhibitor	laboratory strains		Presence of gene increases ADP-ribosylation, increased prevalence of bacteremia during pneumonia and bacterial load	Courjon 2015 (205)	
ess	virulence regulator	clinical	insertion sequence	Increased expression of ESAT^-like secretion system virulence	Altman 2018 (174)	

				factors (when agr-)		
essB	ESAT-6 secretion system component	laboratory strains ST398		Deletion results in decreased neutrophil killing and lethality in blood	Wang 2016 (206)	
fnbA	fibronectin-binding protein A	clinical isolates	Non-synonymous SNPs	Enhanced binding to Fn, associated with cardiac device infections from bacteremia isolates[62]	Hos 2015 (69)	
fusA		clinical isolates		Fusidic acid resistance by target alteration,	Lannergard 2009 (207)	

				some small evidence that one isolate arouse from a SCV phenotype		
fusB		clinical isolates		Fusidic acid resistance by protecting translation apparatus	Lannergard 2009 (207)	
fusC		clinical isolates		Fusidic acid resistance by protecting translation apparatus	Lannergard 2009 (207)	
hglABC		laboratory strians		Presence assists with survival in blood	Malachowa 2011 (208)	
ica	intracellular adhesin locus	laboratory strains		Loss of function shows a	Kropec 2005 (209)	

	conferring poly-N-acetylglucosamine production			greater susceptibility to Ab-dependent killing by leukocytes		
katA	catalase enzyme	clinical isolate	stop codon and truncation	Loss of catalase activity but still results in septic arthritis	Lagos 2016 (210)	
lukED	leucotoxin ED	laboratory strains		Presence target murine phagocytes leading to cytotoxic effects at infection site	Alonzo 2011 (211)	
mgrA	virulence regulator	laboratory strains		Association of loss of function leads to	Li 2019 (212), Rom 2017 (213)	Howden 2008 (83)

				increased susceptibility to host defense response cells via *mprF* and *dltA* expression, possibly loss of *fnbA* expression, and general decreased impact on host health in a bacteremia model; mutation leads to increased virulence in a		

				mouse model		
mprF	cell membrane structure	clinical isolates	non-synonymous, deletions,	Increased daptomycin resistance	Ji 2020 (214), Baek 2015 (179), Chen 2015 (197)	
mpsB	cation translocation in cell membrane	clinical isolates; laboratory strains		Small colony phenotype, suppression of agr activation due to lowered membrane potential	Douglas 2021 (181)	
mspA	membraine protein	clinical and laboratory		Role in toxin production, resistance to innate immue cells, and iron homeostasis	Duggan 2020 (215)	

parC	topoisomerase IV	clinical isolate	insertion	confers quinolone resistance (eg CIP)	Gao 2010 (158)	
psm-mec	phenol-soluble modulin alpha type	clinical isolates	promoter SNP	Decreased biofim formation and increased PMSa3 and Hld expression	Aoyagi 2014 (216)	
purR	purine biosynthesis regulation, and regulation of fibronectin binding protein	laboratory strain	Non-synonymous snp	Increased clumping in blood related to fibronectin binding, increases in SarA expression which has other	Goncheva 2019 (217), Alkam 2021 (218)	

				virulence factor down stream events		
PVL (lukS/F)	Panton-Valentine leukocidin	laboratory strains		Increased pathogenesis in early stages of bacteremia	Diep 2008ab (219)	
rel	synthesis of (p)ppGpp during AA starvation	clinical isolates	Non-synonymous (D134Y, A301T, E384K, V670G), and the fifth is a 4-bp deletion encompassing codon N697 that results in a frameshift causing a premature stop codon at position 701; Gao showed a Phe 128 Tyr	Shortened lag phase, increased fitness in nutrient-poor conditions; increased resistance to antimicrobials and defensins; possibly related to agr	Chen 2023 (220), Bryson 2020 (22), Gao 2013 (160), Gao 2010 (158)	

			substitution from a nucleotide sub)	upregulation for Gao 2010		
rlmN	ribosomal RNA large subunit	clinical isolate	insertion	Confers linezolid resistance, uniquely from previous 23 rRNA mutations	Gao 2010 (158)	
rot	regulation	laboratory strain		Change in sepsis virulence (ie survival) in mice, background dependent	Rom 2021 (71), Rom 2017 (213)	Howden 2008 (83)
RpiRc		laboratory strain (USA300-LAC)	altering protein expression	Repressed RNAIII to mimick rot deletion, leading to	Balasubramanian 2016 (221)	

				bloodstream infection phenotype		
rpoB	RNA polymerase		Non-synonymous		Giulieri 2018 (170), Baek 2015 (179), Gao 2013 (160), Gao 2010 (158), Villar 2011 (178)	
rpoD (sigA)	RNA polymerase sigma factor	laboratory strain	excision of an IS256 element	Decreased capacity to infect bone and increased virulence to mouse	Suligoy 2020 (222)	
rsp	transcription factor repressor	clinical isolates	Non-synonymous and premature stop codon	Reduced lethality,	Das 2016 (223)	

	of surface protein			reduced cytotoxicity		
SaeRS	regulatory system	laboratory strain	deletion	Increased survival in mouse blood; in balance with sarA protease production for virulence; in lab, lack of SaeRS increases mortality, but down regulation of leukosidins and immunomodulatory genes, and some	Liu 2021 (73), Beenken 2014 (72), Nygaard 2010 (173)	Voyich 2009 (224)

				adhesion genes)		
sar	virulence regulation	laboratory strain		Loss of virulence factors leading to loss of attachment to heart valves; loss of virulence that is connected to protease-mediation	Cheung 1994 (176), Beenken 2014 (72), Zielinska 2012 (225)	
SCCMec type IV	MGE harboring mecA	clinical isolates		Higher association with bacteremia or central line infections	Nakano 2022 (182), Young 2021 (183)	

selw	production of SAg SelW	clinical isolates; laboratory strains		Superantigen activation of T-cell proliferation	Vrieling 2020 (226)	
srtA	sortase A	laboratory strains		Deletion results in reduced mortality and dissemination to tissues after introduction to the bloodstream in an injection introduction	Wang 2015 (227)	
stp					Giulieri 2018 (170)	
thyA	thymidylate synthase gene	clinical isolate	premature stop	Small colony phenotype, related to	de Souza, 2020 (180)	

	needing THF cofactor			adaptation of trimethoprim-sulfamethoxazole resistance?		
xerC	recombinase	laboratory strain (USA300-LAC and UAMS-1)		Mutation results in decreased bacterial load in murine bacteremia, decreased accumulation of alpha toxin decreased biofilm production	Atwood 2016) (228)	
yycH	modulation/downstream of walK	clinical isolates	frameshift, premature termination	VAN and DAP nonsusceptibility,	Chen 2015 (197)	

	which is important for cell wall metabolism			mechanism unknown or multifaceted		
	Wall teichoic acids	laboratory strain		Loss of function results in less adherence to endothelial cells and proliferation to spleen and kidneys	Weidenmaier 2005) (229)	

Chapter 5: Metagenome-wide characterization of shared antimicrobial resistance genes in sympatric people and lemurs in rural Madagascar

Reprinted material from: Talbot, B. M., Clennon, J. A., Rakotoarison, M. F. N., Rautman, L., Durry, S., Ragazzo, L. J., Wright, P. C., Gillespie, T. R., & Read, T. D. (2024). Metagenome-wide characterization of shared antimicrobial resistance genes in sympatric people and lemurs in rural Madagascar. *PeerJ*, *12*, e17805. https://doi.org/10.7717/peerj.17805

Abstract

Background. Tracking the spread of antibiotic resistant bacteria is critical to reduce global morbidity and mortality associated with human and animal infections. There is a need to understand the role that wild animals in maintenance and transfer of antibiotic resistance genes (ARGs). **Methods.** This study used metagenomics to identify and compare the abundance of bacterial species and ARGs detected in the gut microbiomes from sympatric humans and wild mouse lemurs in a forest-dominated, roadless region of Madagascar near Ranomafana National Park. We examined the contribution of human geographic location toward differences in ARG abundance and compared the genomic similarity of ARGs between host source microbiomes. **Results.** Alpha and beta diversity of species and ARGs between host sources were distinct but maintained a similar number of detectable ARG alleles. Humans were differentially more abundant for four distinct tetracycline resistance-associated genes compared to lemurs. There was no significant difference in human ARG diversity from different locations. Human and lemur microbiomes shared 14 distinct ARGs with highly conserved in nucleotide identity. Synteny of ARG-associated assemblies revealed a distinct multidrug-resistant gene cassette carrying *dfrA1* and *aadA1* present in human and lemur microbiomes without evidence of geographic overlap, suggesting that these resistance genes could be widespread in this ecosystem. Further investigation into intermediary processes that maintain drug-resistant bacteria in wildlife settings is needed.

Introduction

The global estimated number of human deaths attributed to antibiotic-resistance among bacterial infections in 2019 alone was 1.27 million, with Sub-Saharan African countries experiencing the highest proportion of the burden (25). Antimicrobial resistance in bacteria is a heterogeneous problem, with multiple organisms, biological mechanisms, and anthropogenic activities contributing to its presence and spread. Pathogen spread is known to play a major role in antimicrobial resistance gene (ARG) distribution, with evidence of enteric infections among symptomatic humans and animals having less antibiotic susceptibility compared to asymptomatic individuals (230). Bacteria can acquire antibiotic resistance through *de novo* mutations, but they may also acquire resistance through horizontal gene transfer on mobile genetic elements (MGEs). MGE movement through a bacterial community depends on the species present, as MGE sharing can be restricted by species compatibility and host range (231–233), but the presence of ARGs and its transference into closely related species can facilitate epidemic spread of pathogens (234). Nearly all bacterial pathogens associated with infectious diseases have been found to contain antimicrobial resistance genes, so it becomes imperative to capture the extent to which illness in an area may drive antibiotic resistance.

Although reducing antimicrobial resistant infections in humans and animals is a global priority, there remain major gaps in measurements of the global prevalence of antibiotic resistant organisms across species and region. Knowledge of transmission dynamics and prevalence of community-acquired antimicrobial resistant species shared between overlapping humans and animals is limited. Detecting the distribution and diversity of specific antimicrobial resistant genes (ARGs) within and between human and animal microbiomes can further identify potential spillover events. Although antibiotics are lifesaving during some infections, agricultural and medical overuse of antibiotics contribute to the current rise of resistant organisms in human and animal populations (24). Further, and consequently, domestic animals, peri-domestic rodents, and wildlife all harbor ARGs, and each group can act uniquely as a sentinel for emerging or increased spread of antibiotic resistance (21,235–237). Comparisons of resistomes are well documented between human and agricultural animals (236,238), agricultural soil(30), and in wastewater (28), showing widespread ARG diversity that is geographically specific. A lesser focus has been on comparative studies of ARGs in wildlife animals overlapping with human communities.

In this paper, we examined human and brown mouse lemur gut microbiomes to investigate the extent of ARG sharing between humans and wildlife in rural Madagascar where there are opportunities for humans and lemur spatial overlap. Whether shared environment could be enough to result in shared microbiomes/resistomes is of interest, given that in Madagascar, lemur species exist across a gradient of human-transformed space, from undisturbed wild to being kept as pets in some households. The gradient of lifestyle has had a parallel effect on pathogen prevalence and ARG abundance. In ring-tailed lemurs, for example, it was shown that ARGs were in greater abundance in captive populations compared to wild, and ARGs that could impact human health were correlated to the level of human disturbance in the location of varying lemur populations (29). Further, mouse lemurs dwelling in more human-disturbed areas harbored pathogenic bacteria also found in nearby dwelling human, rodents, and livestock (239). It is unknown whether the ARGs identified in wild populations share a similar genetic profile to the profile of the human microbiomes present in the area. Information on general human and lemur interactions, even no interactions, could be informative of the dispersal of reservoirs for transmission.

The landscape of genomic analyses capable of comparing bacterial communities ranges from fast but less sensitive 16S sequencing to highly discriminatory but labor intensive metatranscriptomics (240). Application of metagenomic sequencing can strike the balance for understudied microbiomes and allow for comparing diversity at the microbial species scale without a priori assumptions of what species should be expected (240), and capture more gene-level diversity which cannot be evaluated from taxonomy gene marker techniques (241). Although challenges remain for positive identification of rare species from short-read sequencing, it is nevertheless a powerful approach for examining patterns in abundantly present and known species (241), for diagnosis of present pathogens (242), and for identifying compositional differences between environmental samples (28). Additionally, genetic closeness of detected species or genes of interest shared by humans and animals can be used to infer zoonotic transfer and further combined with epidemiological information to identify links between ARG presence, species richness, and risks of illness and transmission. However, the abundance of antibiotic resistance genes may be underestimated, and a latent, or undocumented diversity in established databases, population of ARGs exists within and between different hosts and environments (243). Sequencing metagenomic analysis can help identify the reservoirs of rare or unknown species and offer a starting place for hypothesis generation for complimentary methods, such as functional metagenomics, to fill in the species knowledge gaps (244).

Here, we aimed to identify the ARG burden and diversity between humans and wild lemurs near Ranomafana National Park in Madagascar. This unique system provides an opportunity to examine this interplay in low-resource, rural, tropical communities where exceptional biodiversity and human-wildlife overlap create unusually high potential for novel zoonotic events. Comparison of the respective bacterial species and ARG profile lays the foundation for understanding ecological and evolutionary patterns outside of agricultural and clinical settings. This has implications for documenting potential downstream or indirect selection pressure that anthropogenic drug use has on an ecosystem regardless of direct human and animal interaction.

Materials & Methods

Sample Collection and Demographic Survey

As a component of a One Health research platform in Ifanadiana District, Madagascar, a household survey was conducted from June to August 2017 in eight communities in roadless areas < 5km from Ranomafana National Park to collect information regarding household member demographics, antibiotic usage, household illness, exposure to wildlife, and previous illness with diarrheal disease. Within these communities, we have documented diverse global health challenges including high prevalence of enteric infections and resistance genes regardless of antibiotic class, and zoonotic human-wildlife linkages (239,245–248). Human participation in the study and collection of survey data were approved of and reviewed by the Emory Internal Review Board (IRB00093812). Before survey administration, informed oral consent was gathered and documented. Household members were also asked to voluntarily submit a fecal sample regardless of history of diarrheal illness. Fecal sample IDs were linked to their corresponding household survey responses, and deidentified for downstream analysis. Fecal samples were collected from captured wild brown mouse lemurs (*Microcebus rufus*) along footpaths near the villages. The mouse lemurs were trapped using banana-baited Sherman traps (XLR, Sherman Traps Inc., FL), and set overnight at 16:00 and checked at 05:00. One microliter of fresh fecal samples was collected from individual trapped lemurs by using a sterile tongue depressor and transferring the sample into a cryovial filled with approximately 0.8mL RNAlater. The Emory University Institutional Animal Care and Use Committee provided full approval for this research (#3000417) and the field research procedures were approved by Madagascar's Ministry of Environment, Ecology and Forests (permit nos: 028/17; 083/17; 136/17; 146/17; 164/17).

DNA extraction and sequencing

DNA was extracted from fecal samples using a standard Zymo Inc. bead-beating kit. Whole metagenome shotgun sequencing was performed on the NextSeq 2000 platform using Illumina DNA library preparations. Sequencing produced separate forward and reverse paired-end fastq files, which served as inputs for bioinformatic processing. All mouse lemur metagenomic samples and select human fecal samples from each of the surveyed eight communities were included for metagenomic sequencing and downstream analyses. Human samples were selected through stratified random sampling per village. To be included for selection, households had to include at least one adult, one school-aged child (5-17 years), and one child under 5 years. Eligible households were grouped by their home village and a random three households were selected within each community, sampling without replacement. From the selected households, a sample was chosen for each age group. If only one household member represented an age category, then their sample was selected. If more than one household member was represented by an age group, then a second random sampling was done to choose the representative sample for that age group.

Bioinformatic processing and quality control

An overview of the bioinformatic workflow is shown in Supplemental Figure S1. Data quality was assessed with FastQC (v.0.11.9) before and after adaptor trimming and removal of host reads (249). Read quality trimming was conducted using Kneaddata (v0.10.0) with --*trimmomatic*, which employs Trimmomatic (v0.39-2) (250) and Bowtie2 (251) to remove adaptor reads and reads mapping to the human genome refence GRCh37 (252), keeping reads at or above Phred 33. Reads from lemur microbiomes were additionally mapped against a draft genome assembly of *Microcebus murinus* (GCF_000165445.2) (253), selected for its high level of completeness of assembled chromosomes, representation of male and female chromosomes, and shared ancestry to *M. rufus*. The *M. murinus* assembly was indexed using bowtie2-build. The forward and reverse paired-end reads from lemur microbiomes were mapped to the indexed assembly using Bowtie2 (v2.5.0) and saved as a SAM file. Unmatched reads (and therefore non-host reads) were subsequently removed using SAMtools' sort and fastq functions (254). After filtering, only metagenomic sequences with 1 x 10^6 reads or more were considered for further analysis or assembly. Metagenomes were assembled using SPAdes (v3.15.4) (255) on the trimmed and decontaminated paired-end fastq files using the *–meta* parameter. Separate forward and reverse fastq files were used as input with the *-1* and *-2* flags, respectively. The human DNA-scrubbed analysis sequences are available under the Bioproject PRJNA1008138.

Classification of bacterial species and ARGs

Taxonomic composition of the filtered reads was first calculated using MetaPhlAn4 (256), with flags *--input_type fastq*, *--unclassified_estimation*, and *--bowtie2out*, to estimate relative abundance of the both classified and unclassified reads which did not match gene markers in the database. This was followed by a second analysis against the Bowtie2 indices to calculate relative abundance of bacterial-associated reads only using the flags *--input_type bowtie2out*, *--t rel_ab*, *--ignore_eukaryotes*, *--ignore_archaea*. The vOct22 Bowtie2 database available for MetaPhlAn4 was downloaded using the command: *metaphlan --install --bowtie2db* and used as a reference for taxonomic markers. To calculate the abundance of ARGs, we enumerated the reads per kilobase per million (RPKM) relative to the amount of detected bacterial reads in the sample. We derived this formula from Munk et al (28), but accounting for reads. The formula is as follows:

$$\frac{Gene\ reads}{(Length\ of\ gene,\ kilobases) \times (Total\ bacterial\ reads)} \times 10^9$$

Filtered reads were first processed through KMA (257) using the AMR Finder Plus nucleotide sequence database to identify ARGs and virulence genes (258). ARGs were specifically subset from virulence genes based on classification from the Bacterial Antimicrobial Reference Genes database for sequences related to antibiotic resistance (PRJNA313047). Unique genes were categorized based on annotated gene symbols and unique alleles were categorized based on sequences matching to present NCBI nucleotide reference sequences. Results were included if the detected allele had a template coverage greater than or equal to 60 percent and a query identity greater than or equal to 90 percent, and if they had at least three reads assigned. The results file was then joined with a .mapstat file generated by KMA to quantify the number of reads assigned to each reference sequence. To contextualize the reads relative to bacterial content, the filtered fastq files were also run through Kraken2 (259) using the flags *--paired*, *--report*, *--classified-out,* and *--unclassified-out,* and referencing the Kraken2 Standard database (26 September 2022) to obtain the number of reads rooted at the bacterial level and the number of unclassified reads, or reads unable to be identified using the database classifications, in the sample. RPKM was calculated for each ARG allele.

Statistical analysis of species and ARG diversity

Final statistical analyses were conducted in R (260). Continuous values and counts of discrete data were assessed for normal distribution. The Wilcoxon rank sum test in the stats package (v4.0.4) was used to compare human and lemur metagenomes differences in median total detected reads, proportion of reads mapping to higher order taxa, Shannon indices for bacterial species and ARG allele diversity, RPKM of ARG reads mapping to specific antibiotic classes, and median number of species per sample. Alpha and beta diversity metrics were calculated using the vegan package (261). Shannon diversity was calculated based on the presence and absence of detected species or alleles. For this system, two measures of alpha diversity were used to help to capture a better understanding of detected species. The Shannon diversity index allows for comparison of both the richness and evenness of the community structure thus relying on both the abundance and the overall number of species, while Chao1 gives a greater weight to low abundance species present in a sample to help predict the likely number of missing species (Bo-Ra Kim et al., 2017). Principle component analyses (PCA) were performed on the relative abundances of species and RPKM values of ARG alleles transformed into centered log-ratios to account for the compositional nature of metagenomic data (262). Subsequently, Aitchison distance was calculated to assess between-sample differences in species/allele diversity. Chao1 and rarefaction statistics were calculated using the iNext package using the sum of the presence of each species or allele detected within human or lemur microbiomes as input (263).

Differential gene abundance analysis

Differential gene abundance between humans and lemurs for antibiotic resistance genes was conducted on the read counts, summarizing the allele hits to the level of the gene using ALDEx2 (v.1.35.0) (262). First, the raw read counts were transformed using the command *aldex.clr(),* with Monte-Carlo sampling set to 128 and the measured denominator set to “all”. To account for both composition and scale in the read counts, uncertainty was added to the model using a gamma value of 0.5. A sensitivity analysis for significance of unique features at various values of gamma was conducted for reads summarized at the gene level and per corresponding antibiotic class associated with resistance (See Supplemental Figure S3) (264). To statistically evaluate the transformed abundances, *aldex.effect()* and *aldex.ttest()* were used to perform Welch’s T-test and a Wilcoxon rank sum test, and corrected for false discovery using Benjamini-Hochberg corrected p-values (<0.05). Final results were plotted using the *aldex.plot()* function.

Sequence comparison of antibiotic resistance genes shared between humans and lemurs

AMR Finder Plus (258) was used to identify contigs with ARGs. Contigs with the same ARG detected in at least one human and one lemur microbiome were extracted from their sample assembly using *bedtools getfasta*. These sequences were used to create a custom nucleotide BLAST database (v2.12.0). Each sequence was then queried against the database to identify the pairwise percent identity of each gene compared to other detected genes of the same type. A second BLAST comparison was conducted by extracting the 1000 base pair regions before and after ARGs commonly present in human and lemur samples, and the pairwise percent identities were quantified. To gain additional insight into the genomic contexts of highly similar ARG-regions, ARG-bearing contigs were extracted from the assemblies and annotated using Bakta (v1.7.0) (Schwengers et al., 2021). For ARG-regions in which commonly found ARGs between pairs of subjects had greater than 90% similarity, the gene synteny of the annotated contigs was inspected and visualized using Gggenes (265).

Results

The diversity and abundance of bacterial species and ARGs differ between humans and lemurs.

A total of 73 human-derived samples and 15 lemur-derived samples were selected for shotgun metagenome sequencing. Of these samples, 57 human samples had greater than or equal to 1 x 10^6 total reads after decontamination of human reads. After mapping to the lemur genome assembly, 11 lemur samples had greater than 1 x 10^6 reads for analysis. The metagenomes of these 68 samples were further characterized for bacterial species abundances and presence of antibiotic resistance genes.

There was a higher total number of reads in lemur samples compared to human samples after decontamination (Fig. 5.1A) which could not be explained by a large number of small reads present. Although both human and lemur read libraries had average read lengths within an acceptable range for downstream mapping and assembly, the average sequence length for lemurs was higher and tightly ranged (147 to 139 reads) ($p<0.05$) (Fig. 5.1B). For the majority of human and lemur microbiomes, the most abundant taxa identified belong to kingdom Bacteria (Fig. 5.1C). However, lemur metagenomes noticeably contained higher relative abundances of reads unable to be classified taxonomically. A comparison of unclassifiable reads to taxonomically classified reads did not show significant differences in GC content, Q30 score, average length, or minimum length (Fig. S5.2). This suggests that much of the diversity of taxa in lemur microbiomes is not represented even in large database collections used for widescale taxonomic composition classification.

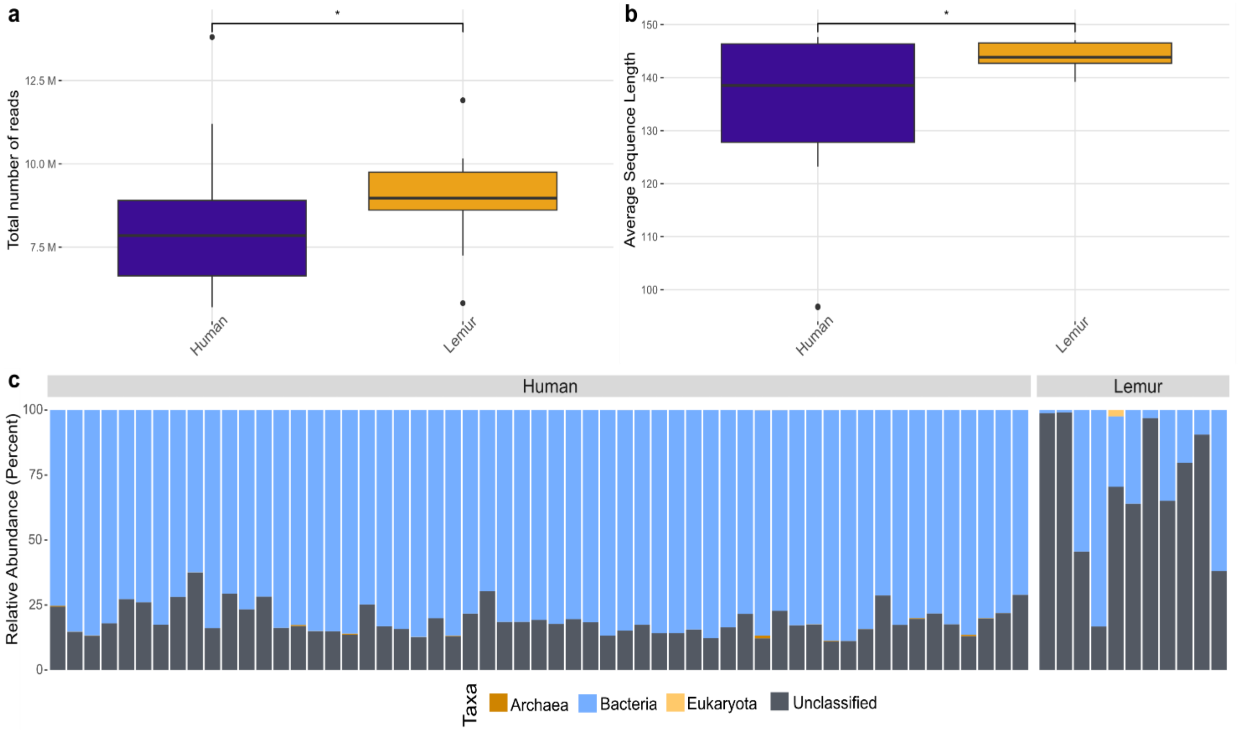

Figure 5.1. Human and lemur metagenomes are different in the classification of sequence reads to higher order taxa. Sequences were filtered for adaptor sequences, tandem repeats, and reads mapping to human or lemur reference assemblies. (a) The number of total read pairs is significantly higher in lemur fecal metagenomes compared to human fecal metagenomes (Wilcoxon Rank Sum, $p<0.05$). (b) The average length of reads within the lemur fecal metagenomes is higher than human fecal metagenomes (Wilcoxon Rank Sum, $p<0.05$). (c) The relative abundances of kingdom-level taxa was quantified using MetaPhlAn4. Reads unable to be identified as belonging to a higher order were designated as “unclassified.” In general, humans and lemurs have the majority of the abundance of taxa assigned to Bacteria or Unclassified; however, lemur fecal metagenomes are characteristically higher in the abundance of unclassified reads.

Although different in rank of abundance, human and lemur fecal metagenomes include a high abundance of bacteria from the phylum Bacillota (Firmicutes), and human metagenomes are also dominated by Bacteriodota. Several lemur samples are contrastingly dominated by species from Pseudomonadota (Proteobacteria) (Fig. 5.2A). Humans and lemurs shared 55 of 452 known bacterial species (12.2%) that had at least 0.01% abundance within a single metagenome and occurred in at least 10% of all samples, suggesting there are many rare species detected in the system. The Shannon index between the two groups was higher in humans (P<0.05) (Fig. 5.2B). Distinct clustering by sample source was also observed when examining the compositional differences between sample sources. Lemur microbiomes grouped more with other lemur microbiomes without overlapping human microbiomes, and human samples overlapped regardless of the village of residence (Fig. 5.2C). By Chao1 estimates, when considering detectable species including those in the lowest abundance (<10% of samples) there were fewer species able to be sampled from lemurs (590, 95% confidence interval (CI): 496-729) compared to humans (1091, 95% CI: 1057-1144) (Fig. 5.2D).

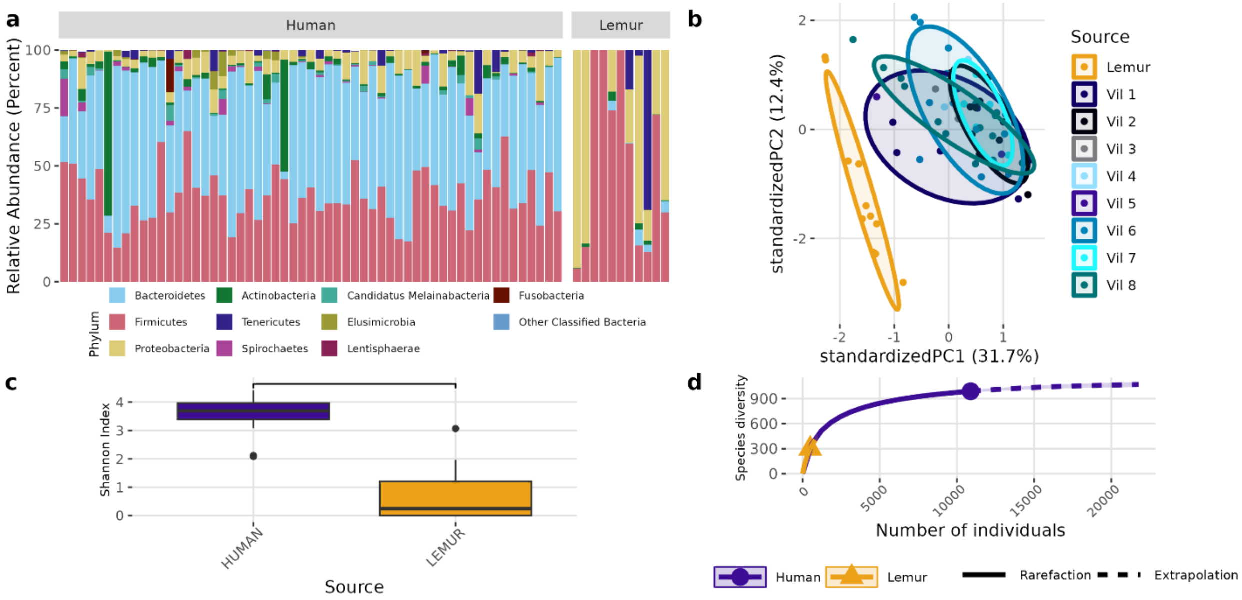

Figure 5.2. Bacterial species communities are distinct between humans and wild lemurs. (a) Relative abundance of bacterial phyla detected in human and lemur derived metagenomes. (c) Principle component analysis using Aitchison distance on centered log-ratio transformed abundances of individual species stratified by the home village for each human sample or if it was sourced from a lemur, and demonstrates component dissimilarity between human and lemur samples. (c) The Shannon diversity indices of the species detected in human and lemur metagenomes show a difference in the mean value between host sources. Statistical significance was determined by Wilcoxon rank sum analysis at $p<0.05$. (d) Rarefaction curves of human- and lemur-associated bacterial species, where the x-axis is the sampling effort of available individual bacterial species and the y-axis is the estimated richness.

A different pattern emerged when examining the overall composition and abundance of ARGs detected in human and lemur metagenomes. We identified reads mapping to 107 unique ARGs, which comprises 217 unique alleles. Two lemur samples had no reads matching ARGs in our database. Individual microbiomes varied in abundance of ARGs grouped by the class of antibiotic to which they confer resistance, but all human microbiomes carried genes associated with tetracycline and trimethoprim, and 55/57 human metagenomes carried resistance genes to beta-lactam antibiotics. In contrast, there was no one shared antibiotic class among detected ARGs in lemur microbiomes, but all classes seen in human microbiomes were represented in at least one lemur microbiome (Fig. 5.3A). Lemur microbiomes had no statistical difference in ARG richness compared to human microbiomes ($P<0.05$) (Fig. 5.3B). Still, lemur microbiomes clustered distinctly in their ARG diversity from human microbiomes, however human microbiome ARG profiles from all resident villages overlapped with one another (Fig. 5.3C). Concordantly, rarefaction estimates, alongside Chao1 calculations suggested that the estimated maximum number of ARG alleles to be sampled are likely similar for lemurs (201, 95% CI: 166-264) and humans (206, 95% CI: 160-302) (Fig. 5.3D).

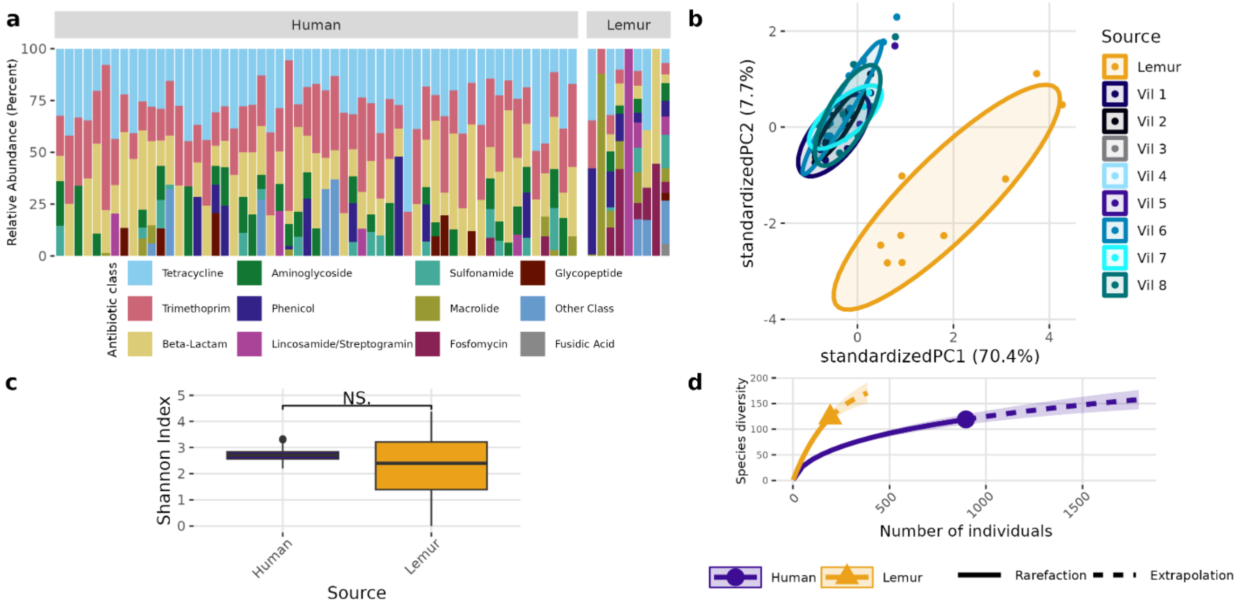

Figure 5.3. Antibiotic resistance gene abundances are distinct between humans and wild lemurs. (a) Relative abundance of genes by their associated antibiotic resistance classes detected in human and lemur derived metagenomes. (b) Principle component analysis using Aitchison distance on the centered log-ratio tranformed abundances of unique ARGs stratified by the home village for each human sample or if it was sourced from a lemur show a distinct grouping of lemur samples with other lemurs and separate from humans. (c) Shannon diversity index of the unique ARG alleles detected in human and lemur metagenomes and demonstrates a difference in mean values between host source. Statistical significance was determined by Wilcoxon rank sum analysis at p<0.05. (d) Rarefaction curves of human- and lemur-associated ARG alleles, where the x-axis is the sampling effort of available alleles and the y-axis is the estimated richness.

Five antibiotic resistance genes were in significantly greater abundance among human microbiomes compared to the lemur microbiomes (Fig. 5.4A-C). These genes consisted of *dfrF* and five separate tetracycline-resistance genes, *tet*(32), *tet*(40), *tet*(W), and *tet*(Q). The effect size of the difference remained significant for these genes when considering the difference in variance in the data points between the two groups (Fig. 4.4B). When grouped by associated antibiotic class of resistance, no associated class groups were differentially abundant between humans and lemurs (Fig. S5.3).

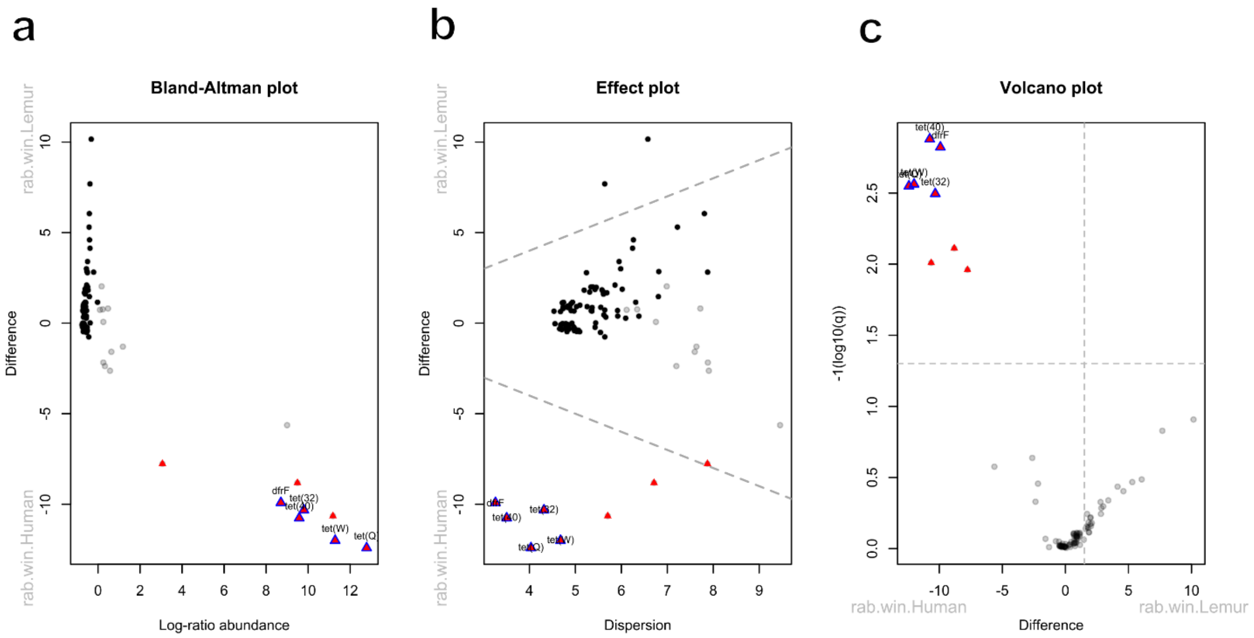

Figure 5.4. Antibiotic resistance genes vary in abundance between human and lemur microbiomes. Raw read counts summarized at the gene level were compared for differential abundance using ALDEx2. Red triangles indicate a significant difference in abundance by an effect value >2. Blue outlining indicates 95% confidence that the value does not intersect zero. Black dots indicate rare and non significant genes while gray dots signify abundant but non-significant genes (a) Bland-Altman plot demonstrating the relationship between the difference between groups in median centered log-ratio (clr) values of each gene and the relative abundance of those genes. (b) An effect plot of the difference between groups in median clr values of each gene and the difference in dispersion, with the dotted lines representing values where dispersion and difference are equal. (c) A

volcano plot demonstrating the abundance in the clr values of each gene. The dotted x-intercept line indicates values at a posterior predictive p-value of 0.001, and the y-intercept line indicates a 1.5-fold difference in log abundance.

Humans and lemurs share highly conserved integron-associated ARGs.

To capture more specific ARG dynamics between humans and lemurs, we quantified and compared assembled ARGs that were detectable in both human and wild lemur metagenomes. A total of 14 ARGs were detected in common between human metagenomes, with all 57 human metagenomes sharing at least one gene with at least one of three lemur metagenomes.

ARGs of the same type were compared between each metagenome containing that gene. Overall, ARGs from different metagenomes were highly similar, with a median nucleotide sequence identity of 99.51 (98.55-99.79) for human-human, 99.67 (99.24-100) for human-lemur pairs, and 100 (100-100) for lemur-lemur comparisons (Figure 5A). Similarly high levels of sequence identity were found within each ARG (Fig. 5.5B). The largest ranges of diversity were among pairwise comparisons of *tet(O)* and *tet(Q)* genes, respectively (Fig. 5.5B). We also compared the genetic context surrounding the shared ARGs to determine the similarity of genomic context regardless of ARG sequence conservation. Sequence identity of the 1000 base-pair flanking regions were also nearly identical between samples. Specifically, the regions around four ARGs (*aadA1, dfrA1*, *qacEdelta1*, and *sul1*) were from seven human metagenomes and one lemur metagenome showed highly similar pairwise sequence comparison of the human-human and human-lemur source pairs (Fig.5.5C). Only one gene, *lsa(D)*, had a pairwise identity score less than 100 percent (Fig. 5.5D).

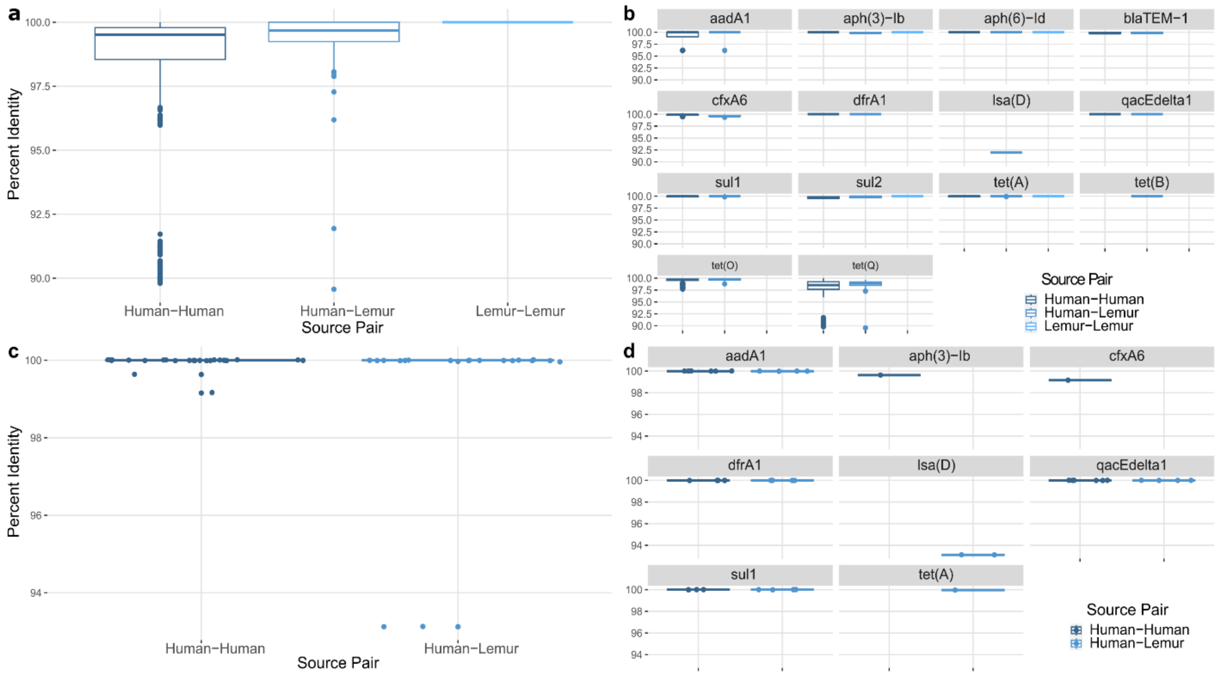

Figure 5.5: ARGs shared between humans and lemurs are highly conserved. (a) Percent identity for each ARG that was detected in at least one human and one lemur. Pairwise comparisons of the gene source, human or lemur metagenome, was calculated using blastn where positive sequence hits were at least 90% of the query length. (b) The pairwise source comparisons of the percent identity were then stratified by the specific ARG query. (c). Percent identity for the 1000 base pair region before and after the query gene of interest that was detected in at least one human and one lemur. Pairwise comparisons of the ARG-region source, human or lemur metagenome, was calculated using blastn where positive sequence hits were at least 90% of the query length. (d) The pairwise source comparisons of the percent identity were then stratified by the specific ARG-region query.

Because the contexts of these genes were so well conserved, we investigated the gene synteny to understand if there were shared AMR genes in close genetic proximity. ARGs that were co-occurring within assembled contigs included *dfrA1* with *aadA1*, *sul1*, and *qacEdelta*. Of the eight samples containing *dfrA1*, seven contained a conserved *aadA1* gene and one sample had a broadly categorized *aadA* region (Fig. 5.6). Five of these samples had a *sul1* downstream of *aadA1*, and three contained *intI1* (encoding class 1 integrase), including the single lemur sample. Of the seven human microbiomes containing a *dfrA1-aadA* pairing, residents were from four different villages, with one pairing from the same household. The single lemur sample was collected closest to a village that none of the human residents carrying this cassette were from.

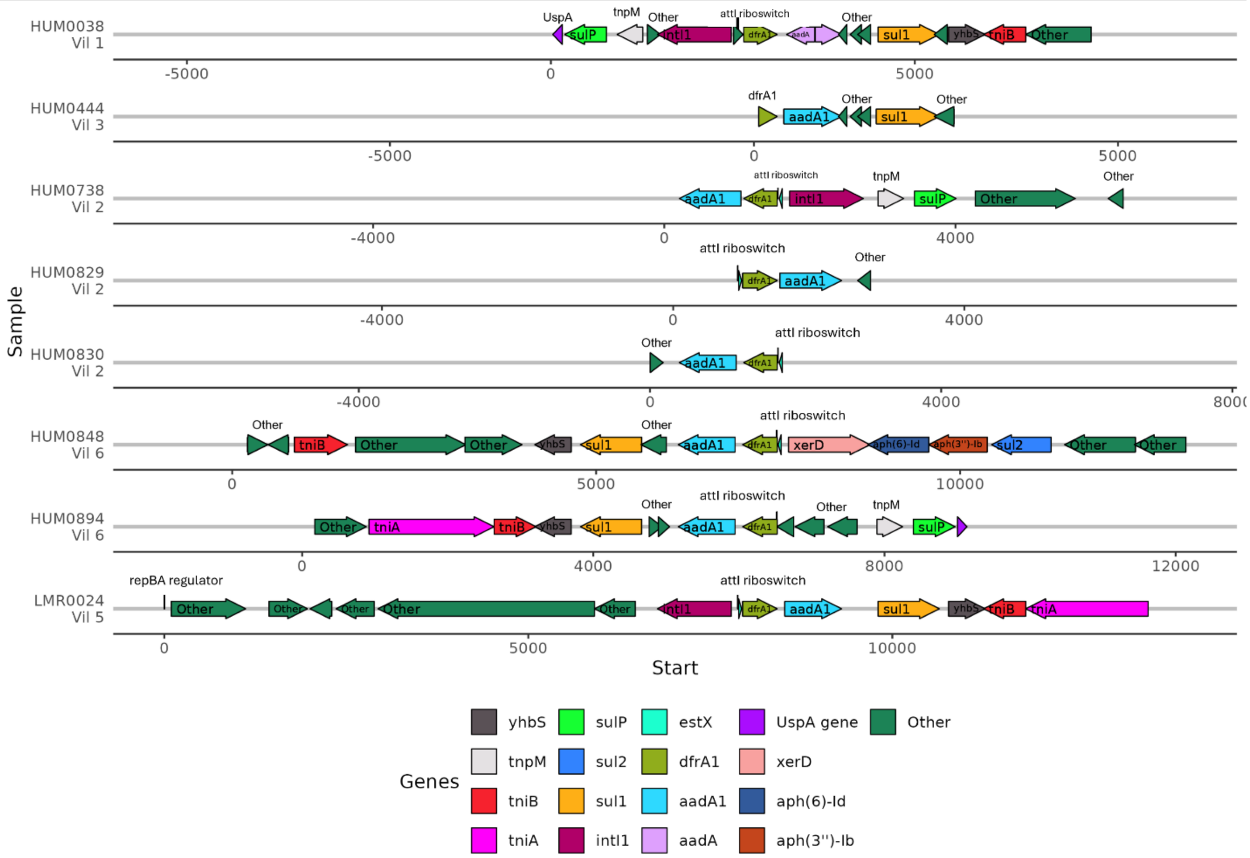

Figure 5.6. Human and lemur dfrA1-aadA1 cassette synteny. Contigs containing highly similar dfrA1 and aadA1 genes were compared from one lemur and four human samples. Sequence annotations were identified using Bakta. Samples from individuals are represented for each line and annotated with the individual and geographic source. Genes present on two or more contigs, or antibiotic resistance-associated genes labeled and represented by colored arrows. Other detected genes unique to the contig are labeled as "other". Sequence coordinates are aligned relative to the present dfrA1 gene on their respective contigs. Arrows indicate the strand direction of the detected gene.

Discussion

To our knowledge, this is the first study to directly compare the antibiotic resistance profiles from human and wild lemur microbiomes from the same geographic space, thus providing further insight into antibiotic resistance gene flow between residential human and wildlife host populations. We quantified and compared the bacterial species and ARG abundances present in human and lemur metagenomes and found their overall profiles to be distinct in both bacterial species and ARG distribution, while human microbiomes from different villages were largely comparable to one another. We also detected some differentially abundant genes among human microbiomes conveying resistance to tetracycline and aminoglycosides. Lastly, we assessed the genomic similarity of ARGs shared between human and lemur microbiomes and found a shared multidrug resistant mobile gene cassette.

Understanding the bacterial composition of microbiomes from hosts within a larger ecological community helps to establish the biological baseline for future surveillance of spillover. In this analysis, humans and lemurs were largely distinct in their microbiome species and ARG abundance and distribution. We found that lemur microbiomes were far less rich in known bacterial species compared to human microbiomes despite the quantity of available DNA in the sample. This is likely explained by a limitation in the detectability of uncharacterized species in our chosen database, which also highlights a larger issue of the current state of curated taxonomic databases available for metagenomic analysis. Even with this bias, though, it is reasonable to conclude that the populations from wildlife, being under-sampled across studies, would likely drive this difference even further from humans.

In contrast the Shannon index for ARGs was not different between lemur microbiomes and human microbiomes, and the number of ARGs detected between the two groups were highly similar. Humans and lemurs shared proportionally few types of ARG alleles, but both groups had a similar absolute number of ARG allele types detected. The diversity of alleles suggests less detection bias that would be preferential toward human microbiomes. It is still possible that the full scope of ARGs is yet to be known (243), but in this system there is evidence to suggest that at least what can be known about antibiotic resistance genes is comparable between humans and wildlife. The structure of resistomes within the gut microbiomes of vertebrates outside of humans are influenced by numerous host-associated factors and environmental factors, including habitat and the threatened status of the wildlife population (236). For this study, mouse lemurs were sampled along roadways specifically to detect patterns in the resistomes in an area of human and wildlife crossover. The unique life histories and diet of non-human primates from human communities would lead to an expectation that gut bacterial species and present ARGs are likely distinct, as has been demonstrated with comparative analyses of the gut microbiomes of humans and non-human apes (266). Our study is consistent with this pattern when comparing human to sympatric lemur microbiomes, as the lemur microbiome is largely divergent in species and ARGs present. Nevertheless, the presence of highly conserved ARGs could be the result of shared host traits selecting for specific microbial functions within the gut or from shared lineages acquired from common overlapping environment.

Some antibiotic-resistance genes were more abundant among humans, though resistance to no one class was more abundant. It is notable that four of the five differentially expressed genes belonged to tetracycline-resistance genes and one aminoglycoside-resistance genes. Phenicol and tetracycline class drugs have been used extensively in agriculture (36) and thus could end up trickling into natural settings, impacting how often wildlife become exposed to these ARGs compared to humans. The synergy of clinical and agricultural use could explain why there are overlaps in top abundance. For example, Since 2009, the World Health Organization has recommended that sulfamethoxazole + trimethoprim, doxycycline, or tetracycline be used as first-line choices for pre- and postexposure treatment to *Yersinia pestis*, the pathogen causing plague and which is endemic in Madagascar and responsible for periodic large outbreaks, including during 2017 (267,268). The diversity of region-specific usages of antibiotics suggests that there is likely no single pressure resulting in the maintenance of the most abundant ARGs, but it does call for a One Health awareness toward the stewardship of different classes so that these drugs can remain effective for interventions, such as management of plague.

Fourteen assembled ARGs were shared between humans and lemurs, though there were 69 distinct ARGs among assembled metagenomes. Presence of shared genes is a signifier of potential ARG reservoirs for human and agricultural pathogens. Among the shared ARGs, several have been detected in pathogen samples with phenotypic resistance to their corresponding antibiotic class, including *aph(3'')-Ib* (269), *aph(6)-Id* (270), *qacEdelta1* (37), and *cfxA6* (271). *DfrA1*, *aadA1*, *aph(3")-Ib*, *aph(6)-Id*, all have a high risk of contributing currently or in the future to pathogen multidrug resistance (244). The *lsa(D)* gene, responsible for lincosamide resistance, was detected in diseased farm-raised fish and attributed to emerging fish pathogens (272). We did not identify a clear village-level association between lemurs and humans that had these shared genes. Therefore, the high similarity could be explained by strong selective pressures within the environment to conserve these gene structures, or it could be explained by ongoing drift of bacteria horboring these genes moving between human and lemur populations via uncharacterized pathways, such as river systems or intermediary contact between wildlife and other domestic or peri-domestic animals. Many of these genes have a prevalence in other global areas where genetic sequences of ARGs sourced from different metagenomes are also highly conserved (273). For either scenario, detection of clinically significant ARGs in wild lemur populations that may not be directly interacting with human communities signifies just how diffuse the community resistome is and may make combating drug resistance more difficult as human and wildlife are brought more and more into contact.

We did identify a common class 1 integron in close genomic context with multiple drug resistance genes present in several human and one lemur microbiome. Common characteristics of a class 1 integron are encoding of intI1 at the 5' coding end, followed by with a variable cassette region and then encoding of *qacEdelta* and *sul1* at 3' coding sequence (274). Other globally distributed gene cassettes harboring trimethoprim-resistant *dfrA* and aminoglycoside-resistant *aadA* genes are known to be associated with class 1 integrons (275–277). Specifically, these gene cassettes have also been found in known patient samples in Madagascar's capital Antananarivo among ESBL-producing Enterobacteriaceae, with the most frequent cassette pairing being *drfA17-aadA5* (275). Contrastingly, we did not identify this specific cassette among any of the microbiomes under consideration in this study from our rural community members. The *dfrA1-aadA1* cassette among our study samples is dispersed between several members of different villages, though more investigation is necessary to understand if its prevalence is hallmark of the specific region. Class 1 integrons harboring *dfrA1* can move between species of gram negative organisms *in vivo* (278). In the context of our study and the growing body of evidence that human-driven antibiotic use drives higher antibiotic resistance profiles in animals and in wildlife, we should be concerned that even non-agriculture animals are maintaining highly similar ARGs to humans in their microbiomes. Stewardship efforts necessary for this system may focus on closing off pathways between human-developed space and wildlife and conservative use in agriculture. Detection of ARGs through metagenomics or other screening is a useful tool for increased surveillance efforts, but it will take additional research to develop meaningful relationships between genetic abundance and the frequency of spread between species or changes in phenotypic resistance to antibiotics. Optimistically, the advent of technologies such as long-read sequencing offer a compliment to identifying species genomes directly as sequences and as reference scaffolds for short read sequences. Given that antibiotic resistance is often a trait maintained when bacteria are consistently exposed to antibacterial chemicals, more work must be done to monitor whether individual genes are continually being reintroduced to wildlife metagenomes from humans to better understand how stable the lemur metagenome niche is as an ARG reservoir.

Conclusions

In this study, we took a metagenomic approach to characterize ARG presence in a specific ecological system and uncover previously unexplored comparisons between humans and wildlife. Our observations add to the growing effort to characterize the global extent of ARG presence, the range of which is still limited especially in lower- and middle-income countries. These findings reflect some known global patterns of drug resistance prevalence and highlight unique patterns for this geographic area. This research supports a continued effort to monitor antibiotic usage for humans and in agriculture, especially effects on non-pathogen members of microbiomes, and their further dissemination into the ecosystem. As more research reveals the extent of ARG transmission through an environment, it is evident that there is an increased needed to investigate intermediary processes beyond individual players' proximities to one another that can lead to drug resistant gene movement through ecological space.

Additional Information and Declarations

Human Ethics

The following information was supplied relating to ethical approvals (*i.e.*, approving body and any reference numbers): The Emory University Institutional Review Board granted approval to carry out this study (reference number IRB00093812)

Animal Ethics

The following information was supplied relating to ethical approvals (*i.e.*, approving body and any reference numbers): The Emory University Institutional Animal Care and Use Committee provided full approval for this research (#3000417)

Field Study Permissions

The following information was supplied relating to field study approvals (*i.e.*, approving body and any reference numbers): Field research procedures were approved by Madagascar's Ministry of Environment, Ecology and Forests (permit nos: 028/17; 083/17; 136/17; 146/17; 164/17).

DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

The metagenomic sequence reads for this project are available in the Sequence Read Archive: ID PRJNA1008138.

Data Availability

The following information was supplied regarding data availability: The analysis code is available on GitHub and Zenodo: https://github.com/bmtalbot/Humans_and_Lemurs_2017.

The accompanying data are available on Zenodo: Talbot, B., Clennon, J., Rakotoarison, F., Rautman, L., Durry, S., Ragazzo, L., Wright, P., Gillespie, T., & Read, T. (2023). Datasets for Metagenome-wide characterization of shared antimicrobial resistance genes in sympatric people and lemurs in rural Madagascar (1.0.0) [Data set].

Zenodo. https://doi.org/10.5281/zenodo.10402843.

Acknowledgements

For logistical and infrastructural support, the authors thank MICET, Centre ValBio, Benjamin Andriamihaja, Pascal Rabeson, Jean Claude Razafimahaimodison, Jean de Dieu Ramanantsoa, Maya Moore, Jesse McKinney, Telo Albert, Victor Tombotiana Aimé, Emilie Redwood, Katherine Noble, Ria Ghai, Wenjie Mei, Joel Hartter, Karen Bailey, Heritiana Anne Louisette, Rasolondraibe, Amelie Marcelle, Solo Justin, and François Zakamanana. The authors also thank Emily Wissel for critical insights and recommendations for metagenomic techniques. The manuscript's contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention.

Supplementary Material

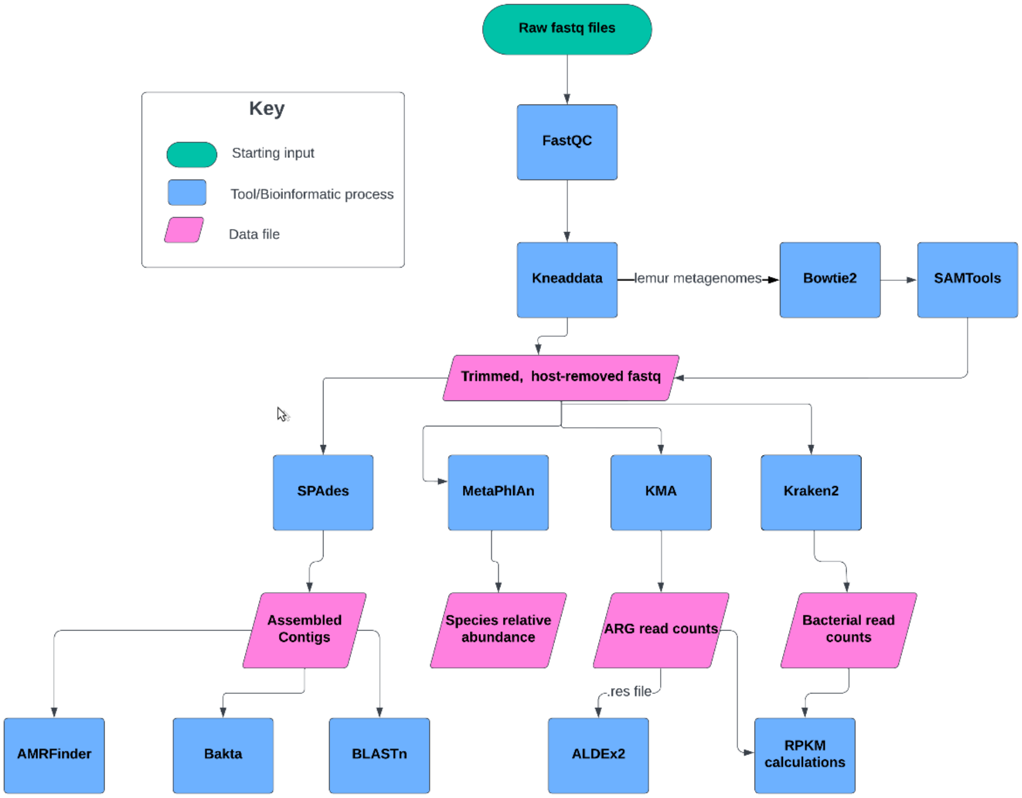

Figure S5.1. A schematic of the overall bioinformatic workflow of sampled human and lemur metagenomes.

Quality comparison of Kraken2-classified Lemur microbiome metagenome reads

The nucleotide sequence quality between unclassified and classified reads was compared to understand its impact on categorize metagenomic reads into taxa for lemur sequence. Reads were run through a comprehensive Kraken2 database (PlusPFP release date 2024-01-12) which includes taxa Refseq markers for bacteria, human, virus, plasmids, plants, protozoa, fungi, and UniVec_Core sequences. Kraken2 was run with the flags --unclassified-out to generate unclassified reads from each read pair sequences, --classified-out for the classified reads per paired sequence, --report to generate the quality report, and --paired to indicate the two paired end sequences for each sample. A quality report was produced summarizing the paired-end sequences of each sample.

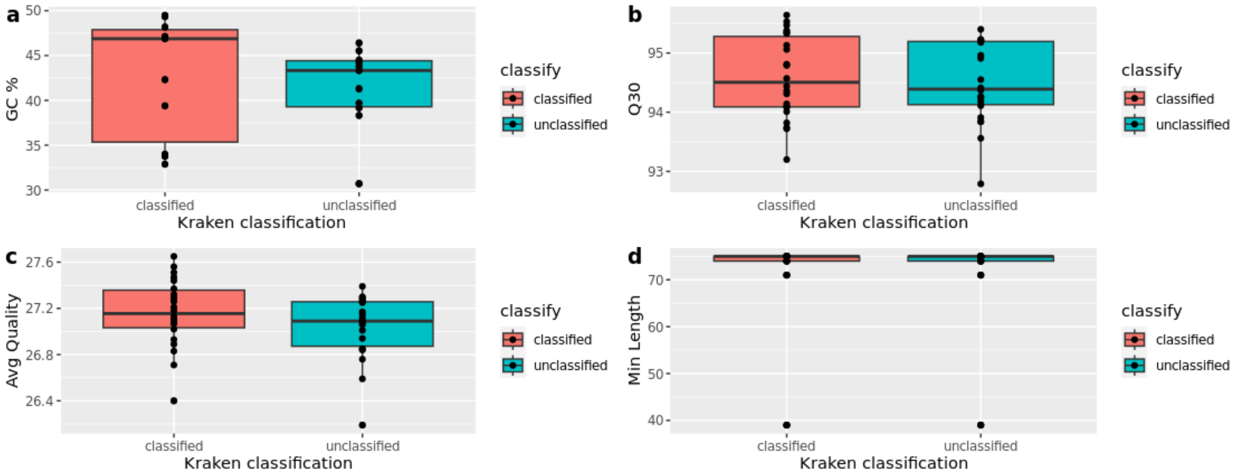

Figure S5.2. Comparison of quality metrics between classified and unclassified reads from lemur microbiomes. Boxplots were generated from the values of each sequence of a read pair per sequence after analysis with Kraken2 against the Kraken PFP database for the following four metrics: (a) The GC percentage per sequence, (b) percentage of reads at Q30 score, (c) the average Phred sequence quality, and (d) the minimum sequence length.

Evaluation of differential abundance calculations using ALDEx2 for ARGs grouped by gene families and primary antibiotic class association

Raw reads of ARGs were assessed with ALDEx2 to identify differential gene abundances between humans and lemurs, accounting for sparseness and composition in the data. Differential abundance grouped at the gene family and the antibiotic class the gene is associated with conferring resistance against. The Aldex2 central log transformation (clr) function was used to generate final analyses. To test the effect of scaling on the outcome of differential abundance, a sensitivity analysis was performed using the function aldex, which introduces different levels of uncertainty using the gamma parameter.

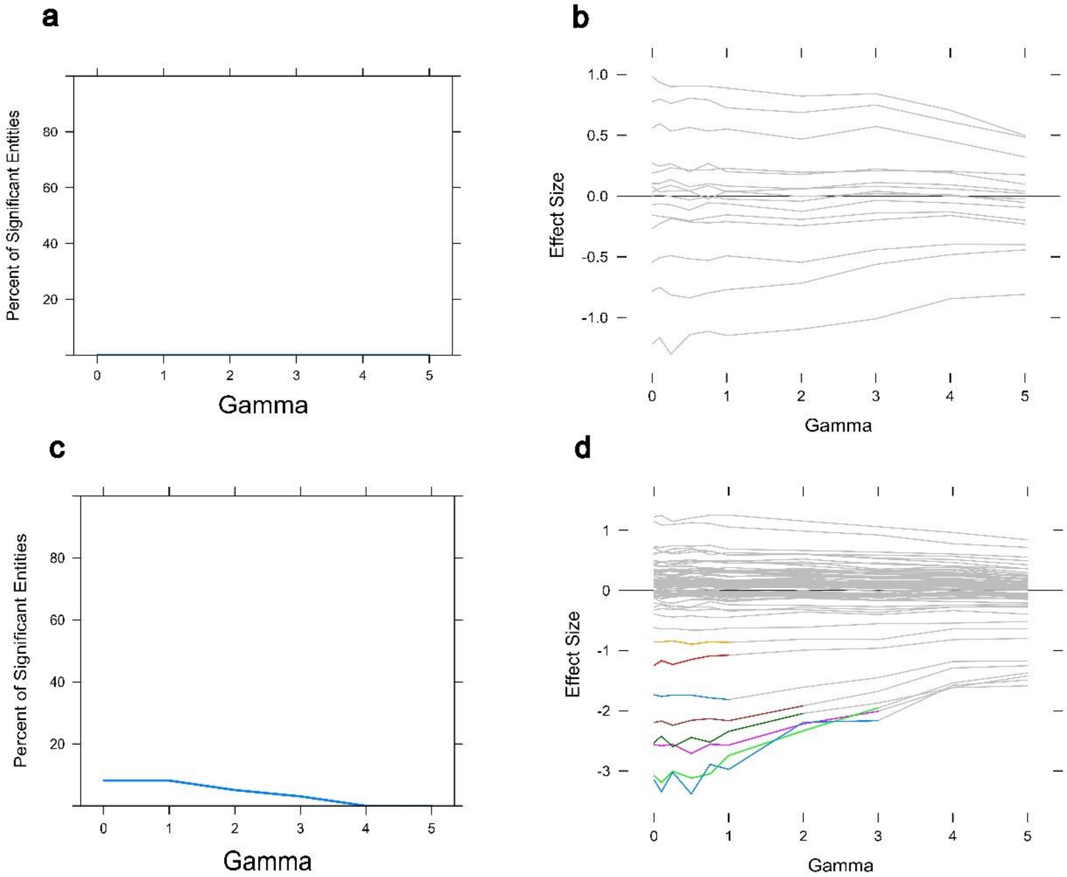

Figure S5.3. Sensitivity of detecting significant differential abundances at different levels of uncertainty using ALDEx2. Raw read abundances of ARGs were summed by either antibiotic class associated with resistance (a-b) or by the gene family associated with the positive allele hit (c-d). Plots were generated using the plotGamma function in ALDEx2 (v.1.35.0). (a) The percent of

significant entities for reads grouped by antibiotic class was not significant for any value of gamma. (b) Individual lines indicate unique entities of antibiotic classes and their corresponding effect size of differential abundance at different values of gamma. Gray indicates that an effect size is not significant. (c) The percent of significant entities for reads grouped by gene family was significant between gamma values of 0 and 3. (d) Individual colored segments indicate significant effect sizes for differential expression of eight unique gene families up through a gamma value of 1. Five entities remain significant up through gamma = 2, and two remain positive up through gamma = 3

Chapter 6: Conclusion

Summary and Discussion

The future of pathogenic disease detection and transmission relies on revealing pathogen genetic and evolutionary biology through genomics. Therefore, the goal of this thesis was to test the boundaries of genomics for pathogen surveillance in and outside of healthcare settings and identify opportunities to expand our understanding bacterial pathogen host adaptation. In Chapter 2, I advocated for a reconsideration of "SNP threshold" as "SNP ranges" to better account for the diverse evolutionary forces that produce genetic differences. In Chapter 3, I tested the strength of association between genetic distance and hospital exposure and identified stable clusters of putative transmission of MRSA. I looked at the issue of genetic clustering of MRSA strains further in Chapter 4 by identifying risks of relapsing strains and differentiating new infections from persistent infections occurring for the same host. Finally, in Chapter 5 I described the antimicrobial resistome for an under sampled geographic area and revealed that there are highly conserved AMR genes shared between humans and lemurs. Collectively, these chapters demonstrate how genetic distance, content, and context reveal the likelihood of bacterial strain transmission and persistence.

Mutations (in the broad sense DNA changes including SNPs, indels, and inversions) is the fundamental mechanism of generating genomic change in bacteria, which is why de novo mutation accumulation is foundational for estimating the amount of change at a calendar scale for epidemiological investigations. Mutation rates are dependent on the fidelity of the cellular replication machinery. Whether a new mutation ultimately survives in a bacterial population ("fixation") is dependent on the interplay of strength of selection and the chance of stochastic loss. Building phylogenies based on number of mutations accumulating at homologous sites common across bacterial genomes (the "core") is highly reliable for inferring evolutionary divergence, and this principle grounds for each chapter. However, comparative genomics at increasingly short time scales (days, for example) reveal evolutionary patterns associated with non-neutral selection for traits that affect pathogenesis. Additionally, there may ways in which the molecular clock can be confounded because of the genetic architecture and behavior of bacterial chromosomes, including selection for mutations that affect the DNA replication rate and homologous recombination which can bring in multiple novel genetic changes all in one event. In Chapter 3 we demonstrated that core SNP distances were stable when the population diversity was considered and showed that epidemiological exposures were more predictable among isolates with SNP distances smaller than 13. However, we also noticed that many clusters of a similar difference in time having widely different cumulative pairwise distances. Therefore, our evidence supports that SNP accumulation is a good marker for between host exposures, but more work is necessary to reveal associations of disease and transmission at even smaller time scales than our investigation period. One potential way to investigate time and SNP distance is to collect more infection colony samples to account for the genetic diversity. It is also important to monitor how the individual host factors can shift the strength of evolutionary forces driving genetic differences. In Chapter 4, we were able to look more closely at lineages that continually seed the blood of the same host and cause relapse and identified a trend in positive selection, not just neutral accumulation of SNPs. It is also documented that an invasive state can increase the mutation rate (80) and can result in specific mutations that are specific to pathogenesis (169). All the isolates in this study were derived from an invasive state, blood, and we detected convergent mutations in genes associated with bacteremia and antibiotic resistance. Therefore, non-neutral changes are important for identifying potential risks of relapse within a patient and could be used as markers that help reduce the burden of disease that could contribute to future outbreaks. We were limited to single bacterial isolates at each time point and likely missed some of the standing variability within a patients' infection. In a healthcare setting, it is possible to collect more than one isolated colony from a single patient infection and doing so can help pick up on rare genes that are of clinical interest for patient care and infection prevention within the clinic (279). Altogether, future models that generate thresholds of relatedness for cluster detection should incorporate information about expectations of the standing genomic variation ("the cloud of diversity")(79), screen for potential host-adaptation, and consider the patient state (a carrier or someone experiencing disease) to better identify clusters of infections with common epidemiological linkages.

Throughout this research, I demonstrated that genomic and metagenomic technologies are useful for simplifying signatures of pathogen detection. In a clinical setting, diagnosis of relapse requires many different pieces of information related to exposures, medical history, and is subject to differences in clinical discretion. The lack of concordance between clinical and genomic definitions of relapse in Chapter 4 demonstrates that often clinical information alone is not accurate enough to identify strains that caused a previous infection for the same person. Our evidence suggests that combining genomics with clinical information can help reduce the noise of some more complicated aspects of diagnosis or clue-in clinicians to persistent infections when the original infection source is unknown.

The relationship between resistant organisms spreading between humans and wildlife is an important topic, especially as changes in climate and land use change the geographic movement of humans and wildlife. In Chapter 5, we were limited in how we could characterize the overall network of interactions that could result in shared AMR genes in humans and lemurs. However, the complexity of this community network is not unlike the complexity of hospitalized patients who seem to cluster with related infections but do not have an apparent overlap, or who have loose overlaps that could suggest transmission. Given that there is increasing interest in using metagenomics at a diagnostic level, including the ability to detect novel pathogens and a lack of need to produce cultures, much can be learned about effective methods and limitations from studies involving unique ecosystems like the one in Chapter 5. For example, there is still much that needs to be done to relate the quantity of AMR genes detected to species spread, pathogenesis, and phenotypic expression. How to best normalize and quantify metagenomic sequences, and how many to sample, is still an ongoing dialogue.

As a technical argument, this body of work largely supports the impressive detective power that genomics has for clinical and public health practice. Why is it that the technology is not widely adopted for all pathogens in all spaces where disease detection occurs? Like the adoption of most technologies, it's a combination of issues. Cost and access to equipment still play a role in how much sequencing can be done, even as the overall price of next-generation sequencing technologies has decreased. The expertise to understand how to best use and analyze data also plays a role. Expertise in both evolutionary principles and comparative genomics is not typically in core curricula of accredited Schools of Public Health (280). The information in Chapter 2 offers an opportunity to fill in gaps in understanding for public health practitioners about biological mechanisms for genomic differences. Furthermore, sequences alone are not enough to come to good scientific conclusions or important decisions on healthcare. Infrastructure, including databases, information management systems, and server capacity, also impact who can use genomics for public health practice. Chapters 3, 4, and 5 all focused on sequences from a single community, but to answer important comparative analysis questions within each chapter, it was necessary to draw upon the curation of reference data available in public databases. As a global community, concerted efforts have been made to create large databases of sequences, which offers ongoing opportunities to query these databases and look for larger trends in pathogen spread and convergence of rare or complicated traits across species and/or strains. This requires interdisciplinary teams and increased support for technical infrastructure globally so that communities can put these data to public health use.

Future Directions

There is still much that can be learned about the genomic changes that occur during disease progression. Identifying or connecting phenotypic differences to genomic variation is frequently a vital next step when significant genetic mutations are detected. In this research, relapse of infection is a complex trait that we demonstrated does not have one singular strain background, nor one gene, and occur in different host niches, but nevertheless there is a general trend of positive selection on the whole genome. We also collected a comprehensive list of specific genes with known associations with bacteremia. This suggests that, overall, relapsing lineages, which are these persistent populations closely associated with the host, are responding and adapting in important ways that allow for ongoing survival. Two potential future directions for this work include exploration of changes in expression of genes and statistical modeling to identify environmental and genetic combinations that predict relapse.

To test if there are expression differences, a set of phenotypic tests could be done to assess the relationship between the mutations that arose in the same genes for multiple lineages. Some of these mutations occurred in correlation with previous antibiotic exposure and are known to confer a resistance phenotype. Others, however, did not demonstrate a specific association to a resistance phenotype. To test if these mutations confer different resistance profiles, future studies could be designed to challenge the ancestral strains without the mutation and the progeny strains with the mutations with different antibiotics to see if there is a difference in growth or survival. Several other genes also had mutations in multiple relapse lineages that are involved in virulence directly or in regulatory pathways. Here, future work could assess the association between these mutations and virulence expression by comparing them to strains that do not carry these mutations. Traits of special interest would be growth rate and important traits for persistence that contribute to the evolution of drug resistance, such as biofilm formation and adhesion.

Another opportunity for advancement is to perform a boosted regression tree analysis to assess the combination of pathogen and host traits that predict relapse infection. Boosted regression trees are useful for predictive analysis because they combine tree-based analysis with ensemble models. Decision trees are useful when there are complex or non-linear relationships between the predictors and the outcome of interest, and because their outcomes are easier to interpret without extensive training in statistics or mathematical modeling. Ensemble approaches, then, combine smaller, weaker "building blocks" together to create a new predictive model with more power. They work well for scenarios where the individual components contribute a weak effect (281). Boosted regression trees for this data would be useful not only based on the biological and clinical trends that I identified, but also because exploring this space can test the method's utility for downstream use and interpretation in a clinical setting, where quick decisions need to be made for patient care and infection control. Other machine learning approaches have shown utility in a clinical setting, especially for cluster detection (7). I have so far tested a boosted regression model to assess the demographic characteristics from patient data and in the future will add information about MRSA isolate strain and mutational profiles.

The evolution of pathogenesis and disease ecology are fundamental to epidemiological practice. Genomics and genetic sequencing expand the availability of information that practitioners can use to make epidemiological inferences. The expansion of larger sequencing datasets means that scientists can screen for convergence of disease-associated traits across a diverse landscape of hosts and environments. Future work should focus on evaluating and curating these data so that we can make new predictions about disease and find new solutions to prevent illness globally.

References

1. Thompson RA. Molecular epidemiology of infectious diseases. Arnold; 2000.
2. Riley Lee W., Blanton Ronald E., Sadowsky Michael. Advances in Molecular Epidemiology of Infectious Diseases: Definitions, Approaches, and Scope of the Field. Microbiol Spectr. 2018 Nov 2;6(6):6.6.01.
3. Eybpoosh S, Haghdoost AA, Mostafavi E, Bahrampour A, Azadmanesh K, Zolala F. Molecular epidemiology of infectious diseases. Electron Physician. 2017;9(8):5149.
4. Black A, MacCannell DR, Sibley TR, Bedford T. Ten recommendations for supporting open pathogen genomic analysis in public health. Nat Med. 2020 Jun 1;26(6):832–41.
5. Brown B, Allard M, Bazaco MC, Blankenship J, Minor T. An economic evaluation of the Whole Genome Sequencing source tracking program in the U.S. PLOS ONE. 2021 Oct 6;16(10):e0258262.
6. Martak D, Meunier A, Sauget M, Cholley P, Thouverez M, Bertrand X, et al. Comparison of pulsed-field gel electrophoresis and whole-genome-sequencing-based typing confirms the accuracy of pulsed-field gel electrophoresis for the investigation of local Pseudomonas aeruginosa outbreaks. J Hosp Infect. 2020 Aug 1;105(4):643–7.
7. Sundermann AJ, Chen J, Kumar P, Ayres AM, Cho ST, Ezeonwuka C, et al. Whole-Genome Sequencing Surveillance and Machine Learning of the Electronic Health Record for Enhanced Healthcare Outbreak Detection. Clin Infect Dis. 2021 Nov 12;ciab946.
8. Maiden MCJ, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol. 2013/09/02 ed. 2013 Oct;11(10):728–36.

9. Griffiths EJ, Timme RE, Page AJ, Alikhan NF, Fornika D, Maguire F, et al. The PHA4GE SARS-CoV-2 contextual data specification for open genomic epidemiology. 2020;

10. Fan Y, Feng Y, Xiao L. Comparative genomics: how has it advanced our knowledge of cryptosporidiosis epidemiology? Parasitol Res. 2019 Dec;118(12):3195–204.

11. Gordis L. The Dynamics of Disease Transmission. In: Epidemiology. 5th ed. Philidelphia, PA: Elsevier Saunders; p. 19–37.

12. Holmes AH, Moore LSP, Sundsfjord A, Steinbakk M, Regmi S, Karkey A, et al. Understanding the mechanisms and drivers of antimicrobial resistance. Lancet Lond Engl. 2016 Jan 9;387(10014):176–87.

13. Monserrat-Martinez A, Gambin Y, Sierecki E. Thinking Outside the Bug: Molecular Targets and Strategies to Overcome Antibiotic Resistance. Int J Mol Sci. 2019 Mar 13;20(6).

14. Cox G, Wright GD. Intrinsic antibiotic resistance: Mechanisms, origins, challenges and solutions. Spec Issue Antibiot Resist. 2013 Aug 1;303(6):287–92.

15. Marvig RL, Dolce D, Sommer LM, Petersen B, Ciofu O, Campana S, et al. Within-host microevolution of Pseudomonas aeruginosa in Italian cystic fibrosis patients. BMC Microbiol. 2015 Oct 19;15(1):218.

16. Buhl M, Kästle C, Geyer A, Autenrieth IB, Peter S, Willmann M. Molecular Evolution of Extensively Drug-Resistant (XDR) Pseudomonas aeruginosa Strains From Patients and Hospital Environment in a Prolonged Outbreak. Front Microbiol. 2019;10:1742.

17. Castillo-Ramírez S, Ghaly T, Gillings M. Non-clinical settings – the understudied facet of antimicrobial drug resistance. Environ Microbiol. 2021 Dec 1;23(12):7271–4.

18. Novick RP. Plasmid incompatibility. Microbiol Rev. 1987;51(4):381–95.

19. Evans DR, Griffith MP, Sundermann AJ, Shutt KA, Saul MI, Mustapha MM, et al. Systematic detection of horizontal gene transfer across genera among multidrug-resistant bacteria in a single hospital. eLife. 2020 Apr 14;9.

20. Baker KS, Dallman TJ, Thomson NR, Jenkins C. An outbreak of a rare Shiga-toxin-producing Escherichia coli serotype (O117:H7) among men who have sex with men. Microb Genomics. 2018 Jul;4(7).

21. Peng Z, Hu Z, Li Z, Zhang X, Jia C, Li T, et al. Antimicrobial resistance and population genomics of multidrug-resistant Escherichia coli in pig farms in mainland China. Nat Commun. 2022 Mar 2;13(1):1116.

22. Bryson D, Hettle AG, Boraston AB, Hobbs JK. Clinical Mutations That Partially Activate the Stringent Response Confer Multidrug Tolerance in Staphylococcus aureus. Antimicrob Agents Chemother. 2020 Feb 21;64(3).

23. Larsen J, Raisen CL, Ba X, Sadgrove NJ, Padilla-González GF, Simmonds MSJ, et al. Emergence of methicillin resistance predates the clinical use of antibiotics. Nature. 2022 Feb 1;602(7895):135–41.

24. Nadimpalli M, Delarocque-Astagneau E, Love DC, Price LB, Huynh BT, Collard JM, et al. Combating Global Antibiotic Resistance: Emerging One Health Concerns in Lower- and Middle-Income Countries. Clin Infect Dis. 2018 Mar 5;66(6):963–9.

25. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. Lancet Lond Engl. 2022 Feb 12;399(10325):629–55.

26. Liu J, Gefen O, Ronin I, Bar-Meir M, Balaban NQ. Effect of tolerance on the evolution of antibiotic resistance under drug combinations. Science. 2020;367(6474):200–4.

27. Murray LM, Hayes A, Snape J, Kasprzyk-Hordern B, Gaze WH, Murray AK. Co-selection for antibiotic resistance by environmental contaminants. Npj Antimicrob Resist. 2024 Apr 1;2(1):9.
28. Munk P, Brinch C, Møller FD, Petersen TN, Hendriksen RS, Seyfarth AM, et al. Genomic analysis of sewage from 101 countries reveals global landscape of antimicrobial resistance. Nat Commun. 2022 Dec 1;13(1):7251.
29. Bornbusch SL, Drea CM. Antibiotic resistance genes in lemur gut and soil microbiota along a gradient of anthropogenic disturbance. Front Ecol Evol. 2021;514.
30. Fang L, Chen C, Li S, Ye P, Shi Y, Sharma G, et al. A comprehensive and global evaluation of residual antibiotics in agricultural soils: Accumulation, potential ecological risks, and attenuation strategies. Ecotoxicol Environ Saf. 2023 Sep 1;262:115175.
31. Popovich KJ, Snitkin ES, Hota B, Green SJ, Pirani A, Aroutcheva A, et al. Genomic and Epidemiological Evidence for Community Origins of Hospital-Onset Methicillin-Resistant Staphylococcus aureus Bloodstream Infections. J Infect Dis. 2017;215(11):1640–7.
32. Montoya A, Schildhouse R, Goyal A, Mann JD, Snyder A, Chopra V, et al. How often are health care personnel hands colonized with multidrug-resistant organisms? A systematic review and meta-analysis. Am J Infect Control. 2019;47(6):693–703.
33. Raad I, Narro J, Khan A, Tarrand J, Vartivarian S, Bodey GP. Serious complications of vascular catheter-related Staphylococcus aureus bacteremia in cancer patients. Eur J Clin Microbiol Infect Dis Off Publ Eur Soc Clin Microbiol. 1992 Aug;11(8):675–82.
34. Moloney EM, Deasy EC, Swan JS, Brennan GI, O'Donnell MJ, Coleman DC. Whole-genome sequencing identifies highly related Pseudomonas aeruginosa strains in multiple washbasin U-bends at several locations in one hospital: evidence for trafficking of potential pathogens via wastewater pipes. J Hosp Infect. 2020 Apr 1;104(4):484–91.

35. Sundermann AJ, Chen J, Miller JK, Saul MI, Shutt KA, Griffith MP, et al. Outbreak of Pseudomonas aeruginosa Infections from a Contaminated Gastroscope Detected by Whole Genome Sequencing Surveillance. Clin Infect Dis Off Publ Infect Dis Soc Am. 2021 Aug 2;73(3):e638–42.

36. Roberts MC, Schwarz S. Tetracycline and Phenicol Resistance Genes and Mechanisms: Importance for Agriculture, the Environment, and Humans. J Environ Qual. 2016;45(2):576–92.

37. Chen S, Fu J, Zhao K, Yang S, Li C, Penttinen P, et al. Class 1 integron carrying qacEΔ1 gene confers resistance to disinfectant and antibiotics in Salmonella. Int J Food Microbiol. 2023 Jul 13;404:110319.

38. Region A, Region SEA, Region EM, Region WP. Global action plan on antimicrobial resistance. 2015;

39. Zay Ya K, Win PTN, Bielicki J, Lambiris M, Fink G. Association Between Antimicrobial Stewardship Programs and Antibiotic Use Globally: A Systematic Review and Meta-Analysis. JAMA Netw Open. 2023 Feb 9;6(2):e2253806–e2253806.

40. American Society for Microbiology. Policy Pathways to Combat the Global Crisis of Antimicrobial Resistance [Internet]. American Society for Microbiology; 2023 p. 19. Available from: https://asm.org/getmedia/5f665383-881a-493d-ae05-04a960a25548/AMR-Policy-Paper-2023.pdf

41. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res. 2020 Jan 8;48(D1):D517–25.

42. Ogston A. Micrococcus Poisoning. J Anat Physiol. 1882 Jul;16(Pt 4):526–67.

43. Rosenbach FJ, Rosenbach AJF. Mikro-organismen bei den Wund-infections-krankheiten des Menschen. JF Bergmann; 1884.

44. Park S, Ronholm J. Staphylococcus aureus in Agriculture: Lessons in Evolution from a Multispecies Pathogen. Clin Microbiol Rev. 2021 Mar 17;34(2).

45. Hofstetter K, Jacko NF, Gunoskey JJ, Talbot BM, Read TD, David MZ. 1377. Bloodstream and Skin Infection MRSA Isolates, 2019-2021: Strain Differences and Phylogenetic Clustering in a Single Health System. In Oxford University Press US; 2022. p. ofac492-1206.

46. Shands KN, Schmid GP, Dan BB, Blum D, Guidotti RJ, Hargrett NT, et al. Toxic-shock syndrome in menstruating women: association with tampon use and Staphylococcus aureus and clinical features in 52 cases. N Engl J Med. 1980 Dec 18;303(25):1436–42.

47. Decousser JW, Desroches M, Bourgeois-Nicolaos N, Potier J, Jehl F, Lina G, et al. Susceptibility trends including emergence of linezolid resistance among coagulase-negative staphylococci and meticillin-resistant Staphylococcus aureus from invasive infections. Int J Antimicrob Agents. 2015 Dec;46(6):622–30.

48. Jonsson IM, Arvidson S, Foster S, Tarkowski A. Sigma factor B and RsbU are required for virulence in Staphylococcus aureus-induced arthritis and sepsis. Infect Immun. 2004 Oct;72(10):6106–11.

49. Sharma-Kuinkel BK, Tkaczyk C, Bonnell J, Yu L, Tovchigrechko A, Tabor DE, et al. Associations of pathogen-specific and host-specific characteristics with disease outcome in patients with Staphylococcus aureus bacteremic pneumonia. Clin Transl Immunol. 2019;8(7):e01070.

50. Salvador VBD, Chapagain B, Joshi A, Brennessel DJ. Clinical Risk Factors for Infective Endocarditis in Staphylococcus aureus Bacteremia. Tex Heart Inst J. 2017 Feb;44(1):10–5.

51. Hennekinne JA, De Buyser ML, Dragacci S. Staphylococcus aureus and its food poisoning toxins: characterization and outbreak investigation. FEMS Microbiol Rev. 2012 Jul 1;36(4):815–36.

52. Popovich KJ, Green SJ, Okamoto K, Rhee Y, Hayden MK, Schoeny M, et al. MRSA Transmission in Intensive Care Units: Genomic Analysis of Patients, Their Environments, and Healthcare Workers. Clin Infect Dis Off Publ Infect Dis Soc Am. 2021 Jun 1;72(11):1879–87.

53. Hsu LY, Harris SR, Chlebowicz MA, Lindsay JA, Koh TH, Krishnan P, et al. Evolutionary dynamics of methicillin-resistant Staphylococcus aureus within a healthcare system. Genome Biol. 2015 Apr 23;16(1):81.

54. Centers for Disease Control and Prevention (U.S.), National Center for Emerging Zoonotic and Infectious Diseases (U.S.). Division of Healthcare Quality Promotion. Antibiotic Resistance Coordination and Strategy Unit ., editors. Antibiotic resistance threats in the United States, 2019. 2019; Available from: https://stacks.cdc.gov/view/cdc/82532

55. Salgado CD, Farr BM, Calfee DP. Community-acquired methicillin-resistant Staphylococcus aureus: a meta-analysis of prevalence and risk factors. Clin Infect Dis Off Publ Infect Dis Soc Am. 2003 Jan 15;36(2):131–9.

56. Souli M, Ruffin F, Choi SH, Park LP, Gao S, Lent NC, et al. Changing characteristics of Staphylococcus aureus bacteremia: results from a 21-year, prospective, longitudinal study. Clin Infect Dis. 2019;69(11):1868–77.

57. Yao Z, Peng Y, Chen X, Bi J, Li Y, Ye X, et al. Healthcare Associated Infections of Methicillin-Resistant Staphylococcus aureus: A Case-Control-Control Study. PLOS ONE. 2015 Oct 15;10(10):e0140604.

58. Junnila J, Hirvioja T, Rintala E, Auranen K, Rantakokko-Jalava K, Silvola J, et al. Changing epidemiology of methicillin-resistant Staphylococcus aureus in a low endemicity area-new challenges for MRSA control. Eur J Clin Microbiol Infect Dis Off Publ Eur Soc Clin Microbiol. 2020 Dec;39(12):2299–307.

59. Fan J, Shu M, Zhang G, Zhou W, Jiang Y, Zhu Y, et al. Biogeography and Virulence of Staphylococcus aureus. PLOS ONE. 2009 Jul 13;4(7):e6216.

60. Talan DA, Krishnadasan A, Gorwitz RJ, Fosheim GE, Limbago B, Albrecht V, et al. Comparison of Staphylococcus aureus From Skin and Soft-Tissue Infections in US Emergency Department Patients, 2004 and 2008. Clin Infect Dis. 2011 Jul 15;53(2):144–9.

61. Mediavilla JR, Chen L, Mathema B, Kreiswirth BN. Global epidemiology of community-associated methicillin resistant Staphylococcus aureus (CA-MRSA). Antimicrob • Genomics. 2012 Oct 1;15(5):588–95.

62. Ochoa SA, Cruz-Córdova A, Mancilla-Rojano J, Escalona-Venegas G, Esteban-Kenel V, Franco-Hernández I, et al. Control of Methicillin-Resistant Staphylococcus aureus Strains Associated With a Hospital Outbreak Involving Contamination From Anesthesia Equipment Using UV-C. Front Microbiol. 2020;11:3236.

63. Uhlemann AC, Otto M, Lowy FD, DeLeo FR. Evolution of community- and healthcare-associated methicillin-resistant Staphylococcus aureus. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis. 2014 Jan;21:563–74.

64. David MZ, Daum RS. Community-associated methicillin-resistant Staphylococcus aureus: epidemiology and clinical consequences of an emerging epidemic. Clin Microbiol Rev. 2010 Jul;23(3):616–87.

65. Fowler VGJ, Nelson CL, McIntyre LM, Kreiswirth BN, Monk A, Archer GL, et al. Potential associations between hematogenous complications and bacterial genotype in Staphylococcus aureus infection. J Infect Dis. 2007 Sep 1;196(5):738–47.

66. Klein E, Smith DL, Laxminarayan R. Community-associated methicillin-resistant Staphylococcus aureus in outpatients, United States, 1999–2006. Emerg Infect Dis. 2009;15(12):1925.

67. Roberts JC. Community-associated methicillin-resistant Staphylococcus aureus epidemic clone USA100; more than a nosocomial pathogen. Springerplus. 2013;2(1):1–3.

68. Otto M. Staphylococcus aureus toxins. Curr Opin Microbiol. 2014 Feb;17:32–7.

69. Hos NJ, Rieg S, Kern WV, Jonas D, Fowler VG, Higgins PG, et al. Amino acid alterations in fibronectin binding protein A (FnBPA) and bacterial genotype are associated with cardiac device related infection in Staphylococcus aureus bacteraemia. J Infect. 2015 Feb;70(2):153–9.

70. Thomer L, Schneewind O, Missiakas D. Pathogenesis of Staphylococcus aureus Bloodstream Infections. Annu Rev Pathol Mech Dis. 2016 May 23;11(1):343–64.

71. Rom JS, Beenken KE, Ramirez AM, Walker CM, Echols EJ, Smeltzer MS. Limiting protease production plays a key role in the pathogenesis of the divergent clinical isolates of Staphylococcus aureus LAC and UAMS-1. Virulence. 2021 Dec;12(1):584–600.

72. Beenken KE, Mrak LN, Zielinska AK, Atwood DN, Loughran AJ, Griffin LM, et al. Impact of the functional status of saeRS on in vivo phenotypes of Staphylococcus aureus sarA mutants. Mol Microbiol. 2014 Jun;92(6):1299–312.

73. Liu Y, Gao W, Yang J, Guo H, Zhang J, Ji Y. Contribution of Coagulase and Its Regulator SaeRS to Lethality of CA-MRSA 923 Bacteremia. Pathog Basel Switz. 2021 Oct 28;10(11).

74. Gill SR, McIntyre LM, Nelson CL, Remortel B, Rude T, Reller LB, et al. Potential Associations between Severity of Infection and the Presence of Virulence-Associated Genes in Clinical Strains of Staphylococcus aureus. PLOS ONE. 2011 Apr 26;6(4):e18673.

75. Su M, Lyles JT, Petit III RA, Peterson J, Hargita M, Tang H, et al. Genomic analysis of variability in Delta-toxin levels between Staphylococcus aureus strains. Hoyles L, editor. PeerJ. 2020 Mar 24;8:e8717.

76. Naimi TS, LeDell KH, Como-Sabetti K, Borchardt SM, Boxrud DJ, Etienne J, et al. Comparison of community-and health care–associated methicillin-resistant Staphylococcus aureus infection. Jama. 2003;290(22):2976–84.

77. Young BC, Wu CH, Gordon NC, Cole K, Price JR, Liu E, et al. Severe infections emerge from commensal bacteria by adaptive evolution. Holden MT, editor. eLife. 2017 Dec 19;6:e30637.

78. Read TD, Petit RA, Yin Z, Montgomery T, McNulty MC, David MZ. USA300 Staphylococcus aureus persists on multiple body sites following an infection. BMC Microbiol. 2018 Dec 5;18(1):206.

79. Coll F, Raven KE, Knight GM, Blane B, Harrison EM, Leek D, et al. Definition of a genetic relatedness cutoff to exclude recent transmission of meticillin-resistant Staphylococcus aureus: a genomic epidemiology analysis. Lancet Microbe. 2020 Dec;1(8):e328–35.

80. Murray GGR, Balmer AJ, Herbert J, Hadjirin NF, Kemp CL, Matuszewska M, et al. Mutation rate dynamics reflect ecological change in an emerging zoonotic pathogen. PLOS Genet. 2021 Nov 8;17(11):e1009864.

81. Planet PJ. Life After USA300: The Rise and Fall of a Superbug. J Infect Dis. 2017 Feb 15;215(suppl_1):S71–7.

82. Raghuram Vishnu, Petit Robert A., Karol Zach, Mehta Rohan, Weissman Daniel B., Read Timothy D. Average nucleotide identity-based Staphylococcus aureus strain grouping allows identification of strain-specific genes in the pangenome. mSystems. 2024 Jun 27;9(7):e00143-24.

83. Howden BP, Smith DJ, Mansell A, Johnson PDR, Ward PB, Stinear TP, et al. Different bacterial gene expression patterns and attenuated host immune responses are associated with the evolution of low-level vancomycin resistance during persistent methicillin-resistant Staphylococcus aureus bacteraemia. BMC Microbiol. 2008 Feb 27;8:39.

84. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res. 2018;3:124.

85. Raghuram V, Alexander AM, Loo HQ, Petit RA 3rd, Goldberg JB, Read TD. Species-Wide Phylogenomics of the Staphylococcus aureus Agr Operon Revealed Convergent Evolution of Frameshift Mutations. Microbiol Spectr. 2022 Feb 23;10(1):e0133421.

86. Uehara Y. Current Status of Staphylococcal Cassette Chromosome mec (SCCmec). Antibiot Basel Switz. 2022 Jan 11;11(1).

87. Frénay HM, Bunschoten AE, Schouls LM, van Leeuwen WJ, Vandenbroucke-Grauls CM, Verhoef J, et al. Molecular typing of methicillin-resistant Staphylococcus aureus on the basis of protein A gene polymorphism. Eur J Clin Microbiol Infect Dis Off Publ Eur Soc Clin Microbiol. 1996 Jan;15(1):60–4.

88. Köser CU, Holden MTG, Ellington MJ, Cartwright EJP, Brown NM, Ogilvy-Stuart AL, et al. Rapid Whole-Genome Sequencing for Investigation of a Neonatal MRSA Outbreak. N Engl J Med. 2012 Jun 14;366(24):2267–75.

89. Harris SR, Cartwright EJ, Török ME, Holden MT, Brown NM, Ogilvy-Stuart AL, et al. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. Lancet Infect Dis. 2013;13(2):130–6.

90. Eyre DW, Golubchik T, Gordon NC, Bowden R, Piazza P, Batty EM, et al. A pilot study of rapid benchtop sequencing of Staphylococcus aureusand Clostridium difficile for outbreak detection and surveillance. BMJ Open. 2012 Jan 1;2(3):e001124.

91. Oakeson KF, Wagner JM, Mendenhall M, Rohrwasser A, Atkinson-Dunn R. Bioinformatic Analyses of Whole-Genome Sequence Data in a Public Health Laboratory. Emerg Infect Dis. 2017 Sep;23(9):1441–5.

92. Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet. 2018 Jan 1;19(1):9–20.

93. Armstrong J, Fiddes IT, Diekhans M, Paten B. Whole-Genome Alignment and Comparative Annotation. Annu Rev Anim Biosci. 2019 Feb 15;7(1):41–64.

94. Stimson J, Gardy J, Mathema B, Crudu V, Cohen T, Colijn C. Beyond the SNP Threshold: Identifying Outbreak Clusters Using Inferred Transmissions. Mol Biol Evol. 2019 Mar 1;36(3):587–603.

95. Azarian T, Maraqa NF, Cook RL, Johnson JA, Bailey C, Wheeler S, et al. Genomic Epidemiology of Methicillin-Resistant Staphylococcus aureus in a Neonatal Intensive Care Unit. PLOS ONE. 2016 Oct 12;11(10):e0164397.

96. Zhongming Z, Linong L, Wangqiang Z, Wei L. Low quality healthcare is increasing the burden of illness and health costs globally. 2018;

97. Saleem Z, Godman B, Hassali MA, Hashmi FK, Azhar F, Rehman IU. Point prevalence surveys of health-care-associated infections: a systematic review. Pathog Glob Health. 2019/06/19 ed. 2019 Jun;113(4):191–205.

98. Shukla Sanjay K., Pantrang Madhulatha, Stahl Buffy, Briska Adam M., Stemper Mary E., Wagner Trevor K., et al. Comparative Whole-Genome Mapping To Determine Staphylococcus aureus Genome Size, Virulence Motifs, and Clonality. J Clin Microbiol. 2012 Nov 1;50(11):3526–33.

99. Poulsen BE, Yang R, Clatworthy AE, White T, Osmulski SJ, Li L, et al. Defining the core essential genome of <em>Pseudomonas aeruginosa</em>. Proc Natl Acad Sci. 2019 May 14;116(20):10072.

100. Senn Laurence, Clerc Olivier, Zanetti Giorgio, Basset Patrick, Prod'hom Guy, Gordon Nicola C., et al. The Stealthy Superbug: the Role of Asymptomatic Enteric Carriage in Maintaining a Long-Term Hospital Outbreak of ST228 Methicillin-Resistant Staphylococcus aureus. mBio. 7(1):e02039-15.

101. Harris Simon R., Feil Edward J., Holden Matthew T. G., Quail Michael A., Nickerson Emma K., Chantratita Narisara, et al. Evolution of MRSA During Hospital Transmission and Intercontinental Spread. Science. 2010 Jan 22;327(5964):469–74.

102. Cramer N, Klockgether J, Wrasman K, Schmidt M, Davenport CF, Tümmler B. Microevolution of the major common Pseudomonas aeruginosa clones C and PA14 in cystic fibrosis lungs. Environ Microbiol. 2011 Jul 1;13(7):1690–704.

103. Willmann M, Bezdan D, Zapata L, Susak H, Vogel W, Schröppel K, et al. Analysis of a long-term outbreak of XDR Pseudomonas aeruginosa: a molecular epidemiological study. J Antimicrob Chemother. 2015 May 1;70(5):1322–30.

104. Smith EE, Buckley DG, Wu Z, Saenphimmachak C, Hoffman LR, D'Argenio DA, et al. Genetic adaptation by <em>Pseudomonas aeruginosa</em> to the airways of cystic fibrosis patients. Proc Natl Acad Sci. 2006 May 30;103(22):8487.

105. Vella V, Galgani I, Polito L, Arora AK, Creech CB, David MZ, et al. Staphylococcus aureus Skin and Soft Tissue Infection Recurrence Rates in Outpatients: A Retrospective Database Study at 3 US Medical Centers. Clin Infect Dis Off Publ Infect Dis Soc Am. 2021 Sep 7;73(5):e1045–53.

106. Ward DV, Hoss AG, Kolde R, van Aggelen HC, Loving J, Smith SA, et al. Integration of genomic and clinical data augments surveillance of healthcare-acquired infections. Infect Control Hosp Epidemiol. 2019 Jun;40(6):649–55.

107. Kristinsdottir I, Haraldsson A, Thorkelsson T, Haraldsson G, Kristinsson KG, Larsen J, et al. MRSA outbreak in a tertiary neonatal intensive care unit in Iceland. Infect Dis Lond Engl. 2019 Dec;51(11–12):815–23.

108. Berbel Caban A, Pak TR, Obla A, Dupper AC, Chacko KI, Fox L, et al. PathoSPOT genomic epidemiology reveals under-the-radar nosocomial outbreaks. Genome Med. 2020 Nov 16;12(1):96.

109. Halstead FD, Quick J, Niebel M, Garvey M, Cumley N, Smith R, et al. Pseudomonas aeruginosa infection in augmented care: the molecular ecology and transmission dynamics in four large UK hospitals. J Hosp Infect. 2021 May;111:162–8.

110. Quick J, Cumley N, Wearn CM, Niebel M, Constantinidou C, Thomas CM, et al. Seeking the source of Pseudomonas aeruginosa infections in a recently opened hospital: an observational study using whole-genome sequencing. BMJ Open. 2014;4(11):e006278.

111. Diaz Caballero Julio, Clark Shawn T., Coburn Bryan, Zhang Yu, Wang Pauline W., Donaldson Sylva L., et al. Selective Sweeps and Parallel Pathoadaptation Drive Pseudomonas aeruginosa Evolution in the Cystic Fibrosis Lung. mBio. 6(5):e00981-15.

112. Magalhães B, Valot B, Abdelbary MMH, Prod'hom G, Greub G, Senn L, et al. Combining Standard Molecular Typing and Whole Genome Sequencing to Investigate Pseudomonas aeruginosa Epidemiology in Intensive Care Units. Front Public Health. 2020;8:3.

113. Parcell BJ, Oravcova K, Pinheiro M, Holden MTG, Phillips G, Turton JF, et al. Pseudomonas aeruginosa intensive care unit outbreak: winnowing of transmissions with molecular and genomic typing. J Hosp Infect. 2018 Mar;98(3):282–8.

114. Snyder L, Loman N, Faraj L, Levi K, Weinstock G, Boswell T, et al. Epidemiological investigation of Pseudomonas aeruginosa isolates from a six-year-long hospital outbreak using high-throughput whole genome sequencing. Eurosurveillance. 2013;18(42):20611.

115. Blanc DS, Magalhães B, Koenig I, Senn L, Grandbastien B. Comparison of Whole Genome (wg-) and Core Genome (cg-) MLST (BioNumerics(TM)) Versus SNP Variant Calling for Epidemiological Investigation of Pseudomonas aeruginosa. Front Microbiol. 2020;11:1729.

116. Davis RJ, Jensen SO, Van Hal S, Espedido B, Gordon A, Farhat R, et al. Whole Genome Sequencing in Real-Time Investigation and Management of a Pseudomonas aeruginosa Outbreak on a Neonatal Intensive Care Unit. Infect Control Hosp Epidemiol. 2015 Sep;36(9):1058–64.

117. Talbot BM, Jacko NF, Petit RA, Pegues DA, Shumaker MJ, Read TD, et al. Unsuspected clonal spread of Methicillin-resistant <em>Staphylococcus aureus</em> causing bloodstream infections in hospitalized adults detected using whole genome sequencing. medRxiv. 2021 Jan 1;2021.12.23.21268338.

118. Kimura M. Evolutionary rate at the molecular level. Nature. 1968;217(5129):624–6.

119. Oliver A, Cantón R, Campo P, Baquero F, Blázquez J. High frequency of hypermutable Pseudomonas aeruginosa in cystic fibrosis lung infection. Science. 2000 May 19;288(5469):1251–4.

120. Kenna DT, Doherty CJ, Foweraker J, Macaskill L, Barcus VA, Govan JRW. Hypermutability in environmental Pseudomonas aeruginosa and in populations causing pulmonary infection in individuals with cystic fibrosis. Microbiol Read Engl. 2007 Jun;153(Pt 6):1852–9.

121. Didelot X, Maiden MC. Impact of recombination on bacterial evolution. Trends Microbiol. 2010;18(7):315–22.

122. González-Torres Pedro, Rodríguez-Mateos Francisco, Antón Josefa, Gabaldón Toni, Heitman Joseph. Impact of Homologous Recombination on the Evolution of Prokaryotic Core Genomes. mBio. 10(1):e02494-18.

123. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2015;43(3):e15–e15.

124. Didelot X, Wilson DJ. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. PLOS Comput Biol. 2015 Feb 12;11(2):e1004041.

125. 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 May 1;34(5):1167–82.

126. Kishimoto T, Ying BW, Tsuru S, Iijima L, Suzuki S, Hashimoto T, et al. Molecular Clock of Neutral Mutations in a Fitness-Increasing Evolutionary Process. PLOS Genet. 2015 Jul 15;11(7):e1005392.

127. Young BC, Golubchik T, Batty EM, Fung R, Larner-Svensson H, Votintseva AA, et al. Evolutionary dynamics of <em>Staphylococcus aureus</em> during progression from carriage to disease. Proc Natl Acad Sci. 2012 Mar 20;109(12):4550.

128. Sullivan Mitchell J., Altman Deena R., Chacko Kieran I., Ciferri Brianne, Webster Elizabeth, Pak Theodore R., et al. A Complete Genome Screening Program of Clinical Methicillin-Resistant Staphylococcus aureus Isolates Identifies the Origin and Progression of a Neonatal Intensive Care Unit Outbreak. J Clin Microbiol. 57(12):e01261-19.

129. Golubchik T, Batty EM, Miller RR, Farr H, Young BC, Larner-Svensson H, et al. Within-Host Evolution of Staphylococcus aureus during Asymptomatic Carriage. PLOS ONE. 2013 May 1;8(5):e61319.

130. Lees JA, Kendall M, Parkhill J, Colijn C, Bentley SD, Harris SR. Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study. Wellcome Open Res. 2018;3:33.

131. Bayliss SC, Thorpe HA, Coyle NM, Sheppard SK, Feil EJ. PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. GigaScience. 2019 Oct 1;8(10):giz119.

132. Kourtis AP, Hatfield K, Baggs J, Mu Y, See I, Epson E, et al. Vital Signs: Epidemiology and Recent Trends in Methicillin-Resistant and in Methicillin-Susceptible Staphylococcus aureus Bloodstream Infections - United States. MMWR Morb Mortal Wkly Rep. 2019 Mar 8;68(9):214–9.

133. Raineri EJM, Altulea D, van Dijl JM. Staphylococcal trafficking and infection—from ‘nose to gut’ and back. FEMS Microbiol Rev [Internet]. 2021 Jul 14 [cited 2021 Aug 6];(fuab041). Available from: https://doi.org/10.1093/femsre/fuab041

134. Yang ES, Tan J, Eells S, Rieg G, Tagudar G, Miller LG. Body site colonization in patients with community-associated methicillin-resistant Staphylococcus aureus and other types of S. aureus skin infections. Clin Microbiol Infect. 2010 May 1;16(5):425–31.

135. Methicillin-resistant staphylococcus aureus infections among competitive sports participants--Colorado, Indiana, Pennsylvania, and Los Angeles County, 2000-2003. MMWR Morb Mortal Wkly Rep. 2003 Aug 22;52(33):793–5.

136. Marks LR, Calix JJ, Wildenthal JA, Wallace MA, Sawhney SS, Ransom EM, et al. Staphylococcus aureus injection drug use-associated bloodstream infections are propagated by community outbreaks of diverse lineages. Commun Med. 2021 Nov 30;1(1):52.

137. Leopold Shana R., Goering Richard V., Witten Anika, Harmsen Dag, Mellmann Alexander, Tang Y.-W. Bacterial Whole-Genome Sequencing Revisited: Portable, Scalable, and Standardized Analysis for Typing and Detection of Virulence and Antibiotic Resistance Genes. J Clin Microbiol. 2014 Jul 1;52(7):2365–70.

138. Armstrong GL, MacCannell DR, Taylor J, Carleton HA, Neuhaus EB, Bradbury RS, et al. Pathogen genomics in public health. N Engl J Med. 2019;381(26):2569–80.

139. CLSI. Performance Standards for Antimicrobial Susceptibility Testing. Wayne, PA: Clinical and Laboratory Standards Institute; 2019. Report No.: CLSI supplement M100.

140. Petit Robert A., Read Timothy D., Segata Nicola. Bactopia: a Flexible Pipeline for Complete Analysis of Bacterial Genomes. mSystems. 5(4):e00190-20.

141. Seemann T. snippy: fast bacterial variant calling from NGS reads [Internet]. 2015. Available from: https://github.com/tseemann/snippy

142. 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 Nov 19;15(11):524.

143. Seemann T. Source code for snp-dists software [Internet]. Zenodo; 2018. Available from: https://doi.org/10.5281/zenodo.1411986

144. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020 May 1;37(5):1530–4.

145. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol. 2017 Jan 1;8(1):28–36.

146. Jombart T. Assessing the quality of a phylogeny. In: Introduction to phylogenetics using R [Internet]. 2016 [cited 2021 Dec 9]. p. 16–7. Available from: https://adegenet.r-forge.r-project.org/files/Glasgow2015/practical-introphylo.1.0.pdf

147. Briand S, Dessimoz C, El-Mabrouk N, Lafond M, Lobinska G. A generalized Robinson-Foulds distance for labeled trees. BMC Genomics. 2020 Nov 18;21(10):779.

148. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinforma Oxf Engl. 2019 Feb 1;35(3):526–8.

149. RStudio Team,. RStudio: Integrated Development Environment for R [Internet]. RStudio,PBC; 2021. Available from: http://www.rstudio.com/

150. Inkscape Project. Inkscape [Internet]. 2020. Available from: https://inkscape.org

151. Bowers JR, Driebe EM, Albrecht V, McDougal LK, Granade M, Roe CC, et al. Improved Subtyping of Staphylococcus aureus Clonal Complex 8 Strains Based on Whole-Genome Phylogenetic Analysis. mSphere. 2018 Jun;3(3).

152. Smith JT, Eckhardt EM, Hansel NB, Eliato TR, Martin IW, Andam CP. Genomic epidemiology of methicillin-resistant and -susceptible Staphylococcus aureus from bloodstream infections. BMC Infect Dis. 2021 Jun 21;21(1):589.

153. Slingerland BCGC, Vos MC, Bras W, Kornelisse RF, De Coninck D, van Belkum A, et al. Whole-genome sequencing to explore nosocomial transmission and virulence in neonatal methicillin-susceptible Staphylococcus aureus bacteremia. Antimicrob Resist Infect Control. 2020 Feb 22;9(1):39.

154. Kanamori H, Rutala WA, Weber DJ. The Role of Patient Care Items as a Fomite in Healthcare-Associated Outbreaks and Infection Prevention. Clin Infect Dis Off Publ Infect Dis Soc Am. 2017 Oct 15;65(8):1412–9.

155. Mattner F, Biertz F, Ziesing S, Gastmeier P, Chaberny IF. Long-term persistence of MRSA in re-admitted patients. Infection. 2010 Oct 1;38(5):363–71.

156. Gorrie CL, Da Silva AG, Ingle DJ, Higgs C, Seemann T, Stinear TP, et al. Systematic analysis of key parameters for genomics-based real-time detection and tracking of multidrug-resistant bacteria. bioRxiv [Internet]. 2020; Available from: https://www.biorxiv.org/content/early/2020/09/25/2020.09.24.310821

157. Minter DJ, Appa A, Chambers HF, Doernberg SB. Contemporary Management of Staphylococcus aureus Bacteremia-Controversies in Clinical Practice. Clin Infect Dis Off Publ Infect Dis Soc Am. 2023 Nov 30;77(11):e57–68.

158. Gao W, Chua K, Davies JK, Newton HJ, Seemann T, Harrison PF, et al. Two novel point mutations in clinical Staphylococcus aureus reduce linezolid susceptibility and switch on the stringent response to promote persistent infection. PLoS Pathog. 2010 Jun 10;6(6):e1000944.

159. Holmes NE, Turnidge JD, Munckhof WJ, Robinson JO, Korman TM, O'Sullivan MVN, et al. Genetic and molecular predictors of high vancomycin MIC in Staphylococcus aureus bacteremia isolates. J Clin Microbiol. 2014 Sep;52(9):3384–93.

160. Gao W, Cameron DR, Davies JK, Kostoulias X, Stepnell J, Tuck KL, et al. The RpoB $H_{481}Y$ rifampicin resistance mutation and an active stringent response reduce virulence and increase resistance to innate immune responses in Staphylococcus aureus. J Infect Dis. 2013 Mar 15;207(6):929–39.

161. Wertheim HFL, Vos MC, Ott A, van Belkum A, Voss A, Kluytmans JAJW, et al. Risk and outcome of nosocomial Staphylococcus aureus bacteraemia in nasal carriers versus non-carriers. Lancet Lond Engl. 2004 Aug 21;364(9435):703–5.

162. Hofstetter KS, Jacko NF, Shumaker MJ, Talbot BM, Petit RA III, Read TD, et al. Strain Differences in Bloodstream and Skin Infection: Methicillin-Resistant Staphylococcus aureus Isolated in 2018–2021 in a Single Health System. Open Forum Infect Dis. 2024 Jun 1;11(6):ofae261.

163. Small PM, Chambers HF. Vancomycin for Staphylococcus aureus endocarditis in intravenous drug users. Antimicrob Agents Chemother. 1990 Jun;34(6):1227–31.

164. Ehni WF, Reller LB. Short-course therapy for catheter-associated Staphylococcus aureus bacteremia. Arch Intern Med. 1989 Mar;149(3):533–6.

165. Choi SH, Dagher M, Ruffin F, Park LP, Sharma-Kuinkel BK, Souli M, et al. Risk Factors for Recurrent Staphylococcus aureus Bacteremia. Clin Infect Dis. 2021 Jun 1;72(11):1891–9.

166. Fowler VGJ, Kong LK, Corey GR, Gottlieb GS, McClelland RS, Sexton DJ, et al. Recurrent Staphylococcus aureus bacteremia: pulsed-field gel electrophoresis findings in 29 patients. J Infect Dis. 1999 May;179(5):1157–61.

167. Welsh KJ, Skrobarcek KA, Abbott AN, Lewis CT, Kruzel MC, Lewis EM, et al. Predictors of relapse of methicillin-resistant Staphylococcus aureus bacteremia after treatment with vancomycin. J Clin Microbiol. 2011 Oct;49(10):3669–72.

168. Chang FY, MacDonald BB, Peacock JEJ, Musher DM, Triplett P, Mylotte JM, et al. A prospective multicenter study of Staphylococcus aureus bacteremia: incidence of endocarditis, risk factors for mortality, and clinical impact of methicillin resistance. Medicine (Baltimore). 2003 Sep;82(5):322–32.

169. Giulieri SG, Guérillot R, Duchene S, Hachani A, Daniel D, Seemann T, et al. Niche-specific genome degradation and convergent evolution shaping Staphylococcus aureus adaptation during severe infections. eLife. 2022 Jun 14;11.

170. Giulieri SG, Baines SL, Guerillot R, Seemann T, Gonçalves da Silva A, Schultz M, et al. Genomic exploration of sequential clinical isolates reveals a distinctive molecular signature of persistent Staphylococcus aureus bacteraemia. Genome Med. 2018 Aug 23;10(1):65.

171. Hachani A, Giulieri SG, Guérillot R, Walsh CJ, Herisse M, Soe YM, et al. A high-throughput cytotoxicity screening platform reveals agr-independent mutations in bacteraemia-associated Staphylococcus aureus that promote intracellular persistence. eLife. 2023 Jun 8;12.

172. Benoit JB, Frank DN, Bessesen MT. Genomic evolution of Staphylococcus aureus isolates colonizing the nares and progressing to bacteremia. PloS One. 2018;13(5):e0195860.

173. Nygaard TK, Pallister KB, Ruzevich P, Griffith S, Vuong C, Voyich JM. SaeR binds a consensus sequence within virulence gene promoters to advance USA300 pathogenesis. J Infect Dis. 2010 Jan 15;201(2):241–54.

174. Altman DR, Sullivan MJ, Chacko KI, Balasubramanian D, Pak TR, Sause WE, et al. Genome Plasticity of agr-Defective Staphylococcus aureus during Clinical Infection. Infect Immun. 2018 Oct;86(10).

175. Tsuji BT, Harigaya Y, Lesse AJ, Sakoulas G, Mylotte JM. Loss of vancomycin bactericidal activity against accessory gene regulator (agr) dysfunctional Staphylococcus aureus under conditions of high bacterial density. Diagn Microbiol Infect Dis. 2009 Jun;64(2):220–4.

176. Cheung AL, Eberhardt KJ, Chung E, Yeaman MR, Sullam PM, Ramos M, et al. Diminished virulence of a sar-/agr- mutant of Staphylococcus aureus in the rabbit model of endocarditis. J Clin Invest. 1994 Nov;94(5):1815–22.

177. Chong YP, Kim ES, Park SJ, Park KH, Kim T, Kim MN, et al. Accessory gene regulator (agr) dysfunction in Staphylococcus aureus bloodstream isolates from South Korean patients. Antimicrob Agents Chemother. 2013 Mar;57(3):1509–12.

178. Villar M, Marimón JM, García-Arenzana JM, de la Campa AG, Ferrándiz MJ, Pérez-Trallero E. Epidemiological and molecular aspects of rifampicin-resistant Staphylococcus aureus isolated from wounds, blood and respiratory samples. J Antimicrob Chemother. 2011 May;66(5):997–1000.

179. Bæk KT, Thøgersen L, Mogenssen RG, Mellergaard M, Thomsen LE, Petersen A, et al. Stepwise decrease in daptomycin susceptibility in clinical Staphylococcus aureus isolates associated with an initial mutation in rpoB and a compensatory inactivation of the clpX gene. Antimicrob Agents Chemother. 2015 Nov;59(11):6983–91.

180. de Souza DC, Cogo LL, Palmeiro JK, Dalla-Costa LM, de Oliveira Tomaz AP, Riedi CA, et al. Thymidine-auxotrophic Staphylococcus aureus small-colony variant bacteremia in a patient with cystic fibrosis. Pediatr Pulmonol. 2020 Jun;55(6):1388–93.

181. Douglas EJA, Duggan S, Brignoli T, Massey RC. The MpsB protein contributes to both the toxicity and immune evasion capacity of Staphylococcus aureus. Microbiol Read Engl. 2021 Oct;167(10).

182. Nakano Y, Murata M, Matsumoto Y, Toyoda K, Ota A, Yamasaki S, et al. Clinical characteristics and factors related to infection with SCCmec type II and IV Methicillin-resistant Staphylococcus aureus in a Japanese secondary care facility: a single-center retrospective study. J Glob Antimicrob Resist. 2022 Dec;31:355–62.

183. Young BC, Wu CH, Charlesworth J, Earle S, Price JR, Gordon NC, et al. Antimicrobial resistance determinants are associated with Staphylococcus aureus bacteraemia and adaptation to the healthcare environment: a bacterial genome-wide association study. Microb Genomics. 2021 Nov;7(11).

184. Culyba MJ, Van Tyne D. Bacterial evolution during human infection: Adapt and live or adapt and die. PLoS Pathog. 2021 Sep;17(9):e1009872.

185. CLSI. Performance Standards for Antimicrobial Susceptibility Testing. 29th ed. CLSI supplement M100. Wayne, PA: Clinical and Laboratory Standards Institute; 2019.

186. David MZ, Daum RS. Treatment of Staphylococcus aureus Infections. In: Bagnoli F, Rappuoli R, Grandi G, editors. Staphylococcus aureus: Microbiology, Pathology, Immunology, Therapy and Prophylaxis [Internet]. Cham: Springer International Publishing; 2017. p. 325–83. Available from: https://doi.org/10.1007/82_2017_42

187. Talbot BM, Jacko NF, Petit III RA, Pegues DA, Shumaker MJ, Read TD, et al. Unsuspected clonal spread of methicillin-resistant Staphylococcus aureus causing bloodstream infections in hospitalized adults detected using whole genome sequencing. Clin Infect Dis. 2022;75(12):2104–12.

188. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017 Jun 1;14(6):587–9.

189. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46.

190. Hazra A, Gogtay N. Biostatistics Series Module 7: The Statistics of Diagnostic Tests. Indian J Dermatol. 2017 Feb;62(1):18–24.

191. Joseph SJ, Li B, Petit Iii RA, Qin ZS, Darrow L, Read TD. The single-species metagenome: subtyping Staphylococcus aureus core genome sequences from shotgun metagenomic data. PeerJ. 2016;4:e2571.

192. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. Microb Genomics. 2021 Nov;7(11).

193. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature. 1991 Jun 20;351(6328):652–4.

194. Abbas M, Rossel A, de Kraker MEA, von Dach E, Marti C, Emonet S, et al. Association between treatment duration and mortality or relapse in adult patients with Staphylococcus aureus bacteraemia: a retrospective cohort study. Clin Microbiol Infect Off Publ Eur Soc Clin Microbiol Infect Dis. 2020 May;26(5):626–31.

195. Horino T, Hori S. Metastatic infection during Staphylococcus aureus bacteremia. J Infect Chemother. 2020 Feb 1;26(2):162–9.

196. Bouiller K, Jacko NF, Shumaker MJ, Talbot BM, Read TD, David MZ. Factors associated with foreign body infection in methicillin-resistant Staphylococcus aureus bacteremia. Front Immunol. 2024;15:1335867.

197. Chen CJ, Huang YC, Chiu CH. Multiple pathways of cross-resistance to glycopeptides and daptomycin in persistent MRSA bacteraemia. J Antimicrob Chemother. 2015 Nov;70(11):2965–72.

198. Thitiananpakorn K, Aiba Y, Tan XE, Watanabe S, Kiga K, Sato'o Y, et al. Association of mprF mutations with cross-resistance to daptomycin and vancomycin in methicillin-resistant Staphylococcus aureus (MRSA). Sci Rep. 2020 Sep 30;10(1):16107.

199. Watanabe Yukiko, Cui Longzhu, Katayama Yuki, Kozue Kishii, Hiramatsu Keiichi. Impact of rpoB Mutations on Reduced Vancomycin Susceptibility in Staphylococcus aureus. J Clin Microbiol. 2020 Dec 21;49(7):2680–4.

200. Strauß M, Vitiello C, Schweimer K, Gottesman M, Rösch P, Knauer SH. Transcription is regulated by NusA:NusG interaction. Nucleic Acids Res. 2016 Jul 8;44(12):5971–82.

201. Miller WR, Bayer AS, Arias CA. Mechanism of Action and Resistance to Daptomycin in Staphylococcus aureus and Enterococci. Cold Spring Harb Perspect Med. 2016 Nov 1;6(11).

202. Diep BA, Stone GG, Basuino L, Graber CJ, Miller A, des Etages SA, et al. The arginine catabolic mobile element and staphylococcal chromosomal cassette mec linkage: convergence of virulence and resistance in the USA300 clone of methicillin-resistant Staphylococcus aureus. J Infect Dis. 2008 Jun 1;197(11):1523–30.

203. Young BC, Golubchik T, Batty EM, Fung R, Larner-Svensson H, Votintseva AA, et al. Evolutionary dynamics of Staphylococcus aureus during progression from carriage to disease. Proc Natl Acad Sci U S A. 2012 Mar 20;109(12):4550–5.

204. Iwata Y, Satou K, Furuichi K, Yoneda I, Matsumura T, Yutani M, et al. Collagen adhesion gene is associated with bloodstream infections caused by methicillin-resistant Staphylococcus aureus. Int J Infect Dis IJID Off Publ Int Soc Infect Dis. 2020 Feb;91:22–31.

205. Courjon J, Munro P, Benito Y, Visvikis O, Bouchiat C, Boyer L, et al. EDIN-B Promotes the Translocation of Staphylococcus aureus to the Bloodstream in the Course of Pneumonia. Toxins. 2015 Oct 15;7(10):4131–42.

206. Wang Y, Hu M, Liu Q, Qin J, Dai Y, He L, et al. Role of the ESAT-6 secretion system in virulence of the emerging community-associated Staphylococcus aureus lineage ST398. Sci Rep. 2016 Apr 26;6:25163.

207. Lannergård J, Norström T, Hughes D. Genetic determinants of resistance to fusidic acid among clinical bacteremia isolates of Staphylococcus aureus. Antimicrob Agents Chemother. 2009 May;53(5):2059–65.

208. Malachowa N, Whitney AR, Kobayashi SD, Sturdevant DE, Kennedy AD, Braughton KR, et al. Global changes in Staphylococcus aureus gene expression in human blood. PloS One. 2011 Apr 15;6(4):e18617.

209. Kropec A, Maira-Litran T, Jefferson KK, Grout M, Cramton SE, Götz F, et al. Poly-N-acetylglucosamine production in Staphylococcus aureus is essential for virulence in murine models of systemic infection. Infect Immun. 2005 Oct;73(10):6868–76.

210. Lagos J, Alarcón P, Benadof D, Ulloa S, Fasce R, Tognarelli J, et al. Novel nonsense mutation in the katA gene of a catalase-negative Staphylococcus aureus strain. Braz J Microbiol Publ Braz Soc Microbiol. 2016 Mar;47(1):177–80.

211. Alonzo F 3rd, Benson MA, Chen J, Novick RP, Shopsin B, Torres VJ. Staphylococcus aureus leucocidin ED contributes to systemic infection by targeting neutrophils and promoting bacterial growth in vivo. Mol Microbiol. 2012 Jan;83(2):423–35.

212. Li L, Wang G, Cheung A, Abdelhady W, Seidl K, Xiong YQ. MgrA Governs Adherence, Host Cell Interaction, and Virulence in a Murine Model of Bacteremia Due to Staphylococcus aureus. J Infect Dis. 2019 Aug 9;220(6):1019–28.

213. Rom JS, Atwood DN, Beenken KE, Meeker DG, Loughran AJ, Spencer HJ, et al. Impact of Staphylococcus aureus regulatory mutations that modulate biofilm formation in the USA300 strain LAC on virulence in a murine bacteremia model. Virulence. 2017 Nov 17;8(8):1776–90.

214. Ji S, Jiang S, Wei X, Sun L, Wang H, Zhao F, et al. In-Host Evolution of Daptomycin Resistance and Heteroresistance in Methicillin-Resistant Staphylococcus aureus Strains From Three Endocarditis Patients. J Infect Dis. 2020 Mar 16;221(Suppl 2):S243–52.

215. Duggan S, Laabei M, Alnahari AA, O'Brien EC, Lacey KA, Bacon L, et al. A Small Membrane Stabilizing Protein Critical to the Pathogenicity of Staphylococcus aureus. Infect Immun. 2020 Aug 19;88(9).

216. Aoyagi T, Kaito C, Sekimizu K, Omae Y, Saito Y, Mao H, et al. Impact of psm-mec in the mobile genetic element on the clinical characteristics and outcome of SCCmec-II methicillin-resistant Staphylococcus aureus bacteraemia in Japan. Clin Microbiol Infect Off Publ Eur Soc Clin Microbiol Infect Dis. 2014 Sep;20(9):912–9.

217. Goncheva MI, Flannagan RS, Sterling BE, Laakso HA, Friedrich NC, Kaiser JC, et al. Stress-induced inactivation of the Staphylococcus aureus purine biosynthesis repressor leads to hypervirulence. Nat Commun. 2019 Feb 15;10(1):775.

218. Alkam D, Jenjaroenpun P, Ramirez AM, Beenken KE, Spencer HJ, Smeltzer MS. The Increased Accumulation of Staphylococcus aureus Virulence Factors Is Maximized in a purR Mutant by the Increased Production of SarA and Decreased Production of Extracellular Proteases. Infect Immun. 2021 Mar 17;89(4).

219. Diep BA, Palazzolo-Ballance AM, Tattevin P, Basuino L, Braughton KR, Whitney AR, et al. Contribution of Panton-Valentine leukocidin in community-associated methicillin-resistant Staphylococcus aureus pathogenesis. PloS One. 2008 Sep 12;3(9):e3198.

220. Chen E, Shaffer MG, Bilodeau RE, West RE 3rd, Oberly PJ, Nolin TD, et al. Clinical rel mutations in Staphylococcus aureus prime pathogen expansion under nutrient stress. mSphere. 2023 Oct 24;8(5):e0024923.

221. Balasubramanian D, Ohneck EA, Chapman J, Weiss A, Kim MK, Reyes-Robles T, et al. Staphylococcus aureus Coordinates Leukocidin Expression and Pathogenesis by Sensing Metabolic Fluxes via RpiRc. mBio. 2016 Jun 21;7(3).

222. Suligoy CM, Díaz RE, Gehrke AK, Ring N, Yebra G, Alves J, et al. Acapsular Staphylococcus aureus with a non-functional agr regains capsule expression after passage through the bloodstream in a bacteremia mouse model. Sci Rep. 2020 Aug 24;10(1):14108.

223. Das S, Lindemann C, Young BC, Muller J, Österreich B, Ternette N, et al. Natural mutations in a <em>Staphylococcus aureus</em> virulence regulator attenuate cytotoxicity but permit bacteremia and abscess formation. Proc Natl Acad Sci. 2016 May 31;113(22):E3101.

224. Voyich JM, Vuong C, DeWald M, Nygaard TK, Kocianova S, Griffith S, et al. The SaeR/S gene regulatory system is essential for innate immune evasion by Staphylococcus aureus. J Infect Dis. 2009 Jun 1;199(11):1698–706.

225. Zielinska AK, Beenken KE, Mrak LN, Spencer HJ, Post GR, Skinner RA, et al. sarA-mediated repression of protease production plays a key role in the pathogenesis of Staphylococcus aureus USA300 isolates. Mol Microbiol. 2012 Dec;86(5):1183–96.

226. Vrieling M, Tuffs SW, Yebra G, van Smoorenburg MY, Alves J, Pickering AC, et al. Population Analysis of Staphylococcus aureus Reveals a Cryptic, Highly Prevalent Superantigen SElW That Contributes to the Pathogenesis of Bacteremia. mBio. 2020 Oct 27;11(5).

227. Wang L, Bi C, Wang T, Xiang H, Chen F, Hu J, et al. A coagulase-negative and non-haemolytic strain of Staphylococcus aureus for investigating the roles of SrtA in a murine model of bloodstream infection. Pathog Dis. 2015 Aug;73(6):ftv042.

228. Atwood DN, Beenken KE, Loughran AJ, Meeker DG, Lantz TL, Graham JW, et al. XerC Contributes to Diverse Forms of Staphylococcus aureus Infection via agr-Dependent and agr-Independent Pathways. Infect Immun. 2016 Apr;84(4):1214–25.

229. Weidenmaier C, Peschel A, Xiong YQ, Kristian SA, Dietz K, Yeaman MR, et al. Lack of wall teichoic acids in Staphylococcus aureus leads to reduced interactions with endothelial cells and to attenuated virulence in a rabbit model of endocarditis. J Infect Dis. 2005 May 15;191(10):1771–7.

230. Parisi A, Le Thi Phuong T, Mather AE, Jombart T, Thanh Tuyen H, Phu Huong Lan N, et al. Differential antimicrobial susceptibility profiles between symptomatic and asymptomatic non-typhoidal Salmonella infections in Vietnamese children. Epidemiol Infect. 2020 May 26;148:e144.

231. Moller Abraham G., Winston Kyle, Ji Shiyu, Wang Junting, Hargita Davis Michelle N., Solís-Lemus Claudia R., et al. Genes Influencing Phage Host Range in Staphylococcus aureus on a Species-Wide Scale. mSphere. 6(1):e01263-20.

232. Jiang X, Hall AB, Xavier RJ, Alm EJ. Comprehensive analysis of chromosomal mobile genetic elements in the gut microbiome reveals phylum-level niche-adaptive gene pools. PloS One. 2019;14(12):e0223680.

233. Cury J, Oliveira PH, de la Cruz F, Rocha EPC. Host Range and Genetic Plasticity Explain the Coexistence of Integrative and Extrachromosomal Mobile Genetic Elements. Mol Biol Evol. 2018 Sep 1;35(9):2230–9.

234. Baker KS, Dallman TJ, Field N, Childs T, Mitchell H, Day M, et al. Horizontal antimicrobial resistance transfer drives epidemics of multiple Shigella species. Nat Commun. 2018 Apr 13;9(1):1462.

235. Gwenzi W, Chaukura N, Muisa-Zikali N, Teta C, Musvuugwa T, Rzymski P, et al. Insects, Rodents, and Pets as Reservoirs, Vectors, and Sentinels of Antimicrobial Resistance. Antibiot Basel Switz. 2021 Jan 12;10(1).

236. Huang G, Qu Q, Wang M, Huang M, Zhou W, Wei F. Global landscape of gut microbiome diversity and antibiotic resistomes across vertebrates. Sci Total Environ. 2022 Sep 10;838:156178.

237. Aivelo T, Laakkonen J, Jernvall J. Population-and individual-level dynamics of the intestinal microbiota of a small primate. Appl Environ Microbiol. 2016;82(12):3537–45.

238. Bloomfield S, Duong VT, Tuyen HT, Campbell JI, Thomson NR, Parkhill J, et al. Mobility of antimicrobial resistance across serovars and disease presentations in non-typhoidal Salmonella from animals and humans in Vietnam. Microb Genomics. 2022 May;8(5).

239. Bublitz DC, Wright PC, Rasambainarivo FT, Arrigo-Nelson SJ, Bodager JR, Gillespie TR. Pathogenic enterobacteria in lemurs associated with anthropogenic disturbance. Am J Primatol. 2015 Mar;77(3):330–7.

240. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, et al. Best practices for analysing microbiomes. Nat Rev Microbiol. 2018 Jul 1;16(7):410–22.

241. Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem Biophys Res Commun. 2016 Jan 22;469(4):967–77.

242. Naccache SN, Federman S, Veeraraghavan N, Zaharia M, Lee D, Samayoa E, et al. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome Res. 2014 Jul;24(7):1180–92.

243. Inda-Díaz JS, Lund D, Parras-Moltó M, Johnning A, Bengtsson-Palme J, Kristiansson E. Latent antibiotic resistance genes are abundant, diverse, and mobile in human, animal, and environmental microbiomes. Microbiome. 2023 Mar 8;11(1):44.

244. Zhang AN, Gaston JM, Dai CL, Zhao S, Poyet M, Groussin M, et al. An omics-based framework for assessing the health risk of antimicrobial resistance genes. Nat Commun. 2021 Aug 6;12(1):4765.

245. Bublitz DC, Wright PC, Bodager JR, Rasambainarivo FT, Bliska JB, Gillespie TR. Epidemiology of pathogenic enterobacteria in humans, livestock, and peridomestic rodents in rural Madagascar. PloS One. 2014;9(7):e101456.

246. Zohdy S, Grossman MK, Fried IR, Rasambainarivo FT, Wright PC, Gillespie TR. Diversity and Prevalence of Diarrhea-Associated Viruses in the Lemur Community and Associated

Human Population of Ranomafana National Park, Madagascar. Int J Primatol. 2015 Feb 1;36(1):143–53.

247. Ragazzo LJ, Zohdy S, Velonabison M, Herrera J, Wright PC, Gillespie TR. Entamoeba histolytica infection in wild lemurs associated with proximity to humans. Vet Parasitol. 2018 Jan 15;249:98–101.

248. Bodager JR, Parsons MB, Wright PC, Rasambainarivo F, Roellig D, Xiao L, et al. Complex epidemiology and zoonotic potential for Cryptosporidium suis in rural Madagascar. Vet Parasitol. 2015 Jan 15;207(1–2):140–3.

249. Andrews S, Krueger F, Segonds-Pichon A, Biggins L, Krueger C, Wingett S. FastQC. Babraham, UK; 2012.

250. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114–20.

251. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Apr 1;9(4):357–9.

252. Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, et al. Modernizing reference genome assemblies. PLoS Biol. 2011 Jul;9(7):e1001091.

253. Larsen PA, Harris RA, Liu Y, Murali SC, Campbell CR, Brown AD, et al. Hybrid de novo genome assembly and centromere characterization of the gray mouse lemur (Microcebus murinus). BMC Biol. 2017 Nov 16;15(1):110.

254. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021 Feb 1;10(2):giab008.

255. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017 May;27(5):824–34.

256. Blanco-Míguez A, Beghini F, Cumbo F, McIver LJ, Thompson KN, Zolfo M, et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. Nat Biotechnol [Internet]. 2023 Feb 23; Available from: https://doi.org/10.1038/s41587-023-01688-w

257. Clausen PTLC, Aarestrup FM, Lund O. Rapid and precise alignment of raw reads against redundant databases with KMA. BMC Bioinformatics. 2018 Aug 29;19(1):307.

258. Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep. 2021 Jun 16;11(1):12728.

259. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019 Nov 28;20(1):257.

260. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2013. Available from: http://www.R-project.org/

261. Oksanen J, Simpson GL, Blanchet FG, Kindt R, Legendre P, Minchin PR, et al. vegan: Community Ecology Package [Internet]. 2022. Available from: https://CRAN.R-project.org/package=vegan

262. Gloor GB, Macklaim JM, Fernandes AD. Displaying Variation in Large Datasets: Plotting a Visual Summary of Effect Sizes. J Comput Graph Stat. 2016 Jul 2;25(3):971–9.

263. Hsieh TC, Ma KH, Chao A. iNEXT: Interpolation and Extrapolation for Species Diversity [Internet]. 2022. Available from: http://chao.stat.nthu.edu.tw/wordpress/software_download/

264. Nixon MP, Letourneau J, David LA, Lazar NA, Mukherjee S, Silverman JD. Scale Reliant Inference. 2023.

265. Wilkins D. gggenes: Draw Gene Arrow Maps in “ggplot2” [Internet]. 2020. Available from: https://CRAN.R-project.org/package=gggenes

266. Amato KR, Mallott EK, McDonald D, Dominy NJ, Goldberg T, Lambert JE, et al. Convergence of human and Old World monkey gut microbiomes demonstrates the importance of human ecology over phylogeny. Genome Biol. 2019 Oct 8;20(1):201.

267. World Health Organization. WHO guidelines for plague management: revised recommendations for the use of rapid diagnostic tests, fluoroquinolones for case management and personal protective equipment for prevention of post-mortem transmission [Internet]. Geneva: World Health Organization; 2021. Available from: https://iris.who.int/handle/10665/341505

268. Nguyen VK, Parra-Rojas C, Hernandez-Vargas EA. The 2017 plague outbreak in Madagascar: Data descriptions and epidemic modelling. Epidemics. 2018 Dec;25:20–5.

269. Ojdana D, Sieńko A, Sacha P, Majewski P, Wieczorek P, Wieczorek A, et al. Genetic basis of enzymatic resistance of E. coli to aminoglycosides. Adv Med Sci. 2018 Mar;63(1):9–13.

270. Ikhimiukor OO, Oaikhena AO, Afolayan AO, Fadeyi A, Kehinde A, Ogunleye VO, et al. Genomic characterization of invasive typhoidal and non-typhoidal Salmonella in southwestern Nigeria. PLoS Negl Trop Dis. 2022 Aug;16(8):e0010716.

271. Binta B, Patel M. Detection of cfxA2, cfxA3, and cfxA6 genes in beta-lactamase producing oral anaerobes. J Appl Oral Sci Rev FOB. 2016 Apr;24(2):142–7.

272. Shi YZ, Yoshida T, Fujiwara A, Nishiki I. Characterization of lsa(D), a Novel Gene Responsible for Resistance to Lincosamides, Streptogramins A, and Pleuromutilins in Fish Pathogenic Lactococcus garvieae Serotype II. Microb Drug Resist Larchmt N. 2021 Mar;27(3):301–10.

273. Osei Sekyere John, Reta Melese Abate. Genomic and Resistance Epidemiology of Gram-Negative Bacteria in Africa: a Systematic Review and Phylogenomic Analyses from a One Health Perspective. mSystems. 2020 Nov 24;5(6):10.1128/msystems.00897-20.

274. Deng Y, Bao X, Ji L, Chen L, Liu J, Miao J, et al. Resistance integrons: class 1, 2 and 3 integrons. Ann Clin Microbiol Antimicrob. 2015 Oct 20;14:45.

275. Rakotonirina HC, Garin B, Randrianirina F, Richard V, Talarmin A, Arlet G. Molecular characterization of multidrug-resistant extended-spectrum β-lactamase-producing Enterobacteriaceae isolated in Antananarivo, Madagascar. BMC Microbiol. 2013 Apr 17;13(1):85.

276. Chaturvedi P, Singh A, Chowdhary P, Pandey A, Gupta P. Occurrence of emerging sulfonamide resistance (sul1 and sul2) associated with mobile integrons-integrase (intI1 and intI2) in riverine systems. Sci Total Environ. 2021 Jan 10;751:142217.

277. Adelowo OO, Helbig T, Knecht C, Reincke F, Mäusezahl I, Müller JA. High abundances of class 1 integrase and sulfonamide resistance genes, and characterisation of class 1 integron gene cassettes in four urban wetlands in Nigeria. PloS One. 2018;13(11):e0208269.

278. van Essen-Zandbergen A, Smith H, Veldman K, Mevius D. In vivo transfer of an incFIB plasmid harbouring a class 1 integron with gene cassettes dfrA1-aadA1. Vet Microbiol. 2009 Jun 12;137(3–4):402–7.

279. Raghuram V, Gunoskey JJ, Hofstetter KS, Jacko NF, Shumaker MJ, Hu YJ, et al. Comparison of genomic diversity between single and pooled Staphylococcus aureus colonies isolated from human colonization cultures. Microb Genomics. 2023 Nov;9(11).

280. Jacko AM, Durst AL, Niemchick KL, Modell SM, Ponte AH. Public Health Genetics: Surveying Preparedness for the Next Generation of Public Health Professionals. Genes. 2023;14(2).

281. James G, Witten D, Hastie T, Tibshirani R, others. An introduction to statistical learning. Vol. 112. Springer; 2013.