Influence of Environmental Gradients on Genomic Variation in Pediatric Commensals and Pathogens

Sanjam Sawhney
Influence of Environmental Gradients on Genomic Variation in Pediatric Commensals and Pathogens
by
Sanjam Sawhney

A dissertation presented to
Washington University in St. Louis
in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

St. Louis, Missouri
August 2023
# Table of Contents

**LIST OF FIGURES** .......................................................................................................................... IV
**LIST OF TABLES** ............................................................................................................................ VII
**ACKNOWLEDGMENTS** ..................................................................................................................... VIII
**ABSTRACT OF THE DISSERTATION** ............................................................................................... XV

**CHAPTER 1: INTRODUCTION** ......................................................................................................... 1
  1.1 Microbes as foes ............................................................................................................................. 1
  1.2 Microbes as friends ........................................................................................................................ 2
  1.3 Microbes, as microbes, and how we study them ........................................................................... 3
  1.4 References ..................................................................................................................................... 6

**CHAPTER 2: WEANING ACCELERATES AND TRANSFORMS WITHIN-HOST EVOLUTION IN THE INFANT GUT MICROBIOME** ......................................................................................... 9
  2.1 Abstract ......................................................................................................................................... 10
  2.2 Introduction .................................................................................................................................... 11
  2.3 Results .......................................................................................................................................... 13
  2.4 Discussion ..................................................................................................................................... 22
  2.5 Materials & Methods .................................................................................................................... 26
  2.6 Data availability ............................................................................................................................ 36
  2.7 References ..................................................................................................................................... 37
  2.8 Figures ......................................................................................................................................... 43

**CHAPTER 3: COMPARATIVE GENOMICS OF BORDERLINE OXACILLIN-RESISTANT**
**STAPHYLOCOCCUS AUREUS** **(BORA)** **DETECTED DURING** **A** **PSEUDO-OUTBREAK** **OF**
**MRSA** **IN** **A** **NEONATAL** **INTENSIVE** **CARE** **UNIT** ................................................................................... 59
  3.1 Abstract ......................................................................................................................................... 60
  3.2 Importance ..................................................................................................................................... 61
  3.3 Introduction ................................................................................................................................... 62
  3.4 Results .......................................................................................................................................... 64
  3.5 Discussion ..................................................................................................................................... 74
  3.6 Materials & Methods .................................................................................................................... 79
  3.7 Data Availability ........................................................................................................................... 86
  3.8 References ..................................................................................................................................... 87
  3.9 Figures ......................................................................................................................................... 94
  3.10 Tables ...................................................................................................................................... 108
CHAPTER 4: COMPARATIVE GENOMICS OF DIAGNOSTIC AND COMMENSAL STAPHYLOCOCCUS PSEUDINTERMEDIUS REVEALS NICHE ADAPTATION THROUGH PARALLEL SELECTION OF DEFENSE MECHANISMS ................................................................. 124

4.1 ABSTRACT ..................................................................................................................................... 125
4.2 INTRODUCTION ............................................................................................................................. 126
4.3 RESULTS ......................................................................................................................................... 129
4.4 DISCUSSION................................................................................................................................. 140
4.5 MATERIALS & METHODS ........................................................................................................... 145
4.6 DATA AVAILABILITY ................................................................................................................... 155
4.7 REFERENCES ............................................................................................................................... 156
4.8 FIGURES ......................................................................................................................................... 164

CHAPTER 5: CONCLUSION AND OUTLOOK ...................................................................................... 180

ADDENDUM: THE RESISTANCE WITHIN – ANTIBIOTIC DISRUPTION OF THE GUT MICROBIOME AND RESISTOME DYNAMICS IN INFANCY ................................................................................. 182

A.1 SUMMARY ..................................................................................................................................... 183
A.2 INTRODUCTION ............................................................................................................................. 184
A.3 ANTIBIOTIC EXPOSURE AND HEALTH OUTCOMES IN EARLY CHILDHOOD ......................... 185
A.6 THE RESISTOME IN EARLY LIFE .................................................................................................. 191
A.7 MULTIDRUG RESISTANT ORGANISMS IN THE INFANT GUT .................................................... 194
A.8 MOBILIZATION OF RESISTANCE IN THE INFANT GUT ............................................................... 195
A.9 FUTURE DIRECTIONS .................................................................................................................... 197
A.10 REFERENCES ................................................................................................................................ 200
A.11 FIGURES ...................................................................................................................................... 207

CURRICULUM VITAE ............................................................................................................................ 209
List of Figures

Figure 2.1  3,995 Reconstructed Genomes are recovered from 10 twin pairs and their mothers................................. 43
Figure 2.2  Evaluating MAG assembly approaches................................................................. 44
Figure 2.3  Putative and Reconstructed Genome capture by time................................. 45
Figure 2.4  Putative and Reconstructed Genome representation by assembly source and sequencing type................................. 46
Figure 2.5  Strain persistence, diversity, and co-occurrence from birth through early adolescence............................................. 47
Figure 2.6  Summary statistics for Persisting and Shared RGs............................................. 49
Figure 2.7  Intra-family strain sharing events vary by phylogeny.................................... 51
Figure 2.8  Taxa-specific mutation rates within the infant gut, colored by individual-of-origin.................................................... 52
Figure 2.9  Taxa-specific mutation rates within the infant gut, colored by appearance in relation to weaning................................................. 53
Figure 2.10  Taxa-specific mutation rates within the maternal gut, colored by individual-of-origin.................................................... 54
Figure 2.11  Generalized mutation rates of persisting RGs in mothers and infants........ 55
Figure 2.12  Mutation rates of persisting RGs by weaning status and pre-weaning diet..... 56
Figure 2.13  The weaning period is a mutation-generating hotspot that triggers a shift in mutated gene functions................................................. 57
Figure 2.14  No family effect on mutated gene profile distance between infants and mothers.................................................... 58
Figure 3.1  Experimental schematic................................................................................. 94
Figure 3.2  Comparison of growth on MRSA screening agars........................................ 95
Figure 3.3  Beta-lactamase inhibitor effect........................................................................... 96
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4</td>
<td>BORSA isolates do not cluster by core-genome, accessory-genome, or ARG content.</td>
</tr>
<tr>
<td>3.5</td>
<td>Class distribution by core genome lineage and sequence type (MLST).</td>
</tr>
<tr>
<td>3.6</td>
<td>Ancestral character estimation of borderline oxacillin resistance.</td>
</tr>
<tr>
<td>3.7</td>
<td>BORSA mutations by protein.</td>
</tr>
<tr>
<td>3.8</td>
<td>SNP distance of closely related isolates.</td>
</tr>
<tr>
<td>3.9</td>
<td>Unrefined RFC built with inclusion of the Argudín et al. (2018) dataset.</td>
</tr>
<tr>
<td>3.10</td>
<td>Random forest classifier predicts BORSA phenotype by PBP/GdpP mutational profile with high diagnostic accuracy.</td>
</tr>
<tr>
<td>3.11</td>
<td>Phylogenetically-informed gene-wide mutation association analysis.</td>
</tr>
<tr>
<td>4.1</td>
<td>Workflow schematic.</td>
</tr>
<tr>
<td>4.2</td>
<td>WGS identifies 493 isolates as <em>S. pseudintermedius</em> and 8 as <em>S. delphini</em>.</td>
</tr>
<tr>
<td>4.3</td>
<td>Niche adaptation is not apparent via whole genome architecture.</td>
</tr>
<tr>
<td>4.4</td>
<td>Host-species adaptation is not apparent via whole genome architecture.</td>
</tr>
<tr>
<td>4.5</td>
<td>Relationship between ANI and geographic distance.</td>
</tr>
<tr>
<td>4.6</td>
<td>Diagnostic isolates are distinguished by greater ARG burden and phenotypic resistance.</td>
</tr>
<tr>
<td>4.7</td>
<td>Determination of lineage cluster boundaries.</td>
</tr>
<tr>
<td>4.8</td>
<td>Syntenic gene structures continue to distinguish select diagnostic isolates following strain correction.</td>
</tr>
<tr>
<td>4.9</td>
<td>Household isolates share more spacers with isolates within their niche than with diagnostic isolates, agnostic of pairwise ANI.</td>
</tr>
<tr>
<td>4.10</td>
<td>Persisting strains accrue mutations in defense mechanism genes amidst household MRSA decolonization.</td>
</tr>
<tr>
<td>4.11</td>
<td>COG distributions of genes accruing at least one SNS.</td>
</tr>
</tbody>
</table>
Figure A.1  Indications, acute and long-term effects of early-life antibiotic exposures on gut microbiome trajectory and health outcomes in term and preterm neonates. ................................. 207

Figure A.2  Resistome dynamics in infancy. ................................. 208
List of Tables

Table 3.1  Phenotypic and genotypic characterization of index isolates, including phenotypic susceptibility testing and evaluation of chromogenic culture medium ................................................................. 108

Table 3.2  Compiled data on isolates under investigation and comparator blood isolates ................................................................. 110

Table 3.3  Phenotypic and genotypic characterization of comparator blood isolates, including phenotypic susceptibility testing and evaluation of chromogenic culture medium ................................................ 114

Table 3.4  Sensitivity and specificity of commercially available MRSA screening agars ........................................................................... 116

Table 3.5  Assembly statistics for each sequenced isolate ................................................. 117

Table 3.6  List of amino acid (AA) substitutions by BORSA isolate ...................... 121

Table 3.7  Bivariate association of the top Random Forest Classifier features with BORSA and MSSA status .............................................. 123
Acknowledgments

Nothing in life is preordained – my journey through graduate school among them. I would absolutely not be at this stage without an army of supporters around me. My deepest gratitude:

To my first scientific mentors at UC Davis, Dr. Karen Bales, PhD, Dr. Nicole Baumgarth, DVM, PhD, and Dr. Hannah P. Savage, DVM, PhD. Karen, thank you for taking me on as a young undergrad with no experience. Your research program gave me an excuse to play with (ahem, “observe and collect urine from”) titi monkeys, and helped me learn the valuable lesson that I am not a morning person! Nicole, your lab was the first place I learned basic molecular biology. I am so lucky to have been assigned under your then-graduate student, Hannah, who invested generous time and patience in my training. Hannah, thank you for allowing me to contribute to your 2017 Journal of Experimental Medicine paper. You gave me something to talk about during graduate school interviews, which is no small reason why I am where I am today.

To my thesis advisor, Dr. Gautam Dantas, PhD. Gautam, never in my wildest dreams could I have imagined having someone like you as a research advisor. I’ll never forget meeting you in the hallway outside your office on revisit weekend and seeing your red-rimmed spectacles perfectly matching a bright red sweater under your suit. I was in awe then with how well-dressed and well-spoken you were (and was convinced you had a different pair of eyewear for each outfit!). Over the years, I continued to be amazed by your ability to deeply
follow the individual and shared research projects of your roughly two dozen trainees and post-docs. In such a big lab, I never felt lost at sea. Most importantly, however, were the lessons I’ve learned from you about leadership. You are a leader’s leader, someone whose deliberate decision-making comes off as effortless. You are quick to identify strengths and weaknesses, and create opportunities that play to both. You’ve created a lab environment where people are able to do great science, and have fun. Above all, you are fiercely protective of your family time. Through my career, I hope to never lose sight of what’s truly important.

To my thesis committee, Dr. Carey-Ann Burnham, PhD, D(ABMM), FIDSA, F(AAM), Dr. Megan T. Baldridge, MD, PhD, Dr. Phillip I. Tarr, MD, and Dr. Daniel Goldberg, MD, PhD. Carey-Ann, your passion for clinical microbiology is infectious. I left every interaction with you excited to do more and learn more. I’ve always considered you a second mentor, and cannot wait to see what you accomplish at Pattern. Megan, I’ll never forget our 1-on-1 where we met to discuss a recent thesis update. I was dejected over some collaborator errors I discovered in the interim period that threatened the trajectory of my research project. There was a bounty of scientific advice you could have given me, but instead you made me feel heard and supported. My problems remained, but you provided exactly what I needed in that moment. Phil, you and Dr. Barbara Warner, MD have my deepest gratitude for your generous access to your meticulously curated clinical samples, without which Chapter 2 would be impossible. Dan, you are the steady heart of my committee. As my Chair, your kindness and encouragement always left me energized after thesis updates, and your commitment to supporting me through my graduation timeline is immensely appreciated.
To the support staff in the Edison Center for Genome Sciences and Systems Biology. Jessica Hoisington-Lopez and MariaLynn Crosby, thank you for all your sequencing assistance over the years. Your turnaround times have spoiled us, and I am truly afraid to return to a world where I’ll have to wait weeks for my As, Cs, Gs, and Ts. Brian Koebbe and Eric Martin, my research would not be possible without your maintenance of the High Throughput Computational Facility. Your generosity in always approving my “temporary” requests for disk space increases was hugely appreciated, as was your patience in educating me on some of the most basic principles of computing – I’ll never confuse memory and disk space again! And Bonnie Dee, it’s always a pleasure to chat with you in the halls. I’m happy to have assisted with your own research investigations!

To the “Thänert Lab,” Dr. Robert Thänert, Dr rer nat, Dr. Eric Keen, PhD, and Bejan Mahmud. Robert, you’ve been my “boots on the ground” mentor throughout my time in the Dantas Lab. As a fellow headstrong individual, I hope you’ve enjoyed sparring with me as much as I have with you. Through our spirited “discussions,” you have expanded my worldview scientifically, sociologically, and politically. Eric, you were one of my first friends in the Dantas Lab. From epic table tennis and real tennis battles, to the thrills of stock tracking and trips to East St. Louis, I’m lucky to have spent so much time with you. I hope we continue sharing our Wikipedia binges with each other ad infinitum. Bejan, the last man standing of our group. Your genuine excitement for life itself and all things in it is truly inspiring. I hope you never stop learning, and cannot wait to see what you accomplish in the future. And you have my begrudging thanks for the very, very, very few occasions where your Python scripts have offered a superior solution to my Excel wizardry.
To all my friends in the Dantas Lab, thank you. **Olivia Gorushi**, you are a true friend, someone who is always there to share in my wins and commiserate in my struggles. I’ll always appreciate our trips to the Children’s Hospital cafeteria, and hope there is a couch with my name on it when I visit you in the land north of North Dakota. **Kevin Blake**, one of my favorite moments of each day in lab involved “bickering” with you about the most inane things. Even though your enjoyment for spoiling Star Wars and Marvel content is questionable, thanks for having my back every day in lab. **Dr. Skye Fishbein, PhD**, you are equal parts brilliant and chaotic – a true mad scientist! You keep me on my toes every time I see you. **Dr. Drew Schwartz, MD, PhD**, our conversations span the full gambit from wholly immature to professional medical counsel. Thanks for injecting me with performance-enhancing drugs, and telling me stories about your friend “Sameer” in “Kenya.” **Wes Agee**, I wish you joined the lab years ago! Your ability to do grad school and be a father to two wonderful children is absolutely inspiring. **Dr. Tayte Campbell, PhD**, I’ll never forget shotgunning root beers with you at Float trip and over an autoclave bin in the Dantas lab hallway... **Bin Wang and Jie Ning**, the lab would not be functional without your work. **James Liao**, you are one of the most unique people I have ever met – you remind me to unabashedly be me. **Dr. Aura Ferreiro**, my fellow Aggie who is always down for a good time. **Dr. Robert Potter, PhD**, my rotation mentor who first exposed me to computational research, and **Luke Diorio-Toth**, who heavily influenced my progression from novice to semi-competent Bash scripter. My undergrads **Mallika Tyagi** and **Sahil Athalye** for trusting me with their first forays into academic science. And a hearty thank you to all my present and former lab buddies, **Anna DeVeaux, Rhiannon Vargas, JooHee Choi, PhD, Erin Newcomer, Jian Ryou, Galen Wong, Rehan Mehta, Suryang Kwak, PhD, Jerome Prusa, PhD, and Kim Sukhum, PhD**. There’s no better place to be on a Friday afternoon than the Dantas lab breakroom.
To my DBBS cohort, you were my first friends at WashU. From board game nights to Mission Taco’s infamous 10pm happy hour to Kyle’s never-ending birthday month, we’ve had some truly wonderful memories together. Suhas Bobba, I’ve somehow known you for longer than I’ve been in St. Louis! I’ve really enjoyed the growth of our friendship as we both progressed through our research. Thanks for always being down to hang on a Friday night, whether it’s to enjoy a cocktail (and watch me learn the essentiality of prosecco in a negroni sbagliato), or 1am shirtless karaoke in my living room (still feel bad for my neighbors who had just moved in). Keeley Choy, it was so heartening to have a friend to learn with and commiserate through the case prepping process – I would not have job offers in consulting without you. Ellie Gaylord, if you would have told me when we first met that you only draft Bills players in fantasy football, I never would have expected losing to you in our final season – well played. Jason Colasanti and Joe Lagas, some of my fondest memories (and lack of them) pre-COVID were at Jason’s Jackson Arms apartment. Whenever I hear a Two Friends “Big Bootie Mix,” or anyone rising on Easter, I think of our good times together.

To my roommates Jesús Santiago-Borges, Kyle Vuong, and Vernuli, we’ve stuck together for the last four years and become a family of our own. After long days in lab, there’s nothing better be greeted by one or all of you when I get home. Whether it’s binging Netflix, directing music videos, or having heart-to-heart conversations, thank you for all the great times together. I cannot wait to see the closing chapters of your research careers, and look forward to returning to St. Louis for your defenses. Jesús, I hope to visit you in Puerto Rico one day, and Kyle, it would be an honor to see Vietnam in the future. In the meantime, you both will always have a place to stay with me in the Bay.
To the city of St. Louis, thank you for educating me on what a “flyover state” is really like. In my time here I’ve learned what “seasons” actually are. I’ve lived through snowstorms (and had my first snowball fight!) and tornados (significantly less fun!). I’ve spent hours exploring different neighborhoods. I’ve accidentally run a half-marathon in the absolutely gorgeous Forest Park – no matter what anyone says, this is the best city park in the country. I bought a condo as a graduate student – where else in America can you say that! Sometimes I think I singlehandedly kept the late-hour St. Louis coffee shop scene alive (Thank you to Coma Coffee, Pipers Tea and Coffee, Quarrelsome Coffee, and several Starbucks for housing me as I wrote several manuscripts and this Thesis). It has been an honor to be a St. Louis citizen these last few years. The people here care about the city they call home. Not just as sports fans (though the Cardinals, Blues, and now St. Louis City fan bases are among the best I’ve seen), but as community members – in the home of the late Michael Brown, it was incredibly humbling to be here during the 2020 Black Lives Matter marches. Though my time in St. Louis may be coming to a close, the city will forever remain in my heart.

To my friends in California, thank you for not forgetting about me. Seriously. It’s easy to lose contact with people you no longer see every day. So if we are still talking, laughing, shooting the shit 2,000 miles away, I know it’s real. In the depths of the pandemic, I had the most fun with Alvin Lin, Andrew Westle, and Ben Mok ranking the entire discographies of Frank Ocean (from Not Crying to Fetal Crying), Drake (from Simp to Cocky), Kendrick Lamar (from Chill to Fire), and KanYe West (from Sinner to God and Crying to Cocky). And on any given night, there are decent odds I was gaming with Rohan Dhadwal, Subham Thapa, and
**Charan Bonda.** Thanks for keeping me sane all these years. I mean this when I say I absolutely cannot wait to see you all soon.

To my family. **Dad,** you instilled in me the importance of education at an early age, and gave me the means to pursue it debt-free. My future is better off for it. To **Vinny Mamu**, **Hina Mami,** and **Yuriko,** thank you for all the postcards and for feeding me generously when I come to visit. To **Alexis,** you’ve been my #1 supporter for all these years. We’ve been long distance for what feels like forever – thank you for sticking with me even after I made the crazy decision to go to grad school in the Midwest. If there’s any silver lining to be had from the pandemic, it’s that it brought you to live with us in St. Louis for all of 2020. Even in the darkest moments of society in our lifetimes, I had some of the happiest moments of my life living with you. Here’s to our future lives together. To my **Mom,** you are the reason I am here today. You were my first friend and earliest supporter, and are the personification of perseverance through tremendous adversity. Twenty years ago you applied to and received admission to several PhD programs, but were unable to attend because you needed to take care of your family. It is your love and faith that has driven me to earn this PhD – I dedicate this Thesis to you.

Sanjam Sawhney

*St. Louis, MO*

*August 2023*
Microbes interact with humans in a myriad of ways. They come together in the human gut to form complex communities that collectively transform and individually evolve in response to changing environmental conditions. Individual microbes can also cause human infection and escape antibiotic clearance through acquisition and development of novel resistance mechanisms. In this thesis, I chronicle the relationships between microbes and humans across two axes: from complex communities to single isolates, and from targeted studies of pediatric populations vulnerable to microbiome perturbations and pathogen challenge, to broader human and animal cohorts spanning colonizing and diagnostic environments.

In Chapter 2, I extensively profile individual microbial members of the gut microbiome from birth through early adolescence. This period of early life is essential to stereotypical gut microbiome development, but to date has mainly been characterized at the community level. To dive deeper, I employed deep short- and long-read sequencing of direct and culture-enriched stool to reconstruct thousands of high-quality microbial genomes. I identified hundreds of strain-sharing events within twin pairs, mother-infant dyads, or family triads, and determined Bacteroidales, Oscillospiraceae, and Lachnospiraceae to be vertically transferred from mother to infant, at birth and throughout infancy. I observed over 1,000 strains to persist within ever-changing microbial ecosystems, and catalogued species-level evolutionary rates in a high-throughput manner. Across infant dietary exposures, I identified weaning to be a critical window in which mutation accrual is accelerated among persisting
strains, as well as a key inflection point after which the gene families accruing mutations dramatically shift towards those of the maternal state. Taken together, this study offers an extensive catalogue of metagenome-assembled genomes in the infant gut, sharing patterns between cohabitating family members, and within-host evolution across dozens of species during periods of maximum microbiome perturbation.

In Chapter 3, I shift focus from the diverse microbial community in the pediatric gut to a single pathobiont colonizing the nares of NICU-hospitalized infants. Recent atypical trends observed during surveillance for methicillin-resistant *Staphylococcus aureus* (MRSA) at the St. Louis Children’s Hospital found oxacillin resistance among *S. aureus* isolates that did not carry the *mec* cassette, the canonical beta-lactam resistance-conferring operon. To investigate the genomic underpinnings of this resistance, I performed whole genome sequencing on 101 *S. aureus* isolates that spanned the borderline oxacillin resistance (BORSA), MRSA, and methicillin-susceptible (MSSA) phenotypes. This comparative genomics investigation found BORSA isolates to be phylogenetically diverse, with no common accessory genome that differentiated them from MRSA or MSSA comparators. However, through the creation of a random forest classifier I was able to predict BORSA identity based on beta-lactamase hyperproduction and five substitutions or truncations to key penicillin-binding proteins (PBPs) or the phosphodiesterase GdpP. This work provides evidence against a recent clonal expansion of BORSA and argues against the role of dozens of amino acid substitutions in PBPs previously suggested to contribute to the BORSA phenotype.

In Chapter 4, I direct attention to a close relative of *S. aureus*, *Staphylococcus pseudintermedius*. Study of *S. pseudintermedius* has historically been limited to its role as the dominant skin colonizer of and cause of pyoderma in canines, yet modern clinical diagnostics has revealed the species has long contributed to human morbidity – ironically, while being miscalled as *S. aureus*. Our expanded understanding beckons key questions regarding the ability of *S. pseudintermedius* as a species to adapt across environmental niches and host-types. Towards this, I sequenced over 500 *S. pseudintermedius* isolates captured in and on human and animal patients at the Barnes-Jewish Hospital and Kansas State Veterinary Clinic, as well as on human and pet inhabitants of and abiotic surfaces in households in the St. Louis metropolitan area. Although the whole genome architecture of *S. pseudintermedius* did not differ significantly by host-species, I found isolates originating from households to share
more CRISPR spacers – markers of ancestral phage predation – with each other than with diagnostic isolates. Further, I found diagnostic isolates to harbor extensive resistance gene burden, while isolates from households accrued non-synonymous mutations in defense-related genes amidst MRSA decolonization. Together, this study offers evidence in support of niche adaptation through parallel selection of defense mechanisms among diagnostic and household S. pseudintermedius.

In the Addendum, I discuss antibiotic-mediated microbiome disruption and development of the resistome during infancy. I summarize recent clinical trends of pediatric antibiotic prescriptions, and their associations with long-term pediatric health. I center recent work that interrogates earliest life seeding and development of antibiotic resistance in the gut microbiome, and specifically highlight mobilization and spread of antibiotic resistance between community members.
Chapter 1: Introduction

1.1 Microbes as foes

Over the course of human history – a notoriously small glimpse of time on this planet – the first predators known to Homo sapiens were large, intimidating creatures: sabertoothed cats, carnivorous reptiles, other hominids (1). As humans developed rudimentary spears, iron weapons, and gunpowder, the threat of large carnivores dissipated, and humans came to believe their truest danger to be other humans outside their “in group”. While aggression, defense, and warfare became integral components to the human experience, humans until very recently were not cognizant of the identity of their greatest killers: microbes. Yet Yersinia pestis, the bacteria responsible for the Black Death, killed up to 60% of western Eurasian population in the fourteenth-century (2, 3), potentially reducing world population by 25% (4). In the twentieth century, smallpox killed an estimated 300 million people, over 100 years after the first vaccine was developed for its prevention (5)! And even during World War I, considered to be one of the darkest and most destructive periods of human history, more humans died from influenza than from the violence wrought by man (6, 7).

It was not until 1884 that Robert Koch established strict criteria to identify bacteria as disease-causing agents – that the bacterium must be found in diseased individuals, recovered from them, able to cause disease in healthy individuals, and again recovered (8). Koch’s breakthrough in microbial culture, along with Louis Pasteur’s theory that microbes could
secrete products to kill other microbes ("antibiosis") (9), laid the foundation for an era of medical microbiology profoundly impactful on human health. Selman Waksman and other scientists in the mid-1900s would actualize Koch’s techniques and Pasteur’s theories to mine the microbial world for antibacterials, resulting in the creation of so many new drugs that the period would come to be known as The Golden Age of Antibiotics (10).

Unfortunately, the faltering rate of discovery of novel antibiotic classes since then, coupled with rampant clinical overuse of these drugs, brought upon a new scourge: antibiotic resistance. Bacteria acquire resistance through spontaneous – or de novo – mutations, or through the acquisition of antibiotic resistance genes. These resistance determinants can be rapidly disseminated through vertical transfer to offspring, selection against antibiotic-susceptible populations, or even horizontal transfer. Today, antimicrobial resistance is a global public health threat, directly attributable to over 1 million deaths each year (11, 12).

Yet, ...

1.2 Microbes as friends

...Microbes are everywhere: in the foods we eat, in the beverages we drink. In us. They cannot all be grave threats to humanity! Indeed, they make yogurt sour, make beer alcoholic, and make humans, well, human. As counterintuitive as it sounds, the human body is host to more microbial cells than human cells (13). They are found on our skin and in our mouth, but mostly they are found in our gut. There, they reside in tremendous numbers, forming complex communities composed of hundreds of species that we together call the gut “microbiome.” No two microbiomes are alike, not between people, near or far, or even in the same person on different days. These communities are constantly in flux, with individual members
vying for nutrients to metabolize, responding to perturbations they may or may not be well-suited to withstand, and dividing at exponential pace (14).

As eclectic as they are, these communities serve a purpose. Microbes within our gut help us ferment complex carbohydrates into simple sugars we can more easily digest. When we are young, they educate our immune system, helping us discern “friend” from “foe.” And, just by taking up space, they help exclude pathogens or pathobiont blooms – “bad” microbes that can cause morbidity and mortality (15). Their role is so important that the gut microbiome is sometimes considered a human organ in its own right (16). Like traditional organs, the gut microbiome has a highly choreographed development process. But unlike traditional organs, which are more or less formed during gestation (17), the human microbiome is non-existent until birth (18). A baby’s microbiome is first “seeded” by communities in their mother’s vaginal tract or on her skin, if a C-section is performed (19). From then on, a variety of factors, including prolonged hospitalization, early-life antibiotic exposure, milk type, and weaning, all shape the pace of microbiome maturation and the kinds of communities that develop (20-22). It is generally understood that the gut microbiome does not reach a fully mature state until three years of life (22). During this time, injury towards stereotypical microbiome development have been associated with disorders that develop in later life, including allergies, psoriasis, obesity, and diabetes (23).

1.3 Microbes, as microbes, and how we study them

So, what differentiates the “good” microbes from the “bad” microbes? It is not as simple as those within us being good – microbes that cause bloodstream or urinary tract infections can find refuge in the gut (24) – or those outside us being bad – we owe our existence
as a species to the prehistoric cyanobacteria that oxygenated the Earth (25)! Oftentimes, the
same microbial species is associated with both commensalism and pathogenesis. *Escherichia
coli*, for example, can be benignly carried in the gut, ingested as a probiotic (*E. coli* Nissle),
productively used in laboratory and industrial settings (*E. coli* K12), or cause severe diarrhea
(enterohemorrhagic *E. coli* O157:H7), depending on the strain under consideration (26, 27).

This range of physiologic capacity of and clinical outcomes engendered by a *single
species* is the result of “open pangenomes,” a concept foreign to *H. sapiens*. While humans
share over 98% of all possible gene families with every member of its species (“core ge-
nome”), many microbial species report enormous intraspecies variability due to a highly dy-
namic and interchangeable “accessory genome”, gene families that are found in fewer than
95% of members of that species. Microbial genes more frequently become nonfunctional or
gain new functions through greater reproduction rates and more error-prone DNA replica-
tion. But the greatest driver of open pangenomes among many bacterial species is their abil-
ity to participate in gene sharing with other microbes within and outside their species
through plasmid sharing, phage-mediated transfer, and uptake and integration of exogenous
DNA. As a result, the core genomes of microbial species may comprise anywhere from 84%
of its total pangenome to as little as 3%, as in the case of *E. coli* (28). Simply knowing the
species identity of a given microbe tells us little about what it is capable of.

Fortunately, microbial genomics has progressed at breathtaking pace in the last
twenty years. Whole genome sequencing (WGS) of pure bacterial isolates empowers us to
perform rigorous comparative genomics studies that contrasts individual members of a mi-
crobial species to interrogate their range of encoded functions – not just what nutrients each
strain is best optimized to metabolize, but also what virulence genes they may harbor and
how they may respond to antibiotic challenge. Meanwhile, technical developments during
the course of my own dissertation work have enabled researchers to profile complex micro-
bial communities – like the gut microbiome – at levels of granularity never before achieved.
Seminal research establishing the microbiome field was limited to answering questions like
“Which species are there?” and “What microbial genes are present?” (29). Now, we can reas-
semble dozens of microbial genomes from stool to rigorously tie gene content to bacterial
host and answer more consequential questions of “Which strains persist?” and “How do bac-
teria evolve within dynamic communities under constant perturbation?” (30, 31).

These are the questions I strive to answer in this thesis. In Chapter 2, I reconstruct
thousands of whole bacterial genomes from the guts of longitudinally-sampled infants and
their mothers to track strain persistence, sharing, and within-host evolution in the context
of infant diet and weaning. In Chapter 3, I interrogate Staphylococcus aureus isolates cap-
tured from pediatric patients in the St. Louis Children’s hospital to identify genomic corre-
lates of non-traditional borderline oxacillin resistance. In Chapter 4, I conduct an extensive
comparative genomics investigation among Staphylococcus pseudintermedius isolates col-
lected from diagnostic and household environments across humans, animals, and surfaces
to profile niche adaptation across environmental exposures and host-species. And in the Ad-
dendum, I review development of the gut resistome in infancy across birth modes, gesta-
tional age, and antibiotic exposures. Collectively, this body of work represents the primary
research publications from my Ph.D., and explores the applications of computational micro-
biology on our understanding of microbial evolution, genomics, and infectious disease.
1.4 References


Chapter 1: Introduction


Chapter 2: Weaning accelerates and transforms within-host evolution in the infant gut microbiome

This work was done in collaboration with Thänert, R., Thänert, A., Hall-Moore, C., Ndao, M., Mahmud, B., Warner, B. B., Tarr, P. I., & Dantas, G. A manuscript is under preparation for peer-review.
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

2.1 Abstract

The human gut microbiome is most dynamic in the first few years of life. In this time, sweeping changes to community architecture are understood, but it remains unknown to what extent individual strains persist and how they evolve in respect to traditional milestones of microbiome development and host aging. Here, we conduct deep hybrid and culture-enhanced sequencing on densely sampled twin pairs from birth through 8 years of life to define persistence and evolutionary potential of species in infancy through adolescence. We construct nearly 4,000 strain-resolved metagenome-assembled genomes across 399 taxa, and report 27.4% of these Reconstructed Genomes persist within individuals. We further identify 726 strains to be shared between twin pairs, mother-infant dyads, or family triads, with Bacteroidales, Oscillospiraceae, and Lachnospiraceae, but not Bifidobacteriaceae, vertically transferred. Lastly, we identify weaning as a critical inflection point that accelerates mutation rates in persisting strains and separates functional profiles of genes accruing mutations.
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

2.2 Introduction

The human gut microbiome is essential to human health, playing critical roles in immune priming, nutrient catabolism, and pathogen exclusion (1, 2). Assembly begins with vertical transmission at birth followed by rapid accumulation of microbes from cohabitating and environmental sources (3, 4). The first three years of life are critical to stereotypical gut microbiome maturation, beginning at developmental stage (months 3-14) before moving through transitional stage (months 15-30) at weaning towards a relatively stable, adult-like ecosystem comprised of trillions of cells (≥ 31 months) (5, 6). Each stage is characterized by a representative mix of bacterial species, and transitions through stages are heavily influenced by gestational age at birth, delivery method, diet, and exposure to antibiotics (7-10), with perturbations during this period linked to chronic diseases in later life (11-13).

Though community composition during this time is well-reported, there is a need to go beyond the species level. Intraspecies variation contributes to conflicting reports associating commensal inhabitants like Prevotella copri with both health and disease states (14), and recent studies indicate that subspecies-level dynamics do not correlate with species-level changes in abundance (15, 16). In adults, individual strains can stably colonize the gut for years to decades (17, 18), acquiring de novo mutations during this time, some of which become fixed through conferral of a fitness advantage or co-selection. Adaptive evolution has been reported to affect key functions, including polysaccharide utilization, antibiotic resistance, and intraspecies competition (18, 19), that ultimately improve individual survival within the gut microbiome. Recent reports indicate that bacterial strains may also persist in the far more dynamic infant gut microbiome (10, 20). However, it remains unknown how
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

mutation rates, along with the genes and pathways they occur in, vary across species and conform to the demands and perturbations of the intestinal community in constant flux.

Gut microbiome studies have traditionally relied on 16S rRNA amplicon sequencing, which can profile community composition at or above the species level, or shotgun metagenomics, which sequences all genetic content to ascertain relative abundances of microbial species and gene functions within the gut metagenome. These studies have been instrumental for the broad characterization of species’ succession in the maturing infant gut (5, 7, 21), and to differentiate between “healthy” and “disease” microbiome compositions associated with various pathologies (22, 23). However, intraspecies variation remains hidden through these techniques, a significant limitation given the wide range of physiologic and pathogenic potential of different strains within a species (24). Isolate sequencing is an accurate approach to achieving whole-genome resolution but is low-throughput and biased towards easily culturable microbes (25), rendering it infeasible for extensive characterization of complex microbial communities. Strain-resolved metagenomics has gained traction as a high-throughput, culture-independent solution for strain tacking within the gut microbiomes of adults and infants. However, such studies have defined “identical” strains based on single nucleotide polymorphisms (SNPs) solely in marker genes (26-28), a limiting approach that does not consider whole-genome variation and may confound similar, unrelated lineages (29). The construction of metagenome-assembled genomes (MAGs) from deep short-read sequencing conveys whole genome resolution for accurate strain-tracking within and across complex communities (30-33). Inclusion of long read and culture-enhanced sequencing provides additional benefits towards completing microbial genomes and the capture of low- (0.1 – 1%) and extra-low (< 1%) abundance microbes (34-36).
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

Here, we investigated how diverse selective pressures like infant diet and weaning shape the establishment, persistence, and evolution of the intestinal microbiota. Our study tracked evolution within the gut microbiome of 10 near-term twin pairs (20 infants) born at the same hospital from birth through 8 years of life (YOL). We also sampled mothers at birth through 3 years postpartum for baseline comparisons of mutation accumulation in stable communities and to assess strain sharing within families. We performed dual ultra-deep short- and long-read sequencing on direct stool and culture-enriched communities to construct microbial genomes of low- and high-abundance organisms in the infant and maternal guts. From these, we used whole genome comparisons to rigorously evaluate strain-level persistence and evolution in a high-throughput manner. Through this work, we identify species that frequently persist through the constantly evolving gut microbiome within infants, identify vertical transmission from mothers to infants at birth and in the years after, and determine weaning to be a significant contributor to the rates of mutation accrual and the genes they reside in.

2.3 Results

2.3.1 Recovering 16,392 microbial genomes from 20 longitudinally sampled twins and mothers.

In this study, we followed 10 twin pairs (20 infants) from birth through 8 YOL, and their mothers from delivery through 3 years post-delivery to thoroughly characterize strain persistence and within-host evolution. A total of 214 fecal samples from these infants and mothers were selected and subjected to ultra-deep short- and pooled long-read meta-
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

genomic sequencing, as well as culture-enhanced sequencing for targeted enrichment of obligate anaerobes (~2.8 tera base pairs [Tbp] of total sequence data). To assemble prokaryotic genomes, we employed a custom bioinformatics pipeline that performed dual short- and long-read metagenomic assembly and binning (Methods, Figure 2.1A) (37). This bifurcated approach, which benefited from the superior N50 scores of long-read origin scaffolds with greater total MAG recovery by the short-read assemblies, outperformed published tools applied to our dataset (Figure 2.2). Together, we assembled 16,392 Putative Genomes (12,597 from infants, 3,795 from mothers), 79.3% (n=12,993) of which exceeded our defined thresholds for medium-quality (MQ) MAGs (completeness ≥ 50%, contamination ≤ 10%) (Figure 2.1B). Infant yielded on average 56.4 Putative Genomes (PGs) per sample, though total count was highly correlated with time from birth to 2-3 YOL (range: 4 PGs [MOL 0] – 131 PGs [MOL 35]), reflective of community-level microbiome maturation. PG counts remained constant across longitudinal maternal samples (Figure 2.3A-D).

To account for strains recovered by both assembly approaches and for those that persist across multiple samples, Putative Genomes from each individual that surpassed medium-quality thresholds were dereplicated at 98% whole genome average nucleotide identity (gANI). 11.9% of these PGs were the sole representative of their strain; dereplication of the remainder resulted in 65.8% ± 7.3% decrease in total PG count per individual (Figure 2.3E-F) to achieve 3,995 representative MAGs, termed Reconstructed Genomes (RGs). Individuals averaged 148.8 ± 33.3 RGs/infant and 102.4 ± 17.7 RGs/mother. Median size of Reconstructed Genomes was 2.77 megabases (Mb) (interquartile range, IQR = 2.24 – 3.33 Mb). Most RGs originated from direct stool sequencing rather than culture-enriched communities, and short-read rather than long-read sequencing methods (Figure 2.4). All RGs were at
minimum medium quality, with 2,084 further fulfilling our high-quality (HQ) threshold (completeness ≥ 90%, contamination ≤ 5%, strain heterogeneity = 0%). Notably, the quality of our HQ Reconstructed Genomes was comparable or superior to that of isolate sequencing, with equivalent N50 scores ($q = 0.39$) and contamination levels ($q = 0.22$), along with fewer contigs per genome ($q < 0.0001$) and strain heterogeneity ($q < 0.0001$) relative to matched NCBI species-representative isolate assemblies (Figure 2.1C).

2.3.2 Reconstructed Genomes represent transient and persisting colonizers in the developing infant GM.

To characterize diversity across our Reconstructed Genomes, coverage breadth and pairwise identity was calculated between each RG and its closest hit within the NCBI RefSeq and Type Strain databases. Taxonomic annotation of the NCBI reference was adopted for the RG if pairwise identity of aligned regions was ≥ 95% over ≥ 50% of the query. For RGs without high-identity NCBI matches, taxonomy was inferred with the same parameters against the Genome Taxonomy Database (GTDB) (Figure 2.1A). Through this approach, we confirmed capture of 288 unique species across 2,606 RGs ($n = 1,411$ HQ, 1,195 MQ). The remaining 1,389 RGs ($n = 673$ HQ, 716 MQ) reflect unexplored microbial diversity as they represent species without publicly available isolate genomes (30). Maximum-likelihood placement of each of these RGs in the GTDB reference tree resulted in 89 unique genera annotations across 778 RGs and 22 unique family or class-level annotations across 611 RGs. Together, Reconstructed Genomes represent 399 taxa across 10 known phyla (Figure 2.5A).

A total of 2,974 (74.4%) Reconstructed Genomes representing 349 taxa originated in infants, reflecting our bias towards infant sample inclusion ($n = 176/214$ stool) (Figure
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

2.5B). Of the 208 taxa identified within both infant- and maternal-origin RGs, *Escherichia coli*, *Enterococcus faecalis*, and *Streptococcus sp.* were enriched among infant RGs (> 3 fold-change among taxa with > 0.5% RG representation), reflecting their role as early gut colonizers that decrease in abundance with age. Meanwhile, maternal-origin RG representation was greatest among bacterial families Lachnospiraceae and Oscillospiraceae, which traditionally achieve greatest relative abundance during adult-like microbiome states (Figure 2.5A).

Within-individual strain tracking was achieved via inStrain through comparisons of short reads between longitudinal samples mapped to same Reconstructed Genome, with “strain persistence” called if the compared region of the RG from both samples (minimum 50% coverage breadth) reported ≥ 99.999% population-level ANI (popANI) (20). Through this, we determined that 1,093 Reconstructed Genomes (n = 763 infant-origin, 330 maternal-origin) representing 224 taxa persisted across 2-21 samples within an individual (Figure 2.6A-B). The median residence time for persisters was 10 months (IQR = 4 – 43.5 months) in infants and 13 months (IQR = 7 – 35 months) in mothers. *Faecalibacterium prausnitzii* and *Bifidobacterium longum* were frequent persisters, with over 20 unique persisting RGs identified for each. We additionally observed several strains of Bacteroidales species (*Bacteroides faecis*, *Parabacteroides distasonis*, *Phocaeicola vulgatus*, and *Prevotella copri*) and Bifidobacteriaceae species (*Bifidobacterium bifidum*, *B. longum*, and *Bifidobacterium pseudocatenulatum*) to persist across 16-21 samples within our 4 extensively-sampled infants (Figure 2.5C-D).

Persistence rates trended with phylogeny, with strains of Bacteroidales (*Bacteroides fragilis*, *Bacteroides uniformis*, and *P. vulgatus*) and Bifidobacteriaceae (*B. pseudocatenulatum*, *B. bifidum*, and *Bifidobacterium breve*) more likely to persist within infants than strains
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

of other species (q = 0.039, 0.039, 0.039, 0.046, 0.017, 0.003, respectively; Binomial test). Meanwhile, strains of *Streptococcus salivarius* and *Ruminococcus/Mediterraneibacter torques*, both Lachnospiraceae, were more likely to be transient (q = 0.005, 0.027, respectively; Binomial test) (Figure 2.5C). Length of persistence similarly reflected relatedness, with Bifidobacteriaceae typically persisting through 3 YOL while Bacteroidales remained in most infant guts through the last sampling period at 7-8 YOL (Figure 2.5D). Infants were frequently co-colonized with multiple strains of *F. prausnitzii* and *B. pseudocatenulatum*, with both reporting ≥ 30% strain co-occurrence rate (Figure 2.5E). As a whole, the number of taxa in infant microbiomes reporting multi-strain co-colonization steadily increased with time from ≤ 1 in the first year of life (YOL) before stabilizing at 3 YOL, remarkably resembling well-reported microbiome maturation dynamics at the community level (Figure 2.5F) (5, 6).

**2.3.3 Strains are frequently shared between family members.**

We further dereplicated all individual-origin Reconstructed Genomes within each family unit to catalogue frequency and diversity of intra-family strain sharing over time. This generated a representative reference Reconstructed Genome list for each family, for which we could track strains across family members with inStrain as previously described. We observed sharing events to be widespread, occurring within every interpersonal relationship sampled – twin pairs, mother-infant dyads, and mother-offspring triads – within every family investigated. A total of 28.1% (726/2,586) of all family-level RGs were shared between at least two family members (Figure 2.6C, Figure 2.7A). Of these, we observe 226 RGs to be shared between mothers and their infants, reflecting prior literature reporting the maternal
microbiome to be a known source for strain seeding within the infant GM (20, 38). When such intergenerational strain sharing occurs, it appears to commonly be unrestricted to sole mother-infant dyads, as 70.8% (160/226) of shared RGs in our dataset appear in both infants. However, we find mothers to be secondary to an infant’s twin as the greater cohabitating contributor to infant GM diversity, with 500 RGs shared exclusively between twin pairs (Figure 2.7A). This is most frequent pre-weaning, with an incidence rate of 0.85 – 1.03 new strain sharing events/month in the first YOL that reduces to 0.52 events/month immediately post weaning, and 0.04 – 0.22 events/month during early adolescence (Figure 2.7B).

Sharing patterns further segregate by phylogeny and microbiome maturation periods. During the developmental phase of microbiota maturation, transient pathobionts E. coli and E. faecalis are exclusively or nearly exclusively shared between twin pairs, whereas commensal Bifidobacteriaceae present in maternal breastmilk are mostly shared between mothers and both infants (5, 7). Bacteroidales, which typically increase in abundance during the transitional stage (months 15-30) and are maintained during the stable stage (≥ 31 months), are predominantly shared between mothers and both infants. However, other common taxa during these later stages report mixed patterns, with some Oscillospiraceae (F. prausnitzii and Ruminococcus birculans) and Lachnospiraceae (Anaerostipes caccae, Anaerostipes hadrus, Blautia wexlerae, and Ruminococcus/Mediterraneibacter gnavus) mainly shared between twin pairs, while others (Oscillospiraceae Ruminococcus bromii; Lachnospiraceae Blautia massiliensis, Lacrimispora celerecrescens, Ruminococcus/Mediterraneibacter torques, and Ruminococcus/Mediterraneibacter faecis) are mainly shared between mother-infant dyads or family triads (Figure 2.7C). For each RG shared between a mother and at least one
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

Infant, we plotted its first appearance in each individual. We find that for Bacteroidales, Lachnospiraceae, and Oscillospiraceae, shared RGs appear in mothers before infants (q = 1.17e-08, 2.51e-05, 0.0025, respectively; FDR-corrected pairwise Wilcoxon), indicating mother-to-infant transmission. Interestingly, though the median first appearance of these shared taxa in mothers are at birth and 1-month postpartum for Bacteroidales and Lachnospiraceae, their median first appearance in infants at MOL 14 and 13, respectively, indicate these strains are not vertically transferred at birth or through breastmilk but instead are seeded through cohabitation. This contrasts with Bifidobacteriaceae, whose median first appearance in mothers and infants are 4 months postpartum and 7 MOL, for which directionality cannot be confidently determined (q = 0.174) (Figure 2.7D).

2.3.4 Weaning is the greatest driver of shifts in mutation rates and mutated genes within persisting RGs.

The wealth of persisting RGs presented a unique opportunity to profile within-host evolution in both the developing infant and stable postpartum maternal gut. Towards this, all discordant base calls with ≥ 10x coverage and ≥ 90% major allele frequency across timepoints were recorded for each persisting RG. Total mutation counts were scaled by coverage breadth to attain whole-genome equivalent counts. From this, we were able to estimate taxa-specific rates of evolution in a high-throughput manner. For 23 taxa in infants and 9 taxa in mothers, each with ≥ 5 persisting RGs, linear regression indicated moderate prediction potential (R² ≥ 0.5) of our calculated rates (Figures 2.8 – 2.10). Generalized mutation rates, calculated across 763 persisting RGs (194 taxa) in infants and 330 persisting RGs (130 taxa) in mothers, found mutations to accrue more frequently in the developing infant gut
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

microbiome (4.6 vs. 2.6 SNPs/year) (Figure 2.11A). Individually, RGs persisting in infants had greater mutation rates than those persisting in mothers ($p < 0.0001$, Mann-Whitney).

While mutations accrued stably in the maternal gut over the years post-birth, we observed a distinct biphasic relationship in evolutionary rates within the developing infant gut, in which RGs appeared to accrue mutations more rapidly during their first 9 months of persistence than the remainder of their residency (Figure 2.11B). We first assessed dietary exposure as a possible contributor, but found evolutionary rates to be comparable between persisting RGs in breastfed and formula-fed infants (Figure 2.12). However, when we instead segregated persisting RGs by initial colonization period, we observed the rapid mutation accumulation in early persistence to primarily associate with RGs that colonized prior to weaning, with those that colonized post-weaning harboring steady mutation rates akin to RGs in the maternal gut (Figure 2.13A). Consequently, RGs that colonized prior to weaning had greater rates of evolution than those that colonized post-weaning ($p = 0.0004$, Mann-Whitney) (Figure 2.13B). Given that weaning is the primary driver of the shift from the developmental to transitional stage of microbiome maturation, we hypothesized that persistence through this window would result in an elevated evolutionary rate, and that this may drive the greater rate observed in pre-weaning vs. post-weaning RGs. Towards this, we calculated two mutation rates for all pre-weaning persisting RGs, one from seeding to the final timepoint prior to weaning, and the other between the immediate timepoints flanking weaning. As hypothesized, we observed that evolutionary rates spiked during the weaning window relative to their time of initial colonization ($p < 0.0001$, Wilcoxon matched-pairs) (Figure 2.13C), implicating persistence through weaning as the greatest driver of mutation accumulation in the developing infant gut.
Finally, we ventured to assess whether weaning similarly served as an inflection point for gene functions accruing mutations. Towards this, we annotated all open reading frames (ORFs) that accrued at least one mutation, and binned these by occurrence relative to weaning and infant diet prior to weaning. The distribution of mutated ORFs across gene functional categories was determined by Eggnog for each bin. We plotted the distributions of each bin via principal coordinate analysis, and found weaning status to be the primary driver of variation (45.7% variance captured) (Figure 2.13D). The secondary axis (32.8% variance captured) differentiated pre-weaning mutated ORFs by infant diet, with “intermediate” infants more closely associated with formula-fed infants. While the distribution of pre-weaning mutated ORFs were widespread across dietary exposures, we observe tight clustering of those post-weaning, signifying the dissipating role of pre-weaning milk-type in the post-weaning infant gut. Lastly, we observed the distribution of post-weaning mutated ORFs to more closely resemble that of mothers. For the four extensively-sampled infants, we generated individual-specific pre-weaning and post-weaning distributions of mutated ORFs by gene function, and an overall distribution for each of their mothers. We then calculated pairwise Bray-Curtis distance between each infant and each mother, finding the post-weaning infant distributions to be a significantly better approximation for those of mothers relative to pre-weaning infant distributions ($p = 0.0002$, Mann-Whitney) (Figure 2.13E). Intra-family comparisons were not significantly closer than inter-family comparisons, indicating the impact of weaning on mutated gene functions outweighs cohabitation and mother-infant effects (Figure 2.14).
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

2.4 Discussion

We present a comprehensive, strain-resolved investigation of within-host evolution – and the factors that influence it – of persisting bacteria in the developing gut microbiome. Our cohort comprised 20 infants (10 twin pairs) densely sampled from birth through their first 8 years of life, and their mothers sampled from childbirth through 3 years postpartum. We employed ultra-deep short- and long-read sequencing on direct stool and culture-enriched communities, coupled with stringent whole-genome based definitions for intraspecies diversity, to characterize persistence and transience across hundreds of species that colonize the human gut during infancy and in adulthood, and to chronicle strain sharing between cohabitating family members from birth through early adolescence. For persisting strains, we track SNP accumulation rates within the infant and maternal gut within and across dozens of species, and characterize the influence of infant diet and weaning on the gene families accruing mutations.

Our approach unlocks whole-genome resolution of the gut microbiome, a level of granularity entirely inaccessible to prior analyses performed via 16S amplicon sequencing or shallow shotgun sequencing (24). Strain tracking within and between cohabitating family members has more recently been performed using metagenomic data (26-28), but their restriction to allelic variation of marker genes can obfuscate closely-related lineages and ignores underlying differences in accessory genome content that can dramatically impact physiologic and pathogenic capacity (14). While isolate sequencing presents a compelling alternative that provides whole-genome data (18), its reliance on culturability and low-throughput nature renders it prohibitive to scale (25). Instead, recovery of thousands of high-quality and phylogenetically-diverse genomes from the infant gut is most feasible via
deep sequencing and MAG assembly (33). Our work furthers progress made by the hallmark short-read MAG studies of the human gut microbiome (30-32) through our incorporation of long-read and culture-enriched sequencing, which together bolsters N50 scores and can even close entire bacterial genomes (36). Recent work using similar whole-genome-based MAG tracking has captured mother-to-infant vertical transmission of bacterial strains at birth and their persistence through the first year of life (20). Here, we extend their findings from birth through early adolescence, expand strain sharing to multiple cohabitating family members at birth and beyond, and catalogue within-host evolution of those persisting bacteria through environmental pressures like infant diet and key transition events like weaning.

We recovered 16,392 Putative Genomes from 214 infant and maternal stool, which we rigorously dereplicated to attain 3,995 strain-resolved Reconstructed Genomes. Our high-quality RGs (52.2% of total) were remarkably complete, reporting comparable or even superior genome quality statistics relative to pure isolate assemblies available on NCBI. RGs recover the vast range of microbial diversity that transiently or persistently colonize the human gut in early life and during adulthood, together representing 399 unique taxa across 10 diverse phyla. Further, within-species total RG count reflects known differences in relative bacterial abundance between infants and adults, as our observed enrichment in total *E. coli* and *E. faecalis* RGs in infants follows their reported high prevalence and abundance during the first year of life and general abundance below the limit of detection in healthy adults (5, 7, 21). We similarly identify congruence between the increasing number of species with cohabitating RGs and increasing richness and diversity in the infant gut, with both stabilizing between 2-3 years of life (3, 5).
Most strains in the infant gut appear to be transient colonizers, an expected observation given the rapid ecological turnover influenced by individualized seeding dynamics and dietary exposures that drive a constant restructuring of the gut microbiome. Still, we identify more than one-quarter of RGs to persist across at least two timepoints within an individual. A recent study investigating strains that seed within the first 2 months of life (“early colonizers”) reported a median residency of 9.6 months among persisters, and identified *B. uniformis* and *P. vulgatus* as more likely to persist than other species (20). Remarkably, we similarly observe a median persistence of 10 months in our expanded analysis encompassing strains that seed at any point in the first 3 years of life, and extend support of this duration of persistence from early-colonizers alone to all microbes that colonize during infant gut development. We also validate their findings for *B. uniformis* and *P. vulgatus*, and further identify *B. fragilis* as well as *Bifidobacterium* species *B. pseudocatenulatum*, *B. bifidum*, and *B. breve* as frequent persisters in the infant gut. Bacteroidales and *Bifidobacterium* strains, along with those from the Oscillospiraceae and Lachnospiraceae families, are also frequently shared between mothers and infants. For all but Bifidobacteriaceae, we determine these shared strains to first appear in mothers. Our inability to call sharing directionality for Bifidobacteriaceae is likely due to those species colonizing the infant gut much earlier in infancy, following positive selection from breastmilk rich in human milk oligosaccharides (39).

Whole-genome resolution of persisting strains enabled us to track within host-evolution across dozens of microbial inhabitants in the infant and maternal gut. Our findings can be contextualized against a recent study that calculated the mutation rate for a single taxon in the adult microbiome over 1-2 years across 7 persisting strains. Their rate of 0.9 SNPS//genome/year – determined via the gold-standard of extensive isolate sequencing – reported
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

an $R^2 = 0.63$ (18). Here, we report remarkably congruent correlation coefficients for dozens of taxa across similar strain counts over the same or often longer periods of persistence, demonstrating the value of whole-genome MAG comparisons for the high-throughput characterization of within-host evolution. When generalizing across taxa, we observe greater mutation rates in the infant gut than the maternal gut. This again is likely a reflection of the dynamism of a microbial ecosystem whose members face perturbations that are more frequent and of greater magnitude, which may enhance or remove preexisting fitness advantages. Of these perturbations, we identified weaning to be a major contributor, during which mutation rates spike as persisting microbes are forced to adapt to dramatically differing nutrition forms. These changes, including decreases in HMOs and increases in starch and complex sugars (21), also influence the genes that accrue mutations towards a distribution that more closely resemble mutation-accruing genes within the maternal microbiome, a by-product of both the strain turnover elicited by the cessation of liquid feeding and the new requirements for persistence in the post-weaned infant gut.

Limitations of our study include overrepresentation of a single geographic region and race among our study participants. With a relatively small participant pool, we are underpowered to fully assess potential impact of confounders on our findings. Our work would benefit from rigorous evaluations against future multi-center studies with more diverse racial and ethnic backgrounds across participants. Notably, for 35% of our 3,995 RGs, no species-level reference was identifiable in the NCBI or GTDB databases, emphasizing the extensive microbial diversity that remains unexplored in the human gut microenvironment (30); future work would be well-placed to exploit recent advances in culturomics (40) to generate reference genomes for these fastidious or previously “unculturable” microbes.
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

2.5 Materials & Methods

2.5.1 Study Cohort.

Study approval was obtained from the Human Research Protection Office approval of Washington University in St. Louis School of Medicine, and written informed consent was obtained from all adult participants for themselves and as parents or legal guardians of all minor subjects. Fecal samples from a cohort of 20 healthy near-term twins in St. Louis, MO, taken at 6 MOL, at weaning, 2 months prior to and post weaning, at 3 YOL, and at 7-8 YOL, were frozen at -80°C at collection and utilized for this study. Four of these 20 infants were extensively sampled, with 19 – 21 timepoints sequenced throughout early life (range: 0 – 97 MOL). Likewise, maternal stool taken at birth of infant, and 6, 24, and 36 months post-birth were collected and utilized for this study. Together, our sampling captured 214 samples from 20 infants and 10 mothers in 20 families, spanning birth through 8 years of life, and childbirth through 3 years postpartum. Demographic data are available in Table 2.1.

2.5.2 Illumina and ONT Sequencing.

Fecal metagenomic DNA was extracted via DNeasy PowerSoil Pro (Qiagen, Hilden, Germany) and prepped via a modified Nextera XT library preparation kit (Illumina, San Diego, CA, USA) as previously described (10, 41). Libraries were pooled at equal concentrations and sequenced on an Illumina NextSeq 500 HighOutput platform (Illumina, San Diego, CA, USA) to a target depth of 50 million reads per sample. For the targeted enrichment of obligate anaerobes *Bacteroides*, *Bifidobacterium*, and *Clostridium* spp., the same stool was plated on Blood agar with colistin agar, Wilkins-Chalgren agar, and Laked Blood with Kanamycin and Vancomycin (LKV) agar (Hardy Diagnostics, Santa Maria, CA, USA) (the latter with prior
chopped meat broth with colistin outgrowth) (10-15 mg/plate) and incubated anaerobically for 3 days at 37°C. All bacterial growth on plates were then swept with nuclease-free water and underwent DNA extraction, library preparation, and Illumina short-read sequencing as mentioned above. Each plate was sequenced to a target depth of 5 million reads. Finally, the same stool also underwent phenol-chloroform DNA extraction and BluePippin HighPass high molecular-weight size selection (Sage Science, Inc, Beverly, MA, USA). DNA from three consecutive stool samples per infant were pooled prior to Oxford Nanopore Technology (ONT) library preparation via Ligation Sequencing Kit SQK-LSK109 and Barcoding Expansion Kit EXP-NBD114, and sequenced on ONT R9.4.1 flow cells (Oxford Nanopore Technologies, Oxford, United Kingdom).

2.5.3 Generation, Annotation, and Quality Assessment of Reconstructed Genomes.

Timepoint-specific short-read-origin metagenomic scaffolds were created from Illumina reads using SPAdes (version 3.14.0; flags: --meta) (42, 43). Separately, long-read-origin metagenomic co-assemblies were generated from pooled ONT reads across three consecutive timepoints via flye (version 2.8.1; flags: --meta, --nano_raw, -i 3) (44). Timepoint-specific short reads were aligned against the pooled long read scaffolds using Bowtie2 (version 2.3.5) (45) for deconvolution of the co-assembly by timepoint. Long read scaffolds with \( \geq 85\% \) breadth of coverage by timepoint-specific short reads were binned into timepoint-filtered long read scaffold sets using bedtools (version 2.29.2; workflows: genomecov, merge, and getfasta) and seqtk (workflow: subseq) (46, 47). For the remaining long reads, any regions with 100% alignment were also extracted and added to the timepoint-filtered long read bins. In validation of our approach, 76.1\% \pm 9.1\% of timepoint-specific short reads aligned to
matched long read co-assemblies, while just 22.0% ± 9.9% of short reads aligned to long read co-assemblies from mismatched infants. Timepoint-filtered long read scaffolds were then polished by their matched timepoint-specific short reads using pilon (version 1.22) (48). For each timepoint, DAS Tool (version 1.1.2; flags: --search_engine diamond, -l concoct,max-bin,metabat) was implemented independently on timepoint-specific short-read origin and timepoint-filtered long-read origin metagenomic scaffolds for MAG binning, yielding 16,392 Putative Genomes (49).

Strain-level MAG dereplication was implemented on all Putative Genomes from the same individual via dRep at 98% gANI secondary threshold (version 3.2.2; workflow: dereplicate; flags: --S_algorithm gANI, -sa 0.98, -nc 0.3, -comp 50, -con 10) to attain a set of representative strain-resolved MAGs for each individual (“Reconstructed Genomes [RGs]”), and their prodigal gene predictions (50). Taxonomy of the Reconstructed Genomes was first approximated via Mash (version 2.3; workflow: screen; flags: -w) against the NCBI RefSeq and Type Strain databases (51). For each RG, pyANI (version 0.2.11; flags: -m ANIm, --nocompress) was run against the best Mash hit from each database (52). Passing metrics were set as ANI ≥ 95% and Hadamard (ANI * breadth of coverage) ≥ 0.5. If the best hit from both databases returned passing metrics, if the Type Strain reference reported greater Hadamard than the RefSeq reference, the Type Strain taxon was assigned to the RG. If the RefSeq reference reported greater Hadamard, if both references were ascribed the same genus and the RefSeq reference ended in “sp.”, the Type Strain taxon was ascribed (else the RefSeq taxon was ascribed). If neither database returned passing metrics, GTDB-TK (version: 1.7.0; workflow: classify) was employed to determine taxonomic classification (53). Total taxonomic di-
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

versity was portrayed through generation of a tree file using GTDB-TK (version: 1.7.0; workflow: de_novo_wf; flags: --skip_gtdb_refs) on a subset of RGs (one representative RG per taxon), and was visualized as a phylogenetic tree on ITOL (53, 54). Family and phyla classifications for each taxon were determined using the NCBI Taxonomy Browser and displayed as color strips, with total strain-resolved RGs per taxon displayed as radial bar graphs (55).

Assembly quality (completeness, contamination, strain heterogeneity, N50, and contig count) of all Putative Genomes as well as the post-dRep representative Reconstructed Genomes was determined via CheckM (version 1.0.13; workflow: lineage_wf) and Quast (version 4.5) (56, 57). Putative and Reconstructed Genomes were determined to be High Quality (completeness ≥ 90%, contamination ≤ 5%, strain heterogeneity = 0%), Medium Quality (completeness ≥ 50%, contamination ≤ 10%), or Low Quality (completeness ≤ 50% or contamination ≥ 10%; Putative Genomes only) (Table 2.2). Assembly quality was also determined for the best hit for each RG from the NCBI RefSeq (n=827) and Type Strain (n=223) databases as described. RGs were binned by Quality type and compared against the NCBI isolate references for each assembly metric. Significance for quality comparisons between groups was determined by Kruskal-Wallis test with Dunn’s multiple comparison correction using Prism 9 and comparisons visualized using the ggplot and stat_compare_means(test = “kruskal.test”) functions (R ggplot2 and ggpubr packages) (58, 59).

2.5.4 Strain Diversity, Persistence, and Co-occurrence.

Strain persistence and co-occurrence was determined via inStrain (29). Towards this, all Reconstructed Genomes, and all prodigal gene predictions for each RG, were concatenated
into two multifasta files per individual. A scaffold-to-bin (stb) mapping file was then generated via dRep (version 3.2.2; workflow: parse_stb.py; flags: --reverse), and TS-specific short reads were aligned against the reference RG multifasta for each timepoint within an individual via Bowtie2 (version 2.3.5). The resulting alignment bam file, along with the RG multifasta, the gene prediction multifasta, and the stb were provided as inputs for InStrain-profile (version: 1.5.4; workflow: profile; flags: -g, -s) to determine coverage breadth and depth of each RG at each timepoint. InStrain-compare was then run on all TS-InStrain-profile outputs within each individual (version: 1.6.3; workflow: compare; flags: --store_mismatch_locations, --database_mode, -bams) to make strain-level comparisons across timepoints. Reconstructed Genomes were determined to “persist” across two or more timepoints if the scaffolds at both timepoints overlap over ≥ 50% of the positions assigned to the RG at ≥ 5x coverage each (percent_genome_compared ≥ 0.5) and population ANI (popANI) between the TS-scaffolds ≥ 99.999% (Tables 2.3 – 2.4). RGs that did not meet these criteria were labeled Transient. Total RG count and the transient-persisting ratio was visualized in ggplot for the taxa with the greatest MAG diversity within infants. Persistence rates for each species with ≥ 15 total infant-origin RGs (≥ 0.5% representation among all infant-origin RGs) was compared against that of all other species (meeting inclusion criteria) using Binomial test with FDR correction (binom.test(alternative = “two.sided”) and p.adjust(method = “fdr”) (R STATS package); RGs not classified at the species level were excluded. A tree file plotting the representative reference genome from NCBI Genome for each taxon was created via GTDB-TK (version: 1.7.0; workflow: de_novo_wf; flags: --skip_gtdb_refs, --custom_taxonomy_file) and visualized in ITOL.
An RG was determined to be present at each timepoint if it met the aforementioned definition of persistence, if it reported \( \geq 5x \) coverage over \( \geq 50\% \) of the positions against the reference RG at high identity (inStrain-profile breadth_minCov \( \geq 0.5 \), popANI_reference \( \geq 99.999\% \)), and/or if it was the timepoint of origin for the dereplicated RG. Strain co-occurrence was then determined if a taxon was represented by two or more RGs present at a given timepoint. The number of taxa displaying strain co-occurrence over time was visualized as a scatterplot with trendline (geom_smooth(method="lm")) in ggplot. For each taxon with high MAG diversity in infants, a strain co-occurrence rate was calculated as the total timepoints in which multiple strain-resolved RGs of the taxon were present / the total timepoints in which any RG of that taxon was present.

### 2.5.5 Intra-Family Strain Sharing.

To catalogue strain sharing within families, strain-level dereplication now implemented on all Putative Genomes from the same family (rather than the same individual) as described above. TS-specific short reads of each family member were aligned against a family-level RG multifasta, and the resulting alignment bam files, the family-level RG multifasta, and a new family-level stb were provided as inputs for inStrain-profile all as previously described. To reduce memory requirements of inStrain-compare, the inStrain-profile genome_info.tsv output files were filtered to display only the samples of an infant(s) and/or mother that each RG could potentially be present in (\( \geq 99.0\% \) popANI_reference over \( \geq 0.5 \) coverage breadth). For each individual RG, inStrain-profile was now rerun with only the single RG as the reference genome (rather than the multifasta containing all family-level RGs) and only for samples it could potentially be present in. Of note, the original alignment file
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

against the family-level RG multifasta remained in use for these inStrain-profile runs to retain competitive mapping and reduce mismapped reads. InStrain-compare was then run for each RG on all RG-specific inStrain-profile outputs within each family unit at default parameters.

Intra-family strain sharing was defined as overlap ≥ 50% of the positions assigned to the RG at ≥ 5x coverage (percent_genome_compared ≥ 0.5) and popANI ≥ 99.999% across two or more timepoints across two or more family members (Table 2.5). Sharing relationships for the most frequently shared taxa were visualized in ggplot and sorted by phylogeny. The tree was created and visualized in GTDB-TK and ITOL using the NCBI Genome reference genome for each taxon as previously described. “New” strain sharing events, in which an RG appeared in both individuals at the same timepoint, were tracked within twin pairs, annotated by bacterial Class or Order (as determined by NCBI Taxonomy Browser), and visualized in ggplot. For RGs shared between mother-infant dyads or family triads (Table 2.6), density plots were created in ggplot to visualize their first appearance in mothers and in infants.

2.5.6 popSNP Tracking and Mutation Rates.

To identify open reading frames (ORFs) that accrue population SNPs (popSNPs) over time, the concatenated inStrain-compare pairwise_SNP_locations.tsv output file, which lists each discordant base call between two compared timepoints, was filtered to include only those for RGs that met the criteria for “persisting”, and filtered again to only include the timepoint comparisons the persisting RGs were present in. Positions were again filtered at ≥ 10x coverage threshold and ≥ 90% major allele frequency for high confidence in popSNP
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

calls (Table 2.7) (18). Remaining entries were converted from pairwise timepoint comparisons (i.e., MOL 9 -> 14, T -> C, MOL 9 -> 36, T -> C, MOL 11 -> 14, T -> C, MOL 11 -> 36 T -> C) to temporal nucleotide transitions (same example, MOL 11 -> 14 T -> C) via custom python scripts. Resulting popSNPs were linked to their persisting RG and binned by the length of persistence from seeding of the RG to when each occurred (Table 2.8). The number of new popSNPs of an RG at each unique interval for each unique length of persistence was recorded (i.e., RG1 persists from MOL 6 -> 36; MOL 9 -> 15 and MOL 12 -> 15 timepoint comparisons both describe popSNPs occurring at LoP = 9 months but are different compared intervals). Because the minimum required coverage overlap of an RG to make comparisons across two timepoints is 50%, popSNP count for each unique interval was divided by the percent_genome_compared value across the two timepoints (range: 0.5 - 1) to attain a scaled popSNP count across the entire genome (breadth-adjusted count per interval). Then, for each unique length of persistence, the breadth-adjusted count of each interval with the same end timepoint was summed to achieve the breadth-adjusted new popSNP count at each length of persistence (same example of RG1, new popSNPs at LoP = 9 months, MOL 9 -> 15 and MOL 12 -> 15). To consider all popSNPs from seeding up to each timepoint under investigation (and not just those that happened since the most recent prior timepoint), this value was summed with all breadth-adjusted counts for all prior lengths of persistence (Table 2.9). For each taxon represented by ≥ 5 persisting RGs, breadth-adjusted summed popSNP counts by length of persistence were visualized as scatterplots with linear regression trendline (geom_smooth(method=“lm”)) in ggplot. The slope of the regression, calculated via stat_regline_equation() (R ggpubr package), is the estimated evolutionary rate for each taxon.
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

To compare mutation rates across taxa, all summed breadth-adjusted popSNP counts for each RG were normalized by genome size (Table 2.9). Breadth- and length-adjusted summed popSNP counts by length of persistence was visualized as a scatterplot with trend-line (geom_smooth(method="loess")) in ggplot. Towards this, data point(s) represent each length of persistence of an RG in which there was an increase in aggregate popSNP count. Data points were also included for each recorded length of persistence prior to the first SNP accrual (aggregate popSNP count = 0). Finally, some persisting RGs did not accrue any pop-SNPs throughout their persistence. For these, all unique lengths of persistence that met inclusion criteria (99.999% popANI over ≥ 0.5 breadth) were plotted (aggregate popSNP count = 0). While each persisting RG may contribute multiple data points to the scatter plot, a single overall breadth- and length-adjusted mutation rate was calculated for each persisting RG. Rates were binned and compared by either initial colonization of their corresponding RG (pre- vs. post-weaning), by pre-weaning diet, and by pre-weaning diet for RGs that colonized pre-weaning vs. those that colonized post-weaning. Significance for each was determined by Mann-Whitney test and visualized using Prism 9. For all RGs that colonized pre-weaning, two temporal mutation rates were determined: one from seeding to the last timepoint before weaning, and one between the immediate timepoints flanking weaning. Pairwise shifts in mutation rates prior to and during weaning were compared via Wilcoxon matched-pairs signed rank test and visualized, both via Prism 9.

The positions of these filtered popSNPs were screened against the coordinate list of each ORF within an RG, previously obtained from the prodigal output of dRep. All ORFs that accrued popSNPs in persisting RGs were annotated via eggNOG-mapper webtool (version 2.1.9) at default parameters (Table 2.10) (60). ORFs annotated as “-“ (n=192/3467), “S -
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

Function unknown” (n=640/3467), and those mutated both pre- and post-weaning (n=20/3467) were removed. The remaining ORFs were assigned by their dietary type and appearance of mutation relative to weaning (ex: breastfed infant, pre-weaning mutation). Overall percentage distribution by COG category was determined for all ORFs in each bin. Bray-Curtis distance between each distribution was calculated using the `vegdist` function (R vegan package) (61) and visualized through principal coordinate analysis using the `pcoa` and `ggplot` functions (R ape and ggplot2 packages) (62). For the four extensively sampled infants, overall percentage distribution by COG category was determined for each individual for mutations that accrued pre-weaning and those that accrued post-weaning. A single distribution was also determined for each of their mothers. Pairwise Bray-Curtis distance was calculated between the pre- and post-weaning distributions of each infant and the overall distributions of each mother. Significance was determined by Mann-Whitney test using Prism 9 and comparisons visualized using the `ggplot` and `stat_compare_means` (test = “wilcox.test”) functions. Pairwise infant-mother comparisons were also binned by same vs. different family and visualized with significance testing using the `ggplot` and `stat_compare_means` (test = “kruskal.test”) functions (R ggplot2 and ggpubr packages).
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

2.6 Data availability

All isolate short reads and assemblies will be made available on NCBI SRA and GenBank prior to manuscript publication. All raw data generated in these analyses are available in Tables 2.2 – 2.10.
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

2.7 References


Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM


Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM


Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM


Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM


Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM


63. Created with Biorender.com.
2.8 Figures

Figure 2.1: 3,995 Reconstructed Genomes are recovered from 10 twin pairs and their mothers. (A) Workflow schematic for the rigorous construction of Reconstructed Genomes from deep short- and long-read sequencing of stool and culture-enriched communities (63). (B) Distribution of the 16,324 Putative Genomes by quality status. (C) Completeness, contig count, log(N50), contamination, and strain heterogeneity comparisons between medium-quality (MQ) Reconstructed Genomes, high-quality (HQ) Reconstructed Genomes, and NCBI reference isolate assemblies. **** corresponds to \( p < 0.0001 \).
Figure 2.2: Evaluating MAG assembly approaches. Summary statistics for (A) total high- (HQ) and medium-quality (MQ) MAGs per stool, (B) average contigs per MAG, and (C) average N50 per MAG. The six columns represent (1) timepoint-filtered long-read meta-assemblies, (2) timepoint-specific short-read meta-assemblies, (3) metaSPAdes merged with filtered long-read meta-assemblies, (4) metaSPAdes --nanopore merged with filtered long-read meta-assemblies, (5) unfiltered long-read meta-assemblies merged with timepoint-specific metaSPAdes, and (6) OPERA-MS. Here, HQ is defined as completeness ≥ 90%, contamination ≤ 5%, strain heterogeneity ≤ 0.5%, and MQ is defined as completeness ≥ 50%, contamination ≤ 5%. “AlignedFracs” refers to only the portion of across-timepoint long-read-assembled (metaFlye) contigs with alignment to timepoint-specific short reads, “highcov” refers to full long-read assembled contigs with ≥ 85% timepoint-specific short-read coverage, and “cmb” refers to “highcov” combined with only the short-read aligned regions of long-read assembled contigs that overall have ≤ 85% coverage.
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

Figure 2.3: Putative and Reconstructed Genome capture by time. (A) Putative Genome count per sample increases until approximately 3 years of life (YOL) before stabilizing. (B) Timepoints 5 and 6, covering the 3 YOL and 7-8 YOL infant samples, harbor more Putative Genomes than the early-life timepoints, taken within the first 2 YOL. (C) Putative Genome count per Timepoint is stable longitudinally. (D) Comparison of Putative Genome counts per individual prior to and after quality filtering. (E) Number of quality-filtered Putative Genomes per dRep secondary cluster during MAG dereplication. (F) Comparison of MAG counts for each sample before and after dereplication.
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

Figure 2.4: Putative and Reconstructed Genome representation by assembly source and sequencing type. Percent of all Putative Genomes, medium- and high-quality (quality-filtered [QF]) Putative Genomes, and Reconstructed Genomes by (A) assembly source (direct stool or culture-enriched) or (B) assembly type (short-read [metaSPAdes] or long-read origin [metaFlye]). For each, completeness, contamination, strain heterogeneity, and ln(N50) are plotted.
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

Figure 2.5: Strain persistence, diversity, and co-occurrence from birth through early adolescence. (A) Phylogenetic tree of the 399 taxa represented by Reconstructed Genomes (RGs). Concentric rings from innermost represent phylum, family, total unique RGs from maternal guts, and total unique RGs from infant guts. (B) Breakdown of total stool samples analyzed, total Reconstructed Genomes, and total Persisting RGs by maternal or infant origin. (C) Total count and transient/persisting split for all taxa with ≥ 20 intra-infant RGs and/or ≥ 10 intra-infant persisting RGs. Taxa (D) Median first (green) and last (red) appearances with 90% confidence intervals plotted for RGs of the ten taxa with greatest persisting RG counts. Taxa are sorted by phylogeny, in order of Eubacteriales (n=3), Bifidobacteriaceae (n=3), and Bacteroidales (n=4). Taxa that are enriched for persisting (light brown) or transient (dark brown) RGs are denoted (*) and ** correspond to q < 0.05 and q < 0.01, respectively). (E) Strain co-occurrence rate for all taxa plotted in C. (F) Number of taxa with multiple co-occurring strains per infant stool by month of life.
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

Figure 2.6: Summary statistics for Persisting and Shared RGs. (A) Distribution of taxa by total persisting RGs across all individuals. (B) Distribution of persisting RGs by total intra-individual samples they are found within. (C) Distribution of shared RGs by total intra-family samples they are present in.
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

### A

![Graph showing the count of RGs (family) and shared RGs over different time periods.](image)

- **Count:**
  - Single individual
  - Multiple individuals

### B

**New Strain Sharing Events per Month Between Twin Pairs**

- **Incidence rate:**
  - 0.00
  - 0.25
  - 0.50
  - 0.75
  - 1.00

- **MOL:**
  - 0-6
  - 6-9
  - 9-12
  - 12-15
  - 15-18
  - 18-21

### C

**Most common taxa representing intra-family strain sharing**

- **Sharing type:**
  - Single
  - Family Triad
  - Dyad: Mother-Infant
  - Dyad: Infant

### D

**Dyads:**

- **Bacteroidales**
  - n=50 shared strains
  - **Density:**
    - MOL
    - **Mother**
    - **Infant**

- **Lachnospiraceae**
  - n=51 shared strains
  - **Density:**
    - MOL
    - **Mother**
    - **Infant**

- **Oscillospiraceae**
  - n=21 shared strains
  - **Density:**
    - MOL
    - **Mother**
    - **Infant**

- **Bifidobacteriaceae**
  - n=30 shared strains
  - **Density:**
    - MOL
    - **Mother**
    - **Infant**
Figure 2.7: Intra-family strain sharing events vary by phylogeny. (A) (i) Distribution of intra-family representative RGs by total individuals that carry them. (ii) Distribution of RGs found in multiple individuals by sharing type. (B) Incidence rate of new strain sharing events per month per twin pair. Bins represent the months compared for each rate calculation. Blue vertical line represents weaning. (C) Breakdown of sharing type for each shared RG per taxon. Taxa are displayed if they meet one or more of the following criteria: ≥ 7 shared RGs, ≥ 6 sharing events between infant-infant dyads, ≥ 3 sharing events between mother-infant dyads, and/or ≥ 4 sharing events between family triads. (D) First appearance of a shared strain in mothers and infants by months post-birth. Vertical lines represent average first appearance for each group. Significance is denoted (** and **** represent q < 0.01 and q < 0.0001, respectively).
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

Figure 2.8: Taxa-specific mutation rates within the infant gut, colored by individual-of-origin. Breadth-adjusted popSNP counts by years since seeding for each persisting RG. Persisting RGs are represented at each timepoint they increase in accumulated popSNP count. All taxa with ≥ 5 persisting RGs are plotted.
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

**Figure 2.9: Taxa-specific mutation rates within the infant gut, colored by appearance in relation to weaning.** Breadth-adjusted popSNP counts by years since seeding for each persisting RG. Persisting RGs are represented at each timepoint they increase in accumulated popSNP count. All taxa with ≥ 5 persisting RGs are plotted.
Figure 2.10: Taxa-specific mutation rates within the maternal gut, colored by individual-of-origin. Breadth-adjusted popSNP counts by years since seeding for each persisting RG. Persisting RGs are represented at each timepoint they increase in accumulated popSNP count. All taxa with ≥ 5 persisting RGs are plotted.
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

Figure 2.11: Generalized mutation rates of persisting RGs in mothers and infants. Aggregate breadth- and genome size-adjusted popSNPs per persisting RG plotted by years since seeding with (A) linear regression or (B) local regression trendline.
Figure 2.12: Mutation rates of persisting RGs by weaning status and pre-weaning diet. Breadth- and genome-size adjusted popSNP counts by years since seeding (A) for all persisting RGs, (B) for persisting RGs seeded pre-weaning, or (C) for persisting RGs seeded post-weaning, each binned by pre-weaning diet.
Chapter 2: Weaning accelerates & transforms within-host evolution in the infant GM

Figure 2.13: The weaning period is a mutation-generating hotspot that triggers a shift in mutated gene functions. (A) Aggregate breadth- and genome size-adjusted popSNPs per persisting RG plotted by years since seeding in the infant gut. Local regression trendlines are drawn for persisting RGs that colonized prior to and post weaning. (B) Mutation rates per persisting RG, binned by seeding in respect to weaning. Significance is denoted (** corresponds to \( p < 0.001 \)). (C) Distributions of gene families representing mutated ORFs, binned by occurrence relative to weaning and diet prior to weaning. (D) Pairwise Bray-Curtis dissimilarity between pre-weaning and post-weaning distributions relative to that of each mother.
**Figure 2.14: No family effect on mutated gene profile distance between infants and mothers.** Pairwise Bray-Curtis dissimilarity between pre-weaning and post-weaning distributions relative to that of each mother, binned by intra- vs. inter-family comparisons.
Chapter 3: Comparative genomics of borderline oxacillin-resistant S. aureus detected during a pseudo-outbreak of MRSA in a neonatal intensive care unit

The contents of this chapter are adapted from a manuscript published in mBio:


* = equal contribution
Chapter 3: Comparative genomics of borderline oxacillin-resistant *S. aureus*

3.1 Abstract

Active surveillance for methicillin-resistant *Staphylococcus aureus* (MRSA) is a component of our neonatal intensive care unit (NICU) infection prevention efforts. Recent atypical trends prompted review of 42 suspected MRSA isolates. Species identification was confirmed by MALDI-TOF MS, and methicillin resistance was re-evaluated by PBP2a lateral flow assay, cefoxitin/oxacillin susceptibility testing, *mecA/mecC* PCR, and six commercially-available MRSA detection agars. All isolates were confirmed *S. aureus* but only eight were MRSA (cefoxitin-resistant, PBP2a-positive, *mecA*-positive, growth on all MRSA screening agars). One MRSA isolate was cefoxitin-susceptible but PBP2a- and *mecA*-positive, and the remaining 33 were cefoxitin-susceptible, PBP2a-negative, and *mecA*-negative; interestingly, these isolates grew inconsistently across MRSA screening agars and had susceptibility profiles consistent with Borderline Oxacillin-Resistant *S. aureus* (BORSA). Comparative genomic analyses found these BORSA isolates to be phylogenetically diverse and not representative of clonal expansion or shared gene content, though clones of two NICU strains were infrequently observed over eight months. We identified 6 features – substitutions and truncations in PBP2, PBP4, and GdpP, and beta-lactamase hyperproduction – that were used to generate a Random Forest classifier to distinguish BORSA from methicillin-susceptible *S. aureus* (MSSA) in our cohort. Our model demonstrated robust ability to predict BORSA phenotype among isolates collected across two continents (validation AUC 0.902). Taken together, we observed an unexpected prevalence of BORSA in our NICU, BORSA misclassification by existing MRSA screening methods, and markers that are together discriminatory for BORSA and MSSA within our cohort. This work has implications for epidemiological reporting of MRSA rates for centers using different screening methods.
Chapter 3: Comparative genomics of borderline oxacillin-resistant *S. aureus*

### 3.2 Importance

Here, we found a high prevalence of *Staphylococcus aureus* isolates exhibiting a borderline oxacillin resistance phenotype (BORSA) in our neonatal intensive care unit (NICU) serendipitously due to the type of MRSA screening agar used by our laboratory for active surveillance cultures. Subsequent phenotypic and molecular characterization highlighted an unexpected prevalence and variability of BORSA isolates. Through whole-genome sequencing, we have interrogated core and accessory genome content and generated a Random Forest classification model to identify mutations and truncations in the PBP2, PBP4, and GdpP proteins, and beta-lactamase hyperproduction, correlated with BORSA and MSSA phenotypes among *S. aureus* clinical isolates collected across two continents. Taken together, this work will help clinical microbiology laboratories and clinicians identify MRSA screening shortfalls and draw attention to the non-
*meca*-mediated BORSA phenotype.
3.3 Introduction

Methicillin-resistant *Staphylococcus aureus* (MRSA) is an important cause of morbidity and mortality in both healthcare and community settings (1). The Centers for Disease Control and Prevention (CDC) defines MRSA as *S. aureus* resistant to methicillin, cefoxitin, or oxacillin by standard susceptibility testing methods, or by the detection of laboratory markers of methicillin resistance (2). Methicillin resistance is most often associated with the penicillin-binding protein PBP2a, encoded by *mecA*. The homolog *mecC* can also confer methicillin resistance but is far less common in the United States (3).

Resistance to beta-lactam antibiotics has been observed in some *mecA/C*-negative *S. aureus* isolates; in these cases, treatment with beta-lactam agents can exacerbate disease burden and result in poorer prognoses (4, 5). Early studies from the 1980s attributed this low-level oxacillin- and/or methicillin-resistance to hyperproduction of beta-lactamase (6), with further work revealing these beta-lactamase hyperproducers (BHP), termed BORSA (borderline oxacillin-resistant *S. aureus*), regain susceptibility upon introduction of a beta-lactamase inhibitor (7, 8). However, beta-lactamase hyperproduction alone was found to be insufficient in some isolates for borderline oxacillin resistance (9, 10), and *mecA/C*-negative, non-BHP isolates with borderline resistance were also recovered (11-13).

A second hypothesis for borderline oxacillin-resistance points to ‘modified’ PBP proteins with lowered drug reactivity and elevated PBP4 levels (14). These isolates were found to have multiple unlinked point mutations initially in *pbp2* (11, 15, 16) and later in *pbp1*, *pbp3*, and *pbp4* (12, 13, 17). Originally termed MODSA (modified PBP *S. aureus*) (14), these isolates are increasingly also referred to as BORSA under a mechanism-agnostic nomenclature (16, 18), as is reflected here. Recently, alternate pathways involving GdpP have been
described in relation to borderline oxacillin resistance (17, 19, 20), complicating our understanding of the phenotype and necessitating further investigation.

Although these meca-negative borderline oxacillin-resistant isolates are considered MRSA by the CDC’s MRSA definition, there is remarkably no consensus definition for BORSA. The prevalence of BORSA reported in the literature from <1% to 12.5% (21), which may be partially driven by local microbial epidemiology and the MRSA surveillance screening employed within hospital infection prevention practices.

In early 2020, an investigation led by the infection prevention team at Saint Louis Children’s Hospital identified instances of irregular positivity from MRSA screening cultures collected from the anterior nares of patients in the neonatal intensive care unit (NICU) in 2019. These instances included scenarios such as a patient with a single positive culture flanked by weekly negative cultures, a patient with an initial positive culture at birth followed by only negative cultures, and patients with sporadic positive cultures spanning an extended timeframe. The microbiology laboratory at Barnes-Jewish Hospital characterized these 42 suspicious isolates, along with 60 comparator isolates recovered from positive blood cultures, by antimicrobial susceptibility testing, genotypic and phenotypic MRSA surrogate marker detection (e.g., meca and PBP2a, respectively), and whole-genome sequencing (WGS). Our objective was to profile the genomic relatedness of these isolates for outbreak assessment, generate a computational model to identify genomic correlates of the BORSA phenotype, and evaluate the performance of MRSA screening agars to detect MRSA and BORSA.
3.4 Results

3.4.1 Characterization of isolates from the NICU.

Forty-two isolates were characterized as part of a NICU MRSA screening culture investigation (Figure 3.1). These isolates were originally reported as MRSA based on growth characteristics on Spectra MRSA screening agar. This study began by confirming the 42 isolates were *S. aureus* using MALDI-TOF MS. To confirm methicillin resistance, isolates were screened for PBP2a and the PBP2a-encoding gene *mecA* using a rapid qualitative immunochromatographic assay and PCR, respectively. Interestingly, only 9 isolates tested positive for PBP2a (Table 3.1). In support, *mecA* was detected only in the 9 PBP2a-positive isolates and was not detected in the PBP2a-negative isolates (Table 3.1). Of note, one PBP2a/mecA-positive isolate was from a culture with two *S. aureus* colony morphologies (Table 3.2). An in-house PCR for the *mecA* homolog *mecC* was also performed (22, 23), with no isolates returning a positive result (Table 3.2). Ultimately, 33 isolates remained in question.

An additional 60 *S. aureus* isolates from blood cultures of 60 patients were also included in this study to serve as comparators to contextualize the NICU isolates. These isolates include 50 consecutive *S. aureus* isolates recovered from blood cultures at the Barnes Jewish microbiology laboratory (regardless of age or patient location) and all non-MRSA isolates recovered from blood cultures of NICU patients in 2019 (n=10). All 60 isolates were confirmed as *S. aureus* by MALDI-TOF MS, negative for PBP2a, cefoxitin susceptible, and not detected for *mecA/mecC* PCR (Tables 3.2 – 3.3), initially indicating MSSA status. These isolates are hereafter referred to as the comparator blood isolates.
3.4.2 Spectra MRSA Screening Agar.

Potential explanations for recovering non-MRSA isolates from the MRSA screening agar are faulty media or inaccuracies in interpreting the growth on the culture media. To determine whether these factors contributed, the 42 isolates (including the 9 confirmed MRSA isolates) were retested on Spectra MRSA screening agar, with representative images shown in Figure 3.2A. Interestingly, all but one isolate exhibited growth and demonstrated the MRSA-expected denim-blue pigmentation. The growth pattern was somewhat varied in terms of colony size and abundance (as expected based on further resistance mechanism characterization, see below). In general, the confirmed MRSA isolates had the most robust growth, as expected. However, the majority of the remaining isolates exhibited reduced growth (e.g., smaller colony sizes and/or reduced colony forming units compared to MRSA control strains) (Table 3.1). The discrepancies between the Spectra MRSA screening agar and the findings from the PBP2a and mecA PCR testing were surprising – this was in stark contrast to the 60 comparator blood isolates where 55 isolates had no growth, 3 had robust growth, and 2 had rare growth (Table 3.3).

3.4.3 Cefoxitin and Oxacillin Susceptibility Testing.

According to the Clinical Laboratory Standards Institute (CLSI), resistance to cefoxitin via disk diffusion is a surrogate marker to predict mecA-mediated methicillin resistance (MRSA). To determine cefoxitin resistance, the 42 isolates were tested using the disk diffusion method as described by CLSI. Eight of the 9 MRSA isolates (PBP2a- and mecA-positive) were resistant (Table 3.1). One MRSA isolate was susceptible to cefoxitin (zone size = 25
mm). Two more isolates tested resistant near the breakpoint (both 19 mm) but were PBP2a- and mecA-negative. None of the 60 comparator blood isolates tested cefoxitin resistant.

Susceptibility testing was also performed for oxacillin resistance. First, isolates were tested by the oxacillin disk diffusion method and interpreted using archived 2007 CLSI breakpoints (24). These historical oxacillin breakpoints were archived when the focus shifted specifically to detection of mecA-mediated beta-lactam resistance. Of the 42 isolates, 15 were resistant including the 8 confirmed MRSA isolates (Table 3.1). The lone cefoxitin-susceptible MRSA isolate was oxacillin intermediate. Two additional isolates were also oxacillin intermediate. None of the 60 comparator blood isolates tested oxacillin resistant or intermediate by disk diffusion. Following this, isolates were tested for oxacillin resistance using a previously reported gradient diffusion method that induces resistance by plating to 2% NaCl screening agar (25-27). Thirty-three of the 42 isolates were found to be resistant (Table 3.1). Nine of the 60 comparator blood isolates also tested oxacillin resistant (Table 3.3). For subsequent analyses, BORSA isolates were defined using the following criteria: mecA/PBP2a-negative S. aureus AND (1) resistant or intermediate to oxacillin by disk diffusion (zone size ≤12 mm) OR (2) resistant to oxacillin by gradient diffusion using inducible 2% NaCl agar (consistent MIC ≥4 µg/mL tested in triplicate). This definition captured 24 investigated isolates and 9 comparator isolates as BORSA, with 9 and 51 investigated and comparator isolates designated MSSA, respectively. The nine BORSA isolates from the comparator collection were from 9 patients with an age range of 7-77 years (median age of 49 years) receiving care at three different hospitals, suggesting that the isolates with this phenotype were not restricted to the neonatal ICU. Further investigation found no epidemiological commonalities associated with these nine isolates from the comparator collection.
Beta-lactamase hyperproduction is a non-\textit{mecA} resistance mechanism that could contribute to the above borderline oxacillin-resistant \textit{S. aureus} (BORSA) phenotype (6-8). To detect beta-lactamase production in these \textit{S. aureus} isolates, two CLSI-recommended tests were performed: disk diffusion penicillin zone-edge test and nitrocefin-based test. The disk diffusion method demonstrated 38 investigated isolates were beta-lactamase positive (Table 3.2). These findings reflected the nitrocefin test results, except for isolate 308 which had a negative cefinase test. Of the comparator 60 blood isolates, 44 were positive by both methods, 15 were negative by both methods, and isolate 56 was cefinase-positive and zone-edge test negative (Table 3.2).

Hyperproduction of beta-lactamase has been previously distinguished as a resistance mechanism in BORSA isolates when the addition of a beta-lactamase inhibitor lowers the MIC of the parent drug by at least two dilutions when tested in combination with the beta-lactamase inhibitor (13). All isolates were tested against amoxicillin and amoxicillin-clavulanate acid using the established gradient diffusion method. Six of the 33 isolates defined as BORSA exhibited a four-fold difference (Table 3.2), while just 6 isolates of the 60 isolates defined as MSSA also exhibited a four-fold difference (Table 3.2); taken together, the difference in lactamase inhibitor effect between BORSA and MSSA isolates was significant \((p=0.0030, \text{Mann-Whitney, Figure 3.3})\).

3.4.4 \textit{MRSA} Screening Agar Comparison.

To determine how these \textit{S. aureus} isolates with differing cefoxitin and oxacillin resistance profiles performed on other commercially available MRSA screening media, 5 additional MRSA screening agars were evaluated: MRSASelect II, BBL CHROMagar MRSA II,
Chapter 3: Comparative genomics of borderline oxacillin-resistant S. aureus

chromID MRSA, nonchromogenic MRSA screen agar, and HardyCHROM MRSA. The 8 cefoxitin-resistant MRSA isolates grew on all agars, and were the only isolates to grow on MRSA-Select II, BBL CHROMagar MRSA II, and chromID MRSA. The nonchromogenic MRSA screen agar grew an additional 13 isolates (12 BORSAs and the lone cefoxitin-susceptible MRSA isolate), while the HardyCHROM MRSA agar grew an additional 16 isolates (14 BORSAs, 2 MSSAs) (Table 3.1). For the comparator blood isolates, none of the 60 isolates grew on MRSA-Select II, BBL CHROMagar MRSA II, and chromID MRSA; however, one BORSA isolate grew on the nonchromogenic MRSA screen agar and 3 isolates (2 BORSAs, 1 MSSA) on the HardyCHROM MRSA (Table 3.3). Taken together, these findings suggested differing selective properties across commercially available agars resulting in different isolates flagging as MRSA, irrespective of PBP2a and mecA presence. Representative images are shown in Figure 3.2B.

To evaluate the analytical performance of the different MRSA screening agars to detect MRSA and BORSAs, sensitivity and specificity were calculated using all 102 isolates. As shown in Table 3.4, sensitivity and specificity varied by agar and the ability to detect MRSA and BORSAs. Briefly, for mecA-mediated MRSA detection, BBL CHROMagar MRSA II, MRSA-Select II, and chromID MRSA demonstrated 89% sensitivity and 100% specificity. The nonchromogenic MRSA screen agar and Spectra MRSA had 100% sensitivity and specificities of 86% and 60%, respectively. For detection of mecA-positive isolates and BORSAs isolates using criteria 1, BBL CHROMagar MRSA II, MRSA-Select II, and chromID MRSA had 50% sensitivity and 100% specificity. Spectra MRSA exhibited 100% sensitivity and 65% specificity. For detection of mecA-positive isolates and BORSAs isolates using criteria 2, BBL CHROMagar MRSA II, MRSA-Select II, and chromID MRSA had 20% sensitivity and 100% specificity. The non-chromogenic MRSA screen agar and Spectra MRSA agar had, respectively, sensitivities of
54% and 76% and specificities of 100% and 83%. While no screening agar had perfect analytical performance characteristics, laboratories may opt to use a specific agar depending on the desire to detect MRSA and BORSA isolates or to focus detection specifically on *mecA*-positive strains.

### 3.4.5 WGS reveals that the BORSA phenotype is not linked to the core or accessory genome.

Prior investigations into mechanisms of borderline oxacillin resistance in *S. aureus* isolates have been limited to amplicon sequencing of pre-selected genes following reported phenotypic associations. Whole genome sequencing (WGS) of human clinical BORSA isolates is rarely performed (12, 20, 28) and only one draft assembly is presently available on NCBI (18). Here, we performed WGS on the 33 human clinical BORSA isolates, as well as the 9 MRSA and 60 MSSA isolates. One MSSA isolate (isolate 35) was dropped due to low coverage resulting in 101 high-quality isolate assemblies with >99% completeness and <1% contamination (Table 3.5).

To determine the population structure of BORSA within our *S. aureus* cohort and conduct a gene-unbiased investigation, we annotated all protein coding sequences of each isolate genome with Prokka and constructed a core genome alignment with Roary (29, 30). From this we generated a maximum-likelihood phylogenetic tree reflecting 1,859 genes shared among >99% of isolates at >95% identity, using FastTree and iTOL (31, 32) (Figure 3.4A). All major branch points indicated >90% bootstrap support values, demonstrating high confidence in the phylogenetic reconstruction. Lineages and clusters were identified
Chapter 3: Comparative genomics of borderline oxacillin-resistant *S. aureus*

with hierBAPS and annotated, depicting the evolutionary relationships of the *S. aureus* genomes within our cohort (33). Though five lineages were identified, more than half of the isolates within our dataset were members of Lineage 1. BORSA isolates represented 12 different multilocus sequence types (ST), with those from ST398, ST15, ST97, and ST8 comprising 57.6% of the BORSA cohort. Two of the most prevalent BORSA sequence types, ST97 (n=5) and ST398 (n=5), have historically been contextualized as livestock-associated MRSA (34-38) though recent reports from France and the US have described ST398 isolates of our predominating *spa*-type t1451 as human bloodstream infection (BSI)- and hospital-associated MSSA (39-41), reflecting the demographics of our cohort. Moreover, a German study also found ST398 to be the third-most common sequence type among their BORSA isolates of clinical origin (42). Yet, to the best of our knowledge, only five ST97 isolates have previously been annotated as BORSA (20, 43). Conversely, ST25, often associated with borderline oxacillin resistance in the literature (13, 16), was absent from our dataset. We believe these are the first reports of BORSA within ST27 (n=1) and ST72 (n=1).

Given the longitudinal collection scheme of this cohort, we assessed for both the potential of a same-strain BORSA outbreak within our NICU and the presence of *S. aureus* strains that may be longitudinally detected within our hospital system. For this we employed snp-sites and snippy to identify clusters of isolates that are within 30 whole genome single nucleotide polymorphisms (SNPs) of each other. Of the 101 analyzed isolates, we uncovered two clusters comprising a combined 4 BORSA and 3 MSSA isolates from 7 patients spanning 8 months in the same NICU (Figure 3.4B), highlighting existing but infrequent longitudinal detection of *S. aureus* strains that colonized non-overlapping inpatients within our hospital system, and providing evidence against a same-strain BORSA outbreak.
Chapter 3: Comparative genomics of borderline oxacillin-resistant *S. aureus*

Though the core genome alignment strongly reflected genomic similarity among isolates of shared MLST and spa-types, we strikingly did not observe strong clustering of isolates by oxacillin resistance (Figure 3.4A). Moreover, the distributions of all lineages and ST groups by BORSA/MSSA status were not significantly different ($p>0.30$ and $p>0.35$, respectively; FDR-corrected Fisher’s exact test) (Figure 3.5). Through ancestral state reconstruction of the borderline oxacillin resistance phenotype, we found independent acquisition in 17 isolates, with the remaining 16 isolates represented among six shallow monophyletic clades of 2-3 isolates each (Figure 3.6). Taken together, these data suggest multiple recent and sporadic acquisitions of borderline oxacillin resistance among formerly MSSA isolates, with occasional vertical transfer from an immediate ancestor.

Given the lack of clustering by core genome alignment, we considered whether BORSA isolates may instead be united by a similar accessory genome. Consistent with previous reports (44-46), we observed an open pangenome architecture among our *S. aureus* isolates, implying sufficient depth for isolate discrimination by accessory gene content. Ordination by accessory gene content, however, did not support a distinction among isolates for borderline oxacillin resistance (Figure 3.4C). A more targeted analysis surveying antimicrobial resistance (AMR) gene content validated *meca* PCR results in MRSA isolates and found that resistance gene repertoires correlated with isolate ST. However, AMR analysis similarly found no clustering by borderline oxacillin resistance, implying that aside from this shared phenotype, BORSA isolates have a non-uniform array of resistance repertoires (Figure 3.4D). In summary, we did not observe a common BORSA signature by core genome alignment, accessory genome content, multilocus sequence typing (MLST), or encoded AMR profile.
3.4.6 Diverse substitutions in canonical BORSA-associated proteins.

Oxacillin and/or methicillin resistance in *mecA/C*-negative *S. aureus* isolates has historically been associated with multiple unlinked amino acid (AA) substitutions in the PBP proteins (12, 14-16, 47) and GdpP (17, 19, 20). We used Prokka to identify and extract AA sequences of these proteins from all BORSA and MSSA isolates to generate protein-specific AA alignments, from which we assembled a consensus sequence for each protein of interest and identified isolate-specific substitutions and truncations (48, 49) (Table 3.6). We compared our observations with mutational profiles of proteins of interest of 86 clinical and lab-grown BORSA isolates from prior reports (11-13, 16, 17, 19, 47, 50, 51). Though we similarly observed a breadth of substituted sites across the five proteins (Figure 3.7A), only 37% of unique AA substitutions found in our cohort have been reported in published BORSA-association studies (Figure 3.7B). Remarkably, of these AA substitutions present in our isolates that were previously reported, 96% are jointly or exclusively found in our MSSA cohort as well. Even within our cohort, BORSA isolates did not definitively cluster by their substitutional profile (Figure 3.7A), and a majority of mutated residues within these five proteins were found within both BORSA and MSSA isolates (Figure 3.7B). Both the paucity of overlapping substitution sites across studies and the abundance of overlap between our BORSA and MSSA isolates may be attributable to the geographic distance in isolate collection sites, as most published isolates were collected in Belgium, Canada, or Scotland, while our BORSA and MSSA isolates were collected from St. Louis, USA. These data indicate that even among canonical proteins of interest, BORSA isolates are characterized by a constellation of AA substitutions and protein truncations that are often simultaneously present in MSSA isolates.
3.4.7 A sparse Random Forest Classifier (RFC) identifies correlates and anticorrelates of BORSA phenotype among isolates of varied oxacillin susceptibility from an American and Belgian cohort.

We endeavored to parse the overlapping AA substitutional data to identify a detectable signature specific to our BORSA isolates. Isolates 307, 338, 340, and 342 were removed for this analysis as sequencing data indicate they are clones (≤10 whole genome SNPs) of the first isolate taken from the same patient (n=4) largely at the same sampling event (Figure 3.8, Table 3.2). To increase our sample size, we incorporated an additional dataset of 32 clinical BORSA isolates with published gene sequences for pbp1-4 and gdpP, as well as MLST and beta-lactamase inhibitor effect metadata (13) (Figure 3.9A). Together with our curated BORSA and MSSA cohorts, this sample set was used to construct a supervised machine learning model trained on the presence or absence of 123 unique AA substitutions across PBP1-4 and GdpP, truncations to PBP2 and/or GdpP, presence of blaZ, MLST, and beta-lactamase inhibitor effect (Figure 3.9B-C). We refined our RFC through a two-fold feature elimination process in which 56 redundant features were first removed by correlation analysis. Of the remaining 72 features, ten-fold cross-validation indicated that just six are sufficient for accurate prediction of BORSA status. We then trained a sparse RFC on the six most informative predictors (Figure 3.10B) and achieved an average validation area under receiver-operator curve (AUROC) of 0.902 ± 0.009 over 100 iterations (Figure 3.10A). This model correctly classified 16 of 16 BORSA isolates and 18 of the 21 MSSA isolates (91.9% accuracy, Figure 3.10C). The top BORSA-correlated and anti-correlated features among S. aureus isolates within these cohorts were truncations to GdpP or PBP2, beta-lactamase inhibitor effect, and the AA substitutions GdpP I52V, PBP2 A285P, and PBP4 T189S (Figure 3.10B, Table 3.7),
Chapter 3: Comparative genomics of borderline oxacillin-resistant \textit{S. aureus}

representing a combination of lineage markers that allow accurate discrimination of BORSA and MSSA. Implementation of a phylogenetically informed association tool provided further support for the strong positive association of GdpP truncation with BORSA status, finding trait convergence over four independent genotype transitions within our cohort (Figure 3.11). Our supervised machine learning model trained on a subset of our cohort accurately identified whether a \textit{S. aureus} isolate in the withheld dataset was borderline oxacillin-resistant or methicillin-susceptible based on beta-lactamase hyperproduction, truncated GdpP or PBP2, and three substitutions in PBP2 and PBP4, despite high-level heterogeneity in AA substitutions within canonical BORSA-associated proteins.

3.5 Discussion

Here, we report a pseudo-outbreak of suspected MRSA, and an ensuing diagnostic and genomic characterization of 102 clinical \textit{S. aureus} isolates of varying oxacillin susceptibility. These results confirm prior reports of diverse mutations in PBP proteins of BORSA isolates (11-13, 16, 17, 47, 51); however, the critical inclusion of a MSSA isolate bank from the same institution as comparators reveal that many of the previously reported BORSA-linked amino acid mutations from geographically distant collection sites are either not found in any of our cohorts or are shared by BORSA and MSSA isolates alike. Given we did not observe a core or accessory genome signature among BORSA isolates as distinct from MSSA or MRSA isolates, we leveraged our sequencing data to construct a supervised machine learning model with input features largely based on the mutational profile of historically BORSA-associated proteins. Through this work we generated a Random Forest Classifier trained on just six features that yielded a robust diagnostic accuracy for the detection of BORSA among MSSA isolates.
within our cohorts. Among these features are mutations and/or truncations to PBP2 and PBP4, beta-lactamase hyperproducer effect, which was observed in some but not all our BORSA isolates, and a mutation in and truncation of the GdpP protein.

GdpP is a phosphodiesterase that catalyzes the hydrolysis of intracellular secondary messenger c-di-AMP, and *S. aureus gdpP*-deletion mutants have been shown to have elevated c-di-AMP and PBP4 levels, thicker cell walls, reduced cell size, and reduced susceptibility to beta-lactams and other cell wall-targeting antimicrobials (19, 20, 50, 52). Our study emphasizes its importance to the BORSA phenotype, as the sole feature exclusively present in certain isolates in our BORSA cohort and prior BORSA reports is a truncated GdpP protein (Figure 3.7B, BORSA (Previously Reported) and BORSA (This Study) overlap). Of significance, the lone BORSA clone among the Cluster 2 isolates (Figure 3.4B) harbors a premature STOP codon in GdpP that is not present in any of the three MSSA clones that all together differ by a maximum of 17 whole genome SNPs. Moreover, the two isolates collected from Patient 33 are only 10 whole genome SNPs apart (Figure 3.8E), yet a SNP resulting in a premature STOP codon is present in isolate 344 (BORSA) and absent in 343 (MSSA). 100% of the BORSA isolates that encode premature STOP codons in their *gdpP* genes (n=4, Figure 3.7A) were shown to be oxacillin-resistant by both disk diffusion (zone size ≤ 10 mm) and gradient diffusion (Table 3.1). In addition, one of the 3 isolates with expected full-length GdpP proteins that also met this conservative criteria is the sole isolate in our cohort to harbor a D349Y substitution within the Desert hedgehog (DHH) motif of GdpP, the domain responsible for c-di-AMP phosphodiesterase activity (19, 50). Taken together, our data demonstrate a major-
ity of isolates with the strongest BORSA phenotype (criteria 1 & 2) have truncated or mutated GdpP proteins, and one of the few discriminatory SNPs between nearly identical BORSA and MSSA isolates results in a premature STOP codon in GdpP.

Many BORSA investigations utilize lab-adapted strains (11, 47, 50) or are multicenter studies (13, 16, 20). Among the few large-scale single-center studies (4, 53, 54), ours is remarkable as the first to indicate against a clonal outbreak within our institution, as supported by WGS of isolates. Indeed, we find BORSA isolates within our cohort to be spread across lineages and clades, phylogenetically dispersed among MSSA and MRSA isolates. Our data further support independent acquisition of borderline oxacillin resistance in a majority of our BORSA isolates, with infrequent vertical transfer of resistance from an immediate common ancestor. WGS was critical to these findings, yet there is a critical paucity of human colonizing or infecting BORSA isolate assemblies uploaded to the public database NCBI GenBank. Our work provides the addition of 33 high-quality BORSA assemblies for public use.

A challenge facing clinical microbiology laboratories is identifying and reporting BORSA due to the lack of a consensus definition and an established mechanism. This also complicates comparisons between studies and across previously published work. This work used several methods and two distinct BORSA definitions to help address these variabilities. A limitation of this study is that data were generated from a sole clinical microbiology laboratory. However, having a single laboratory perform these studies enabled us to standardize methods and better compare BORSA isolates to other S. aureus isolates in our microbiology region. In addition, isolates were collected from multiple hospitals and patient populations within our geographic area. Comparing our clinical investigation to previous studies, our
findings are generalizable (11-13, 16, 17, 19, 47, 50, 51). Other limitations include the targeted isolate selection criteria and lack of BORSA prevalence data prior to this investigation.

A common challenge and limitation of machine learning models is their potential to learn lineage markers rather than causal variants. Though we did not find MLST to be a strong contributor in the unrefined RFC (Figure 3.9B) or find BORSA/MSSA status to be significantly different by lineage or sequence type, the potential influence of phylogenetic relatedness cannot be discounted. Given the lack of publicly available sequencing data available for BORSA isolates, we are unable to evaluate the robustness of our model for BORSA isolates beyond those included in this study. Further work reproducing our findings is therefore necessary before this sparse model can be considered generalizable.

MRSA screening can have significant impacts on patient care and clinical workflows. For example, MRSA-colonized infants are placed on contact precautions and may undergo decolonization regimens such as intranasal mupirocin to reduce the risk of subsequent infection. Misidentification of MRSA can lead to unnecessary use of isolation precautions and overuse of antimicrobial agents. In addition, MRSA screening is used to establish accurate hospital-acquired infection rates. In 2017, MRSA rates were included in the reimbursement equation in Medicare’s Hospital Value-Based Purchasing (VBP) program. This can have immense financial implications for hospitals with penalties of a 1% reduction in payments if they are in the bottom 25%. With such high stakes on MRSA rates, it is critical to properly evaluate performance differences between MRSA screening methods. Nearly all molecular MRSA screening approaches look for the genomic markers mecA and/or mecC, yet BORSA is not mecA/C-mediated. Though it is clear BORSA detection varies by MRSA screening method and thus impacts MRSA rates, the degree of significance is dependent on BORSA prevalence,
which is unknown at most institutions. Based on College of American Pathologists MRSA surveys in 2021, laboratories commonly use the following MRSA screening agars assessed in the presented work, all of which were assessed for MRSA and BORSA detection specificity and sensitivity here: BBL CHROMagar MRSA II (109 labs), chromID MRSA (56 labs), MRSA-Select II (100 labs), HardyCHROM MRSA (8 labs), and Spectra MRSA (144 labs).

The resistance mechanisms of the BORSA phenotype are complex and not fully understood. In this study, the phenotypic variability of BORSA was highlighted by growth differences across screening agars, including the initial clinical MRSA surveillance testing. Within our cohort, one isolate (isolate 310) retested inconsistently (Oxacillin gradient diffusion MIC = 2, 4, and 4 µg/mL on repeat testing). While it is unlikely that oxacillin would be used clinically for treatment of infection with such an isolate, we chose to conservatively include its genome among the MSSA cohort due to its inconsistent borderline susceptibility to oxacillin, and to be consistent with our in-text classification of BORSA as consistent MIC ≥4 µg/mL. One explanation for variable “breakthrough” is expression differences of resistance mechanisms within subpopulations. This “breakthrough” should not discourage use of MRSA screening agars, but should warrant follow-up testing.

The present work led to modification of clinical susceptibility testing of \textit{S. aureus} isolates in our microbiology laboratory. Our laboratory uses customizable disk diffusion panels which now include both cefoxitin and oxacillin. Cefoxitin-susceptible isolates are evaluated for an oxacillin non-susceptible phenotype using CLSI 2007 breakpoints (susceptible is >12 mm; intermediate is 11-12 mm; resistant is <11 mm). Isolates consistent with BORSA are reported as such to aid in selection of antimicrobial therapy. This will also help our institution establish BORSA prevalence and evaluate treatment outcomes for future studies.
Screening agars could serve as a reasonable alternative, and clinical microbiology laboratories will likely continue to use screening agars to survey for resistant strains of *S. aureus*.

In summary, our work was prompted by the investigation of a possible MRSA outbreak in our NICU, which was determined to be non-clonal BORSA. We characterized this collection of isolates and subsequently established BORSA definitions based on phenotypic and genotypic findings, reevaluated commonly used MRSA screening agars, and implemented an accurate computational classifier which distinguishes BORSA from MSSA/MRSA using only six genetic and phenotypic features.

### 3.6 Materials & Methods

#### 3.6.1 Study cohort.

Institutional review board (IRB) approval was obtained for this study. MRSA surveillance cultures from anterior nares specimens of all NICU patients are performed as part of standard of care at St. Louis Children’s Hospital (SLCH) upon admission and weekly thereafter. Isolates from positive cultures are frozen as part of routine clinical protocols. Due to increased MRSA colonization rates in 2019, the SLCH Infection Prevention team reviewed all NICU patients with positive MRSA surveillance cultures and identified several cases with irregular results. The 32 cultures in question were taken from 31 subjects (one patient had two cultures that were evaluated); the median age of the subjects was 41 days and an age range of 0-523 days. The microbiology laboratory at Barnes-Jewish Hospital reevaluated the NICU MRSA isolates using frozen stocks. Four of the 32 freezer stocks grew out 2 colony morphologies, resulting in 36 isolates independently characterized from the NICU MRSA screening cultures. This investigation led to characterization of an additional 6 isolates that
were collected from routine MRSA screening of non-NICU patients, which grew on MRSA screening agar despite being PBP2a-negative. In total, there were 42 isolates under investigation (Table 3.1, 301-346). This nonconsecutive order reflects the initial observation of different colony morphologies from freezer stocks that were later determined identical (Figure 3.8).

The 42 isolates under investigation were selected by subject matter experts in Infection Prevention who were concerned about unexpected positive culture results (i.e. the initial cohort represented an enriched sample set biased for an unusual phenotype). To put this enriched sample set into context, a second isolate collection was created using 50 consecutive cefoxitin-susceptible, PBP2a-negative S. aureus isolates (designated 1-50) recovered from blood cultures in the microbiology laboratory at Barnes-Jewish Hospital (this laboratory also serves SLCH) (Table 3.3). To further supplement this comparator collection, all S. aureus isolates recovered from blood cultures from patients in the NICU in 2019 (n=10) were included (designated 53-62). Taken together, the comparator blood isolates totaled 60, bringing the total sample set to 102 isolates.

3.6.2 Identification and Resistance Characterization.

Microbial identification was confirmed by Bruker BioTyper MALDI-TOF MS (Bruker, Billerica, MA). For MRSA designation, PBP2a SA Culture Colony Test (Alere) was performed according to manufacturer’s instructions using 18-24 hour subculture growth. The presence of mecA and the homolog mecC, were determined by in-house PCRs (22, 23). Susceptibility testing included cefoxitin by disk diffusion and oxacillin by disk diffusion and gradient diffusion. Methods followed the procedural guidelines outlined by the Clinical and Laboratory
Standards Institute in the M02 and M23 (55, 56). Disk diffusion testing of cefoxitin (Hardy Diagnostics) and oxacillin (BD) was performed on conventional Mueller-Hinton agar (MHA) (Hardy Diagnostics). Gradient diffusion testing of oxacillin (bioMerieux) was performed on MHA with 2% NaCl agar (Remel). Detection of beta-lactamase production was assessed by the disk diffusion penicillin zone-edge test (Hardy Diagnostics) and nitrocefin-based cefinase disk test (Hardy Diagnostics). Beta-lactamase inhibitor rescue phenotype was determined by assessing the fold change in minimum inhibitory concentration (MIC) of amoxicillin (bioMerieux) and amoxicillin-clavulanic acid (bioMerieux) gradient diffusion testing.

3.6.3 MRSA Screening Agars.

Spectra MRSA screening agar (Remel) is routinely used by the Barnes-Jewish Hospital microbiology laboratory for MRSA active surveillance cultures. For clinical active surveillance cultures, anterior nares samples were collected using the BD E-Swab specimen collection and transport kit, plated to Spectra MRSA agar, and incubated/analyzed using the Kiestra laboratory automation system (BD) (57). Images were captured using Kiestra system and modified by cropping and rotating for alignment. Five other commercially-available MRSA screening agars were also used in this study: chromID MRSA (bioMerieux), MRSASelect II (Bio-Rad), HardyCHROM MRSA (Hardy Diagnostics), BBL CHROMagar MRSA II (BD), and nonchromogenic MRSA screen plate (Hardy Diagnostics). To compare growth, 18-24 hour growth off blood agar plates were suspended at a 0.5 McFarland and 50 µL spotted to each agar. Isolates were cultured two per plate and struck for isolation, except for the nonchromogenic MRSA screen plate. For this, 8 isolates were cultured per plate by spotting 10 µL. Growth of all screening agars was analyzed after 18-24 hours at 35°C in an air incubator.
Images were modified by cropping, rotating for alignment, and adjusting modestly for contrast and brightness (images contain MRSA and MSSA isolates on same plate/image for uniformity). Black lines were superimposed on some images to assist readers with differentiating strains on top and bottom.

3.6.4 Illumina WGS.

Isolate DNA was extracted manually as previously described (58). DNA was quantified with the Quant-iT PicoGreen dsDNA assay (Thermo Fisher Scientific, Waltham, MA, USA). Isolate DNA at 0.5 ng was used as input for Illumina library preparation using a Nextera kit (Illumina, San Diego, CA, USA). Libraries were pooled at equal concentrations and sequenced on the Illumina NextSeq 500 HighOutput platform (Illumina, San Diego, CA, USA) to a depth of 2.5 million reads per sample (2 x 150 bp). Illumina adapter sequences were removed from demultiplexed reads using Trimmomatic (version 0.38) with the following parameters: leading, 10; trailing, 10; sliding window, 4:15; minlen, 60 (59). Potential human read contamination was removed using DeconSeq (version 0.4.3) with default parameters (60). Processed reads were then assembled into draft genomes using the de novo assembler SPAdes (version 3.13.0) with parameters -k 21,33,55,77 –careful (61). The scaffolds.fasta outputs were used for all downstream analyses. Draft genomes were determined to be >99% complete and <1% contaminated by CheckM (version 1.0.13), and assembly quality was calculated using QUAST (version 3.2) (62, 63). High-quality assemblies were annotated using Prokka (version 1.14) with --mincontiglen 500 to identify open reading frames (29).
3.6.5 Core genome analysis.

The general feature format (.gff) files outputted by Prokka were used to construct a core genome alignment for all BORSA, MRSA, and MSSA isolates through Roary (version 3.12.0) with default parameters (30). The alignment, composed of 1,859 genes shared by all isolates at minimum 95% identity, was used to generate an unrooted maximum likelihood tree with FastTree (version 2.1.10) (31). The resulting newick file was visualized in iTOL (32). In silico multilocus sequence typing (MLST) and spa-typing were performed using the webtools MLST 2.0 and spaTyper 1.0 maintained by the Center for Genomic Epidemiology (64, 65). MLST, spa type, and class were viewed as color strips in iTOL. Lineages identified by hierBAPS during fastGEAR analysis were also marked on the trees (33). The ancestral state of borderline oxacillin resistance was estimated as a discrete trait using the maximum likelihood-based ace function (R ape package) (66).

3.6.6 SNP distance determination.

The snp-sites (version 2.4.0) tool was used to call isolate-specific SNPs against the core genome alignment file created by Roary (67). A core genome SNP threshold of 600 was determined to be discriminatory for closely related isolates through an all-vs-all comparison, and isolate pairs with distances below this threshold were binned into clusters. Snippy (version 4.4.3) was then employed to call whole genome SNPs between isolate pairs for each pair in the same cluster (68). Isolates were considered clones of the same persisting strain if their whole genome SNP distance (Snippy VariantTotal) was below 30 (≥ 99.999% ANI) (69). Rooted SNP distance trees were created with the R ape package.
3.6.7 Accessory genome analysis.

The gene_presence_absence.Rtab output file from Roary was purged of core genes and used to calculate the Jaccard distance between all isolates through the `vegdist` function (R vegan package) (70). Clustering by accessory gene content was visualized through principal coordinate analysis using the `pcoa` and `ggplot` functions (R ape and ggplot2 packages) (66, 71).

ARGs were initially annotated in silico against the NCBI comprehensive database of acquired and intrinsic antimicrobial resistance proteins at >90% identity using AMRfinder (version 3.8.4) (72). A presence-absence matrix was generated using the `pheatmap` function (R pheatmap and dendsort packages) (73, 74), in which isolates were clustered by ARG presence, with associated metadata displayed as color strips to represent isolate class, sequence type, and expected antimicrobial resistance (aminoglycoside, beta-lactams, fosfomycin, fusidic acid, lincosamide, macrolide, mupirocin, quaternary ammonium, and tetracycline).

3.6.8 Selected protein investigation.

The PBP1-4 and GdpP proteins were chosen for further investigation based on published literature (12, 13, 16, 17, 19). Amino acid sequences with the following headers were extracted from Prokka.faa output files: Penicillin-binding protein; D-alanyl-D-alanine carboxypeptidase; Cyclic-di-AMP phosphodiesterase GdpP. The PBP1-4 and GdpP amino acid sequences from 32 published BORSA isolates (13) were also incorporated (GenBank MF070915.1 – MF071042.2, MF071075.1 – MF071106.1). Sequences alignments for each protein were performed using Clustal Omega and visualized with Jalview (48, 49). All substitutions, deletions, and protein truncations against the consensus sequence were manually
Chapter 3: Comparative genomics of borderline oxacillin-resistant *S. aureus*

curated. Results were compared against those found in previous BORSA reports (11-13, 16, 17, 19, 47, 50, 51), with overlap visualized with the *euler* function (R eulerr package) (75).

3.6.9 Random Forest classification of BORSA phenotype.

A custom machine learning process employing Random Forest analysis was used to distinguish between the aforementioned BORSA and MSSA isolates. The initial model consisted of 128 features, including amino acid mutations in the PBP1-4 and GdpP proteins, presence of *bla*Z, MLST, and beta-lactamase inhibitor effect, and was trained on 70% of the isolate dataset, with 30% withheld for validation. The *randomForest* function (R randomForest package) was utilized at default parameters with the following adjustments to establish a baseline model for optimization: ntree=5,000, proximity=FALSE, importance=TRUE. A feature elimination step was then performed to improve performance of the subsequent model iterations. To minimize redundancy, we used the *cor* and *findCorrelation* functions (R STATS, caret packages) to calculate Spearman correlation coefficients across all features, with highly correlated features determined as those with >0.75 correlation (76, 77). A representative of each set of correlated features was retained, along with all uncorrelated features. The *rfeControl* and *rfe* functions (R caret package) were used to perform tenfold cross-validation on the 72 remaining features to estimate the minimum number of RFC features required to optimally predict BORSA phenotype (77). The most important features for phenotype classification were identified from the mean decrease in accuracy index as determined over 100 iterations of the *importance* function (R randomForest package) on the decorrelated classifier (78). The top six features as determined by accuracy index were used
to build a sparse model with the following parameters: ntree=5,000, mtry=3, proximity=FALSE, importance=TRUE. These features included 5 amino acid substitutions or protein truncations in PBP2, PBP4, and GdpP, and beta-lactamase inhibitor effect. The sparse model was run over 100 iterations of the 70/30 training/validation dataset split, and model performance was measured through the widely used AUC (Area Under the Receiver-Operator Curve) estimator using the prediction and performance functions (R ROCR package) at default parameters (79). The mean validation AUC value was reported with 95% confidence intervals. The ROC plot was visualized using the predict and roc functions (R pROC package) at default parameters (80). A phylogenetically-informed gene-wide mutation association analysis was performed with hogwash (default parameters, fdr=0.05) on only BORSA and MSSA isolates from our cohort (n=92) (81).

**3.6.10 Statistical Analysis.**

Comparisons of categorical and continuous data were performed by Fisher’s exact test with False Discovery Rate (FDR) correction and Mann-Whitney U test, respectively. All tests were two-tailed, and statistical significance was defined as $p \leq 0.05$ (*) and $p \leq 0.01$ (**).

**3.7 Data Availability**

All isolate assemblies are available on NCBI GenBank under BioProject PRJNA695316. All RFC-relevant code is publicly available at [https://github.com/sanjsawhney/BORSA_RFC](https://github.com/sanjsawhney/BORSA_RFC).
Chapter 3: Comparative genomics of borderline oxacillin-resistant *S. aureus*

3.8 References


Chapter 3: Comparative genomics of borderline oxacillin-resistant *S. aureus*


Chapter 3: Comparative genomics of borderline oxacillin-resistant *S. aureus*


Chapter 3: Comparative genomics of borderline oxacillin-resistant *S. aureus*


51. Greninger AL, Chatterjee SS, Chan LC, Hamilton SM, Chambers HF, Chiu CY. 2016. Whole-Genome Sequencing of Methicillin-Resistant Staphylococcus aureus Resistant to Fifth-
Chapter 3: Comparative genomics of borderline oxacillin-resistant *S. aureus*


Chapter 3: Comparative genomics of borderline oxacillin-resistant *S. aureus*


Chapter 3: Comparative genomics of borderline oxacillin-resistant *S. aureus*


82. Made with Biorender.com.
3.9 Figures

**Figure 3.1: Experimental schematic.** Overview of diagnostic methods employed to interrogate clinical *S. aureus* isolates for oxacillin resistance (82). AMX: amoxicillin; AMC: amoxicillin-clavulanic acid; DD: Disk Diffusion.
Figure 3.2: Comparison of growth on MRSA screening agars. Representative images of MRSA, BORSA, and MSSA isolates grown on Spectra MRSA screening agar (A). Representative images of growth and pigmentation from different screening agars using an MRSA (top) and BORSA (bottom) isolate grown on the same agar plate (B).
Figure 3.3: Beta-lactamase inhibitor effect. Log₂ fold change in MICs between amoxicillin and amoxicillin-clavulanic acid. Lines represent standard deviation, and significance is measured by Mann-Whitney U test.
Chapter 3: Comparative genomics of borderline oxacillin-resistant *S. aureus*
Figure 3.4: BORSA isolates do not cluster by core-genome, accessory-genome, or ARG content.  
A) Core genome alignment of 101 *S. aureus* isolates. Class, MLST, and spa-Type are indicated by color strips, and lineage and clades are indicated by branch color. B) Outgroup-rooted whole genome SNP distance trees of clonal isolates from multiple inpatients with timelines of collection below. Dark blue and sky-blue branch colors represent BORSA and MSSA status, respectively. Cluster 1 isolates 316, 318, and 334 are 10-17 SNPs apart, and 592-608 SNPs apart from nearest phylogenetic neighbor, 061. Cluster 2 isolates 310, 311, 312, and 346 are 6-17 isolates apart, and 394-400 SNPs apart from nearest phylogenetic neighbor, 007. C) Principal coordinate analysis ordination of accessory genome similarity as calculated by Jaccard distance. Axis length is reflective of percent variance captured. D) Isolates clustered by resistance gene content, with MLST and class indicated by color strips. Resistance gene rows are grouped and labeled by antimicrobial class.
Chapter 3: Comparative genomics of borderline oxacillin-resistant *S. aureus*

Figure 3.5: Class Distribution by Core Genome Lineage and Sequence Type (MLST).
Figure 3.6. Ancestral character estimation of borderline oxacillin resistance. Highly probable ancestral states are defined as internal nodes with \( \geq 0.8 \) likelihood of ancestral BORSA status. Shaded clades indicate BORSA isolates stemming from a recent common ancestor expected to be borderline oxacillin resistant.
Figure 3.7: BORSA Mutations by Protein. A) Consensus sequences were generated from multiple sequence alignments of PBP1-4 and GdpP. Amino acid mutations against these consensus sequences are visualized, with isolates clustered by MLST. For each mutation observed only in BORSA isolates within our cohort (“BORSA Exclusive”), the “Unreported” bar indicates whether it has previously been referenced in BORSA-related literature (11-13, 16, 17, 19, 47, 50, 51). B) Venn diagram contrasting BORSA-linked mutations from prior reports (11-13, 16, 17, 19, 47, 50, 51), mutations present in BORSA isolates from this study, and mutations present in MSSA isolates from this study.
Figure 3.8: SNP distance of closely related isolates. Outgroup-rooted whole genome SNP distance trees of clonal isolates collected from the same inpatient on the same sampling date (A, C-E) and from the same inpatient over multiple sampling events (B). Dark blue and sky-blue branch colors represent BORSA and MSSA status, respectively. All isolate pairs are ≤ 10 whole-genome SNPs in distance. B) Isolates 332 and 338 were collected seven days apart.
Chapter 3: Comparative genomics of borderline oxacillin-resistant S. aureus
Figure 3.9: Unrefined RFC built with inclusion of the Argudín et al. (2018) dataset. A) BORSA mutations by protein from isolates in this and the Argudín cohorts. B-C) Prototype RFC preceding removal of highly-correlated features. Data originally included all amino acid substitutions or truncations in the PBP1-4 and GdpP proteins, MLST, $blaZ$ presence, and beta-lactamase inhibitor effect.
Chapter 3: Comparative genomics of borderline oxacillin-resistant *S. aureus*

A) ROC Curve

B) Mean Decrease in Accuracy

C) Confusion Matrix

AUC = 0.902 ± 0.009

Sensitivity

Specificity

1.0

0.8

0.6

0.4

0.2

0.0

1.0

0.8

0.6

0.4

0.2

0.0

0

16

3

0

18

BOSA

MSSA
Figure 3.10: Random Forest Classifier predicts BORSA Phenotype by PBP/GdpP mutational profile with high diagnostic accuracy. A) Receiver operating characteristic (ROC) curves evaluating ability to predict BORSA phenotype using Random Forest classification (RFC). The RFC was run over 100 iterations (denoted in gray) of the validation dataset, each time randomly selected in a 70:30 train/test split, with the mean ROC curve in red. Each ROC curve represents the true-positive and false-positive rates of the supervised machine learning model. B) All remaining features used to train the prototype RFC after removal of highly-correlated features, ranked by importance. The six features used in the final, sparse model (Figure 3.10A) are signified in red. Data are mean ± s.e.m. computed over 100 iterations. C) Confusion matrix for the classification of BORSA status using only these 6 predictors. The sparse model was highly accurate (91.9% classification accuracy).
Figure 3.11: Phylogenetically-informed gene-wide mutation association analysis. **A)** Manhattan plot of all genetic features (substitutions or truncations in PBP1-4 and GdpP) in 92 BORSA and MSSA isolates from this study, plotted against \(-\ln(P\text{-value})\). Features in which the genotype is present or absent in all or all but one isolates, as well as features that do not display 2+ high confidence genotype transition edges are excluded. **B)** Genotype transitions of GdpP premature STOP in red. **C)** Null distribution and observed value of the significant hit “GdpP premature STOP.” All plots generated by Hogwash. Significance threshold set at FDR-corrected \(p\)-value < 0.05.
### 3.10 Tables

**Table 3.1: Phenotypic and genotypic characterization of index isolates, including phenotypic susceptibility testing and evaluation of chromogenic culture medium.**

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Class</th>
<th>Molecular Testing</th>
<th>Susceptibility Testing</th>
<th>MRSA Screening Agars</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PBP2a</td>
<td>mecA</td>
<td>FOX DD, MH</td>
</tr>
<tr>
<td>301</td>
<td>BORSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>302</td>
<td>BORSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>303</td>
<td>MRSA</td>
<td>P</td>
<td>Det</td>
<td>R</td>
</tr>
<tr>
<td>304</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>305</td>
<td>BORSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>306</td>
<td>BORSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>307</td>
<td>BORSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>308</td>
<td>MRSA</td>
<td>P</td>
<td>Det</td>
<td>R</td>
</tr>
<tr>
<td>309</td>
<td>MRSA</td>
<td>P</td>
<td>Det</td>
<td>R</td>
</tr>
<tr>
<td>310</td>
<td>BORSA/MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>311</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>312</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>314</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>315</td>
<td>BORSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>316</td>
<td>BORSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>318</td>
<td>BORSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>320</td>
<td>MRSA</td>
<td>P</td>
<td>Det</td>
<td>R</td>
</tr>
<tr>
<td>321</td>
<td>BORSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>322</td>
<td>MRSA</td>
<td>P</td>
<td>Det</td>
<td>R</td>
</tr>
<tr>
<td>323</td>
<td>BORSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>324</td>
<td>MRSA</td>
<td>P</td>
<td>Det</td>
<td>R</td>
</tr>
<tr>
<td>325</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>327</td>
<td>BORSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>328</td>
<td>BORSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>329</td>
<td>MRSA</td>
<td>P</td>
<td>Det</td>
<td>R</td>
</tr>
<tr>
<td>330</td>
<td>MRSA</td>
<td>P</td>
<td>Det</td>
<td>R</td>
</tr>
<tr>
<td>331</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>332</td>
<td>BORSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>333</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>334</td>
<td>BORSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>335</td>
<td>MRSA</td>
<td>P</td>
<td>Det</td>
<td>S</td>
</tr>
</tbody>
</table>
Chapter 3: Comparative genomics of borderline oxacillin-resistant *S. aureus*

<table>
<thead>
<tr>
<th>Isolate</th>
<th>BORSA</th>
<th>MSSA</th>
<th>FOX</th>
<th>OXA</th>
<th>FOX</th>
<th>OXA</th>
<th>FOX</th>
<th>OXA</th>
</tr>
</thead>
<tbody>
<tr>
<td>336</td>
<td>BORSA</td>
<td>-</td>
<td>S</td>
<td>R</td>
<td>R</td>
<td>+</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>337</td>
<td>BORSA</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>R</td>
<td>***</td>
<td>+</td>
<td>++</td>
</tr>
<tr>
<td>338</td>
<td>BORSA</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>R</td>
<td>+</td>
<td>-</td>
<td>++</td>
</tr>
<tr>
<td>339</td>
<td>BORSA</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>R</td>
<td>***</td>
<td>+++</td>
<td>++</td>
</tr>
<tr>
<td>340</td>
<td>BORSA</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>R</td>
<td>***</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>341</td>
<td>BORSA</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>R</td>
<td>++</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>342</td>
<td>BORSA</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>R</td>
<td>++</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>343</td>
<td>MSSA</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>-</td>
<td>++</td>
<td>-</td>
</tr>
<tr>
<td>344</td>
<td>BORSA</td>
<td>-</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>+++</td>
<td>+++</td>
<td>++</td>
</tr>
<tr>
<td>345</td>
<td>BORSA</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>R</td>
<td>+++</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>346</td>
<td>BORSA</td>
<td>-</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>+++</td>
<td>+++</td>
<td>++</td>
</tr>
</tbody>
</table>

Abbreviations: FOX, cefoxitin; OXA, oxacillin; DD, disk diffusion; GD, gradient diffusion; MH, Mueller-Hinton agar; P, positive; -, negative, not detected, or no growth; Det, detected; S, susceptible; I, intermediate; R, resistant. Plus signs indicate growth abundance: rare (+), few (++), or growth equal to control strain (+++). ^The ads MRSASelect II, BBL CHROMagar MRSA II, and chromID MRSA performed identically. #Isolate shared characteristics consistent with BORSA and MSSA (see Discussion for details).
### Table 3.2: Compiled data on isolates under investigation and comparator blood isolates.

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Class</th>
<th>Patient</th>
<th>PBP 2a</th>
<th>mecA PCR</th>
<th>mecC PCR</th>
<th>Cefoxitin DD, MH (mm)</th>
<th>Oxacillin DD, MH (mm)</th>
<th>Oxacillin, GD 2% NaCl MH (mm)</th>
<th>Pen DD Zone-edge Test</th>
<th>Ce-finase Test</th>
<th>Beta-lactamase Inhibitor Impact (fold)</th>
<th>MRSA Screening Agars</th>
<th>Other 3 Agars*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>I</td>
<td>I</td>
<td>MIC I</td>
<td>I</td>
<td></td>
<td></td>
<td>Spectra MRSA</td>
<td></td>
</tr>
<tr>
<td>301</td>
<td>BORSA</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>24</td>
<td>6</td>
<td>16</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>+++</td>
<td>-</td>
</tr>
<tr>
<td>302</td>
<td>BORSA</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>24</td>
<td>11</td>
<td>1</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>+++</td>
<td>-</td>
</tr>
<tr>
<td>303</td>
<td>MRSA</td>
<td>3</td>
<td>P</td>
<td>Det</td>
<td>-</td>
<td>11</td>
<td>6</td>
<td>&gt;256</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>304</td>
<td>MSSA</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25</td>
<td>21</td>
<td>1</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>+++</td>
<td>-</td>
</tr>
<tr>
<td>305</td>
<td>BORSA</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>24</td>
<td>10</td>
<td>4</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>++</td>
<td>+++</td>
</tr>
<tr>
<td>306</td>
<td>BORSA</td>
<td>5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>24</td>
<td>18</td>
<td>8</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>++</td>
<td>+++</td>
</tr>
<tr>
<td>307</td>
<td>BORSA</td>
<td>5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25</td>
<td>15</td>
<td>4</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>++</td>
<td>-</td>
</tr>
<tr>
<td>308</td>
<td>MRSA</td>
<td>6</td>
<td>P</td>
<td>Det</td>
<td>-</td>
<td>10</td>
<td>6</td>
<td>&gt;256</td>
<td>R</td>
<td>P</td>
<td>-</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>309</td>
<td>MRSA</td>
<td>7</td>
<td>P</td>
<td>Det</td>
<td>-</td>
<td>11</td>
<td>6</td>
<td>&gt;256</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>310</td>
<td>BORSA</td>
<td>8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>27</td>
<td>16</td>
<td>4</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>++</td>
<td>-</td>
</tr>
<tr>
<td>311</td>
<td>MSSA</td>
<td>9</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25</td>
<td>15</td>
<td>2</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>312</td>
<td>MSSA</td>
<td>10</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>24</td>
<td>17</td>
<td>2</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>++</td>
<td>-</td>
</tr>
<tr>
<td>314</td>
<td>MSSA</td>
<td>11</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>27</td>
<td>22</td>
<td>1</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>+++</td>
<td>-</td>
</tr>
<tr>
<td>316</td>
<td>BORSA</td>
<td>12</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25</td>
<td>16</td>
<td>16</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>++</td>
<td>-</td>
</tr>
<tr>
<td>318</td>
<td>BORSA</td>
<td>13</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>27</td>
<td>16</td>
<td>4</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>320</td>
<td>MRSA</td>
<td>15</td>
<td>P</td>
<td>Det</td>
<td>-</td>
<td>10</td>
<td>6</td>
<td>&gt;256</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>321</td>
<td>BORSA</td>
<td>16</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>23</td>
<td>6</td>
<td>16</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>322</td>
<td>MRSA</td>
<td>16</td>
<td>P</td>
<td>Det</td>
<td>-</td>
<td>6</td>
<td>6</td>
<td>&gt;256</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>323</td>
<td>BORSA</td>
<td>17</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>24</td>
<td>13</td>
<td>4</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>+++</td>
<td>++</td>
</tr>
<tr>
<td>324</td>
<td>MRSA</td>
<td>18</td>
<td>P</td>
<td>Det</td>
<td>-</td>
<td>14</td>
<td>10</td>
<td>&gt;256</td>
<td>R</td>
<td>-</td>
<td>-</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>325</td>
<td>MSSA</td>
<td>19</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25</td>
<td>17</td>
<td>2</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>++</td>
<td>-</td>
</tr>
<tr>
<td>327</td>
<td>BORSA</td>
<td>20</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>28</td>
<td>15</td>
<td>8</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>+++</td>
<td>-</td>
</tr>
</tbody>
</table>
## Chapter 3: Comparative genomics of borderline oxacillin-resistant *S. aureus*

<table>
<thead>
<tr>
<th></th>
<th>BORS</th>
<th>MRS</th>
<th>MSS</th>
<th>MRSA</th>
<th>MSS</th>
<th>MRSA</th>
<th>MSS</th>
<th>MRSA</th>
<th>MSS</th>
<th>MRSA</th>
<th>MSS</th>
<th>MRSA</th>
<th>MSS</th>
<th>MRSA</th>
<th>MSS</th>
<th>MRSA</th>
<th>MSS</th>
<th>MRSA</th>
<th>MSS</th>
<th>MRSA</th>
<th>MSS</th>
<th>MRSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>328</td>
<td>BORS</td>
<td>21</td>
<td>-</td>
<td>-</td>
<td>24</td>
<td>11</td>
<td>I</td>
<td>4</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>329</td>
<td>MRS</td>
<td>22</td>
<td>P</td>
<td>Det</td>
<td>-</td>
<td>18</td>
<td>R</td>
<td>8</td>
<td>R</td>
<td>16</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>+++</td>
<td>++</td>
<td>+++</td>
<td>++</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>330</td>
<td>MRS</td>
<td>23</td>
<td>P</td>
<td>Det</td>
<td>-</td>
<td>12</td>
<td>R</td>
<td>6</td>
<td>R</td>
<td>&gt;256</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>4</td>
<td>+++</td>
<td>+</td>
<td>+++</td>
<td>+++</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>331</td>
<td>MSS</td>
<td>24</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>24</td>
<td>S</td>
<td>16</td>
<td>S</td>
<td>2</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>332</td>
<td>BORS</td>
<td>25^</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>24</td>
<td>S</td>
<td>16</td>
<td>S</td>
<td>4</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>333</td>
<td>MSS</td>
<td>26</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>30</td>
<td>S</td>
<td>18</td>
<td>S</td>
<td>2</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>334</td>
<td>BORS</td>
<td>27</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>27</td>
<td>S</td>
<td>15</td>
<td>S</td>
<td>8</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>335</td>
<td>MRS</td>
<td>28</td>
<td>P</td>
<td>Det</td>
<td>-</td>
<td>25</td>
<td>S</td>
<td>12</td>
<td>I</td>
<td>2</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>++</td>
<td>-</td>
<td>+++</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>336</td>
<td>BORS</td>
<td>29</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25</td>
<td>S</td>
<td>10</td>
<td>R</td>
<td>16</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>+</td>
<td>++</td>
<td>++</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>337</td>
<td>BORS</td>
<td>30</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>26</td>
<td>S</td>
<td>14</td>
<td>S</td>
<td>4</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>4</td>
<td>+++</td>
<td>+</td>
<td>++</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>338</td>
<td>BORS</td>
<td>25^</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25</td>
<td>S</td>
<td>17</td>
<td>S</td>
<td>4</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>+</td>
<td>-</td>
<td>++</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>339</td>
<td>BORS</td>
<td>31</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>22</td>
<td>S</td>
<td>16</td>
<td>S</td>
<td>4</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>4</td>
<td>+++</td>
<td>+++</td>
<td>++</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>340</td>
<td>BORS</td>
<td>30</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>23</td>
<td>S</td>
<td>16</td>
<td>S</td>
<td>8</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>+++</td>
<td>++</td>
<td>++</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>341</td>
<td>BORS</td>
<td>32</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>22</td>
<td>S</td>
<td>18</td>
<td>S</td>
<td>8</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>342</td>
<td>BORS</td>
<td>32</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>22</td>
<td>S</td>
<td>18</td>
<td>S</td>
<td>8</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>343</td>
<td>MSS</td>
<td>33</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>22</td>
<td>S</td>
<td>19</td>
<td>S</td>
<td>2</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>4</td>
<td>-</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>344</td>
<td>BORS</td>
<td>33</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>19</td>
<td>R</td>
<td>6</td>
<td>R</td>
<td>32</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>+++</td>
<td>+++</td>
<td>++</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>345</td>
<td>BORS</td>
<td>34</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>22</td>
<td>S</td>
<td>14</td>
<td>S</td>
<td>16</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>+++</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>346</td>
<td>BORS</td>
<td>35</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>19</td>
<td>R</td>
<td>6</td>
<td>R</td>
<td>16</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>+++</td>
<td>+++</td>
<td>++</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>MSS</td>
<td>36</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>27</td>
<td>S</td>
<td>17</td>
<td>S</td>
<td>0.5</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>MSS</td>
<td>37</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25</td>
<td>S</td>
<td>22</td>
<td>S</td>
<td>1</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>MSS</td>
<td>38</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>26</td>
<td>S</td>
<td>21</td>
<td>S</td>
<td>2</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>MSS</td>
<td>39</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>26</td>
<td>S</td>
<td>17</td>
<td>S</td>
<td>0.5</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>MSS</td>
<td>40</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25</td>
<td>S</td>
<td>16</td>
<td>S</td>
<td>1</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>MSS</td>
<td>41</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>23</td>
<td>S</td>
<td>15</td>
<td>S</td>
<td>2</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>MSS</td>
<td>42</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>26</td>
<td>S</td>
<td>19</td>
<td>S</td>
<td>0.5</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>MSS</td>
<td>43</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>24</td>
<td>S</td>
<td>16</td>
<td>S</td>
<td>1</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>MSS</td>
<td>44</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>26</td>
<td>S</td>
<td>14</td>
<td>S</td>
<td>2</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>BORS</td>
<td>45</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>24</td>
<td>S</td>
<td>13</td>
<td>S</td>
<td>4</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>MSS</td>
<td>46</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>26</td>
<td>S</td>
<td>19</td>
<td>S</td>
<td>1</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

111
### Chapter 3: Comparative genomics of borderline oxacillin-resistant *S. aureus*

<table>
<thead>
<tr>
<th></th>
<th>BORSA</th>
<th>MSSA</th>
<th>MSSA</th>
<th>MSSA</th>
<th>MSSA</th>
<th>MSSA</th>
<th>MSSA</th>
<th>MSSA</th>
<th>MSSA</th>
<th>MSSA</th>
<th>MSSA</th>
<th>MSSA</th>
<th>MSSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>47</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>22</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>13</td>
<td>48</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>26</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>14</td>
<td>49</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>27</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>15</td>
<td>50</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>26</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>16</td>
<td>51</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>26</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>17</td>
<td>52</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>26</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>18</td>
<td>53</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>19</td>
<td>54</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>20</td>
<td>55</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>21</td>
<td>56</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>26</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>22</td>
<td>57</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>23</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>23</td>
<td>58</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>26</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>24</td>
<td>59</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>24</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>25</td>
<td>60</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>24</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>26</td>
<td>61</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>27</td>
<td>62</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>27</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>28</td>
<td>63</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>28</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>29</td>
<td>64</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>24</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>30</td>
<td>65</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>24</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>31</td>
<td>66</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>26</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>32</td>
<td>67</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>24</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>33</td>
<td>68</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>27</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>34</td>
<td>69</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>35</td>
<td>70</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>26</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>36</td>
<td>71</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>37</td>
<td>72</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>30</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>38</td>
<td>73</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>27</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>39</td>
<td>74</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>28</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>40</td>
<td>75</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>29</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>41</td>
<td>76</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>22</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>18</td>
</tr>
</tbody>
</table>

112
Chapter 3: Comparative genomics of borderline oxacillin-resistant *S. aureus*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>77</th>
<th>-</th>
<th>-</th>
<th>25</th>
<th>S</th>
<th>16</th>
<th>S</th>
<th>1</th>
<th>S</th>
<th>P</th>
<th>P</th>
<th>2</th>
<th>-</th>
<th>-</th>
<th>-</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>43</td>
<td>MSSA</td>
<td>78</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>27</td>
<td>S</td>
<td>18</td>
<td>S</td>
<td>1</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>44</td>
<td>MSSA</td>
<td>79</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>28</td>
<td>S</td>
<td>17</td>
<td>S</td>
<td>2</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>45</td>
<td>MSSA</td>
<td>80</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>26</td>
<td>S</td>
<td>15</td>
<td>S</td>
<td>2</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>46</td>
<td>MSSA</td>
<td>81</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25</td>
<td>S</td>
<td>16</td>
<td>S</td>
<td>1</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>47</td>
<td>MSSA</td>
<td>82</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>26</td>
<td>S</td>
<td>21</td>
<td>S</td>
<td>2</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>48</td>
<td>MSSA</td>
<td>83</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25</td>
<td>S</td>
<td>14</td>
<td>S</td>
<td>2</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>49</td>
<td>BORSA</td>
<td>84</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25</td>
<td>S</td>
<td>16</td>
<td>S</td>
<td>16</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>50</td>
<td>MSSA</td>
<td>85</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>29</td>
<td>S</td>
<td>17</td>
<td>S</td>
<td>1</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>51</td>
<td>MSSA</td>
<td>86</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>23</td>
<td>S</td>
<td>14</td>
<td>S</td>
<td>2</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>52</td>
<td>MSSA</td>
<td>87</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>23</td>
<td>S</td>
<td>14</td>
<td>S</td>
<td>2</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>53</td>
<td>MSSA</td>
<td>88</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>24</td>
<td>S</td>
<td>17</td>
<td>S</td>
<td>1</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>54</td>
<td>MSSA</td>
<td>89</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>26</td>
<td>S</td>
<td>26</td>
<td>S</td>
<td>0.25</td>
<td>S</td>
<td>-</td>
<td>P</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>55</td>
<td>MSSA</td>
<td>90</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>24</td>
<td>S</td>
<td>19</td>
<td>S</td>
<td>1</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>56</td>
<td>MSSA</td>
<td>91</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25</td>
<td>S</td>
<td>23</td>
<td>S</td>
<td>1</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>57</td>
<td>MSSA</td>
<td>92</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>24</td>
<td>S</td>
<td>18</td>
<td>S</td>
<td>2</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>++</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>58</td>
<td>MSSA</td>
<td>93</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>24</td>
<td>S</td>
<td>16</td>
<td>S</td>
<td>2</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>59</td>
<td>MSSA</td>
<td>94</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>25</td>
<td>S</td>
<td>20</td>
<td>S</td>
<td>2</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>60</td>
<td>MSSA</td>
<td>95</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>24</td>
<td>S</td>
<td>17</td>
<td>S</td>
<td>1</td>
<td>S</td>
<td>P</td>
<td>P</td>
<td>2</td>
<td>++</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 3.3: Phenotypic and genotypic characterization of comparator blood isolates, including phenotypic susceptibility testing and evaluation of chromogenic culture medium.

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Class</th>
<th>Molecular Testing</th>
<th>Susceptibility Testing</th>
<th>MRSA Screening Agars</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PBP2a</td>
<td>mecaPCR</td>
<td>FOX DD, MH</td>
</tr>
<tr>
<td>1</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>2</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>3</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>4</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>5</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>6</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>7</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>8</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>9</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>10</td>
<td>BORSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>11</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>12</td>
<td>BORSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>13</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>14</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>15</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>16</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>17</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>18</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>19</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>20</td>
<td>BORSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>21</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>22</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>23</td>
<td>BORSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>24</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>25</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>26</td>
<td>BORSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>27</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>28</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>29</td>
<td>BORSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>30</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>31</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
<tr>
<td>32</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
</tr>
</tbody>
</table>
### Chapter 3: Comparative genomics of borderline oxacillin-resistant \textit{S. aureus}

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th>S</th>
<th>S</th>
<th>S</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>33</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>34</td>
<td>BORSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>R</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>35</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>36</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>37</td>
<td>BORSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>R</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>38</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>39</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>40</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>41</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>42</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>43</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>44</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>45</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>46</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>47</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>48</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>49</td>
<td>BORSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>R</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>50</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>51</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>52</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>53</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>54</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>55</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>56</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>57</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>58</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>59</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>++</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>60</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>61</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>62</td>
<td>MSSA</td>
<td>-</td>
<td>-</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>++</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Abbreviations: FOX, cefoxitin; OXA, oxacillin; DD, disk diffusion; GD, gradient diffusion; MH, Mueller-Hinton agar; P, positive; -, negative, not detected, or no growth; Det, detected; S, susceptible; I, intermediate; R, resistant. Plus signs indicate growth abundance: rare (+), few (++) or growth equal to control strain (+++). ^The agars MRSASelect II, BBL CHROMagar MRSA II, and chromID MRSA performed identically.
**Table 3.4: Sensitivity and specificity of commercially available MRSA screening agars.**

<table>
<thead>
<tr>
<th></th>
<th>meca-mediated only</th>
<th>+BORSAs Criteria 1*</th>
<th>+BORSAs Criteria 2^</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sens</td>
<td>Spec</td>
<td>Sens</td>
</tr>
<tr>
<td>BBL CHROMagar MRSA II</td>
<td>0.89</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>MRSA Select II</td>
<td>0.89</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>chromID MRSA</td>
<td>0.89</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>Hardy MRSA screen</td>
<td>1</td>
<td>0.86</td>
<td>0.88</td>
</tr>
<tr>
<td>Hardy CHROM MRSA</td>
<td>0.89</td>
<td>0.8</td>
<td>0.88</td>
</tr>
<tr>
<td>Spectra MRSA</td>
<td>1</td>
<td>0.6</td>
<td>1</td>
</tr>
</tbody>
</table>

*Resistant or intermediate to oxacillin by disk diffusion. ^Resistant to oxacillin by gradient diffusion using inducible 2% NaCl agar. There were 102 strains included in this evaluation.
## Table 3.5: Assembly statistics for each sequenced isolate.

<table>
<thead>
<tr>
<th>Assembly</th>
<th>Id</th>
<th>Complete-ness</th>
<th>Contamination</th>
<th>Strain heterogeneity</th>
<th># contigs</th>
<th>Largest contig</th>
<th>Total length</th>
<th>GC (%)</th>
<th>N50</th>
<th>N75</th>
<th>L50</th>
<th>L75</th>
<th># N's per 100 kbp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.25</td>
<td>0</td>
<td>46</td>
<td>318,994</td>
<td>2,811,776</td>
<td>33</td>
<td>175,697</td>
<td>78,744</td>
<td>6</td>
<td>13</td>
<td>10.35</td>
</tr>
<tr>
<td>2</td>
<td>g__Staphylococcus</td>
<td>98.95</td>
<td>0.10</td>
<td>0</td>
<td>19</td>
<td>1,028,528</td>
<td>2,815,296</td>
<td>33</td>
<td>975,452</td>
<td>332,414</td>
<td>2</td>
<td>3</td>
<td>13.78</td>
</tr>
<tr>
<td>3</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>23</td>
<td>713,367</td>
<td>2,654,746</td>
<td>33</td>
<td>314,176</td>
<td>133,824</td>
<td>3</td>
<td>6</td>
<td>18.46</td>
</tr>
<tr>
<td>4</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>25</td>
<td>403,813</td>
<td>2,761,597</td>
<td>33</td>
<td>293,964</td>
<td>169,167</td>
<td>4</td>
<td>7</td>
<td>31.79</td>
</tr>
<tr>
<td>5</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.13</td>
<td>0</td>
<td>20</td>
<td>1,004,201</td>
<td>2,796,406</td>
<td>33</td>
<td>861,117</td>
<td>328,124</td>
<td>2</td>
<td>3</td>
<td>20.67</td>
</tr>
<tr>
<td>6</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>25</td>
<td>548,842</td>
<td>2,682,237</td>
<td>33</td>
<td>410,626</td>
<td>105,417</td>
<td>3</td>
<td>7</td>
<td>36.09</td>
</tr>
<tr>
<td>7</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>23</td>
<td>536,197</td>
<td>2,679,958</td>
<td>33</td>
<td>410,474</td>
<td>136,929</td>
<td>3</td>
<td>6</td>
<td>32.24</td>
</tr>
<tr>
<td>8</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>23</td>
<td>561,524</td>
<td>2,703,471</td>
<td>33</td>
<td>403,753</td>
<td>131,801</td>
<td>3</td>
<td>7</td>
<td>43.46</td>
</tr>
<tr>
<td>9</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.11</td>
<td>0</td>
<td>17</td>
<td>783,112</td>
<td>2,771,902</td>
<td>33</td>
<td>259,214</td>
<td>197,042</td>
<td>3</td>
<td>6</td>
<td>35.21</td>
</tr>
<tr>
<td>10</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.69</td>
<td>0</td>
<td>30</td>
<td>542,115</td>
<td>2,863,792</td>
<td>33</td>
<td>273,510</td>
<td>106,473</td>
<td>4</td>
<td>8</td>
<td>13.65</td>
</tr>
<tr>
<td>11</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.19</td>
<td>50</td>
<td>35</td>
<td>397,430</td>
<td>2,668,436</td>
<td>33</td>
<td>140,899</td>
<td>61,862</td>
<td>6</td>
<td>13</td>
<td>7.46</td>
</tr>
<tr>
<td>12</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.64</td>
<td>0</td>
<td>24</td>
<td>834,827</td>
<td>2,814,929</td>
<td>33</td>
<td>332,185</td>
<td>173,132</td>
<td>3</td>
<td>5</td>
<td>13.82</td>
</tr>
<tr>
<td>13</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>20</td>
<td>756,166</td>
<td>2,771,366</td>
<td>33</td>
<td>311,117</td>
<td>126,750</td>
<td>3</td>
<td>6</td>
<td>38.9</td>
</tr>
<tr>
<td>14</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.13</td>
<td>0</td>
<td>16</td>
<td>987,854</td>
<td>2,720,190</td>
<td>33</td>
<td>497,884</td>
<td>281,703</td>
<td>2</td>
<td>4</td>
<td>28.64</td>
</tr>
<tr>
<td>15</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.64</td>
<td>0</td>
<td>22</td>
<td>986,123</td>
<td>2,850,139</td>
<td>33</td>
<td>820,827</td>
<td>345,457</td>
<td>2</td>
<td>3</td>
<td>13.79</td>
</tr>
<tr>
<td>16</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>26</td>
<td>971,863</td>
<td>2,783,960</td>
<td>33</td>
<td>546,548</td>
<td>293,725</td>
<td>2</td>
<td>4</td>
<td>14.12</td>
</tr>
<tr>
<td>17</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>22</td>
<td>600,595</td>
<td>2,744,822</td>
<td>33</td>
<td>271,203</td>
<td>202,584</td>
<td>4</td>
<td>6</td>
<td>25.03</td>
</tr>
<tr>
<td>18</td>
<td>g__Staphylococcus</td>
<td>99.37</td>
<td>0.22</td>
<td>0</td>
<td>37</td>
<td>497,138</td>
<td>2,826,152</td>
<td>33</td>
<td>155,042</td>
<td>78,704</td>
<td>6</td>
<td>12</td>
<td>13.91</td>
</tr>
<tr>
<td>19</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>30</td>
<td>564,697</td>
<td>2,722,462</td>
<td>33</td>
<td>209,929</td>
<td>134,253</td>
<td>4</td>
<td>8</td>
<td>31.74</td>
</tr>
<tr>
<td>20</td>
<td>g__Staphylococcus</td>
<td>99.23</td>
<td>0.08</td>
<td>0</td>
<td>33</td>
<td>376,004</td>
<td>2,792,492</td>
<td>33</td>
<td>210,835</td>
<td>85,184</td>
<td>5</td>
<td>10</td>
<td>17.58</td>
</tr>
<tr>
<td>21</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.64</td>
<td>0</td>
<td>31</td>
<td>821,239</td>
<td>2,857,438</td>
<td>33</td>
<td>295,823</td>
<td>106,731</td>
<td>3</td>
<td>7</td>
<td>10.11</td>
</tr>
<tr>
<td>22</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>26</td>
<td>562,350</td>
<td>2,699,793</td>
<td>33</td>
<td>411,395</td>
<td>105,417</td>
<td>3</td>
<td>7</td>
<td>39.37</td>
</tr>
<tr>
<td>23</td>
<td>g__Staphylococcus</td>
<td>99.48</td>
<td>0.08</td>
<td>0</td>
<td>30</td>
<td>1,067,405</td>
<td>2,817,064</td>
<td>33</td>
<td>699,993</td>
<td>215,097</td>
<td>2</td>
<td>4</td>
<td>20.8</td>
</tr>
<tr>
<td>24</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>32</td>
<td>355,484</td>
<td>2,707,534</td>
<td>33</td>
<td>208,543</td>
<td>104,310</td>
<td>5</td>
<td>11</td>
<td>10.49</td>
</tr>
</tbody>
</table>
## Chapter 3: Comparative genomics of borderline oxacillin-resistant \textit{S. aureus}

<table>
<thead>
<tr>
<th></th>
<th>g_Staphylococcus</th>
<th>99.51</th>
<th>0.08</th>
<th>0</th>
<th>24</th>
<th>545,657</th>
<th>2,705,073</th>
<th>33</th>
<th>212,729</th>
<th>148,293</th>
<th>5</th>
<th>8</th>
<th>35.86</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>g_Staphylococcus</td>
<td>99.51</td>
<td>0.64</td>
<td>0</td>
<td>21</td>
<td>822,175</td>
<td>2,726,708</td>
<td>33</td>
<td>298,438</td>
<td>203,727</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>26</td>
<td>g_Staphylococcus</td>
<td>99.51</td>
<td>0.64</td>
<td>0</td>
<td>33</td>
<td>987,657</td>
<td>2,867,597</td>
<td>33</td>
<td>313,453</td>
<td>155,052</td>
<td>3</td>
<td>6</td>
<td>10.25</td>
</tr>
<tr>
<td>27</td>
<td>g_Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>29</td>
<td>701,151</td>
<td>2,774,616</td>
<td>33</td>
<td>316,238</td>
<td>88,284</td>
<td>3</td>
<td>9</td>
<td>17.52</td>
</tr>
<tr>
<td>28</td>
<td>g_Staphylococcus</td>
<td>99.51</td>
<td>0.66</td>
<td>40</td>
<td>60</td>
<td>382,792</td>
<td>2,839,650</td>
<td>33</td>
<td>220,128</td>
<td>95,828</td>
<td>5</td>
<td>9</td>
<td>27.57</td>
</tr>
<tr>
<td>29</td>
<td>g_Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>24</td>
<td>448,899</td>
<td>2,695,748</td>
<td>33</td>
<td>291,054</td>
<td>179,127</td>
<td>4</td>
<td>8</td>
<td>18.03</td>
</tr>
<tr>
<td>30</td>
<td>g_Staphylococcus</td>
<td>99.51</td>
<td>0.13</td>
<td>0</td>
<td>24</td>
<td>1,300,055</td>
<td>2,753,737</td>
<td>33</td>
<td>601,983</td>
<td>427,318</td>
<td>2</td>
<td>3</td>
<td>38.53</td>
</tr>
<tr>
<td>31</td>
<td>g_Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>24</td>
<td>514,205</td>
<td>2,713,165</td>
<td>33</td>
<td>420,697</td>
<td>257,823</td>
<td>3</td>
<td>5</td>
<td>28.79</td>
</tr>
<tr>
<td>32</td>
<td>g_Staphylococcus</td>
<td>99.51</td>
<td>0.22</td>
<td>0</td>
<td>44</td>
<td>305,046</td>
<td>2,799,339</td>
<td>33</td>
<td>150,804</td>
<td>61,520</td>
<td>8</td>
<td>15</td>
<td>13.82</td>
</tr>
<tr>
<td>33</td>
<td>g_Staphylococcus</td>
<td>99.51</td>
<td>0.11</td>
<td>0</td>
<td>28</td>
<td>353,579</td>
<td>2,717,258</td>
<td>33</td>
<td>255,954</td>
<td>140,948</td>
<td>5</td>
<td>8</td>
<td>25.06</td>
</tr>
<tr>
<td>34</td>
<td>g_Staphylococcus</td>
<td>99.51</td>
<td>0.1</td>
<td>0</td>
<td>27</td>
<td>589,810</td>
<td>2,732,785</td>
<td>33</td>
<td>382,857</td>
<td>84,371</td>
<td>3</td>
<td>7</td>
<td>21.59</td>
</tr>
<tr>
<td>35</td>
<td>g_Staphylococcus</td>
<td>99.51</td>
<td>0.1</td>
<td>0</td>
<td>27</td>
<td>638,211</td>
<td>2,806,135</td>
<td>33</td>
<td>591,289</td>
<td>159,366</td>
<td>3</td>
<td>5</td>
<td>20.7</td>
</tr>
<tr>
<td>36</td>
<td>g_Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>28</td>
<td>536,844</td>
<td>2,681,076</td>
<td>33</td>
<td>296,685</td>
<td>129,753</td>
<td>4</td>
<td>7</td>
<td>7.12</td>
</tr>
<tr>
<td>37</td>
<td>g_Staphylococcus</td>
<td>99.51</td>
<td>0.22</td>
<td>0</td>
<td>36</td>
<td>484,733</td>
<td>2,794,191</td>
<td>33</td>
<td>174,963</td>
<td>67,681</td>
<td>5</td>
<td>11</td>
<td>10.52</td>
</tr>
<tr>
<td>38</td>
<td>g_Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>24</td>
<td>570,432</td>
<td>2,689,813</td>
<td>33</td>
<td>250,364</td>
<td>118,569</td>
<td>4</td>
<td>8</td>
<td>21.93</td>
</tr>
<tr>
<td>39</td>
<td>g_Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>22</td>
<td>795,063</td>
<td>2,709,583</td>
<td>33</td>
<td>323,776</td>
<td>151,587</td>
<td>3</td>
<td>7</td>
<td>17.94</td>
</tr>
<tr>
<td>40</td>
<td>g_Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>33</td>
<td>463,193</td>
<td>2,721,114</td>
<td>33</td>
<td>158,710</td>
<td>100,885</td>
<td>4</td>
<td>9</td>
<td>35.83</td>
</tr>
<tr>
<td>41</td>
<td>g_Staphylococcus</td>
<td>99.4</td>
<td>0.13</td>
<td>0</td>
<td>17</td>
<td>623,326</td>
<td>2,719,010</td>
<td>33</td>
<td>385,583</td>
<td>199,021</td>
<td>3</td>
<td>5</td>
<td>21.7</td>
</tr>
<tr>
<td>42</td>
<td>g_Staphylococcus</td>
<td>99.51</td>
<td>0.22</td>
<td>0</td>
<td>37</td>
<td>303,985</td>
<td>2,694,056</td>
<td>33</td>
<td>135,743</td>
<td>98,225</td>
<td>7</td>
<td>12</td>
<td>32.85</td>
</tr>
<tr>
<td>43</td>
<td>g_Staphylococcus</td>
<td>99.51</td>
<td>0.64</td>
<td>0</td>
<td>31</td>
<td>956,930</td>
<td>2,888,059</td>
<td>33</td>
<td>547,381</td>
<td>109,381</td>
<td>2</td>
<td>5</td>
<td>6.72</td>
</tr>
<tr>
<td>44</td>
<td>g_Staphylococcus</td>
<td>98.95</td>
<td>0.25</td>
<td>0</td>
<td>46</td>
<td>380,086</td>
<td>2,805,602</td>
<td>33</td>
<td>169,714</td>
<td>76,933</td>
<td>6</td>
<td>12</td>
<td>27.41</td>
</tr>
<tr>
<td>45</td>
<td>g_Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>27</td>
<td>471,603</td>
<td>2,764,961</td>
<td>33</td>
<td>187,299</td>
<td>120,462</td>
<td>5</td>
<td>9</td>
<td>21.16</td>
</tr>
<tr>
<td>46</td>
<td>g_Staphylococcus</td>
<td>99.45</td>
<td>0.08</td>
<td>0</td>
<td>15</td>
<td>805,942</td>
<td>2,757,238</td>
<td>33</td>
<td>728,605</td>
<td>229,130</td>
<td>2</td>
<td>4</td>
<td>21.18</td>
</tr>
<tr>
<td>47</td>
<td>g_Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>12</td>
<td>1,342,337</td>
<td>2,699,319</td>
<td>33</td>
<td>358,605</td>
<td>328,586</td>
<td>2</td>
<td>3</td>
<td>18.04</td>
</tr>
<tr>
<td>48</td>
<td>g_Staphylococcus</td>
<td>99.51</td>
<td>0.22</td>
<td>0</td>
<td>39</td>
<td>406,726</td>
<td>2,849,670</td>
<td>33</td>
<td>169,490</td>
<td>85,069</td>
<td>6</td>
<td>11</td>
<td>6.91</td>
</tr>
<tr>
<td>49</td>
<td>g_Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>25</td>
<td>564,656</td>
<td>2,696,533</td>
<td>33</td>
<td>410,568</td>
<td>131,753</td>
<td>3</td>
<td>6</td>
<td>32.41</td>
</tr>
</tbody>
</table>
Chapter 3: Comparative genomics of borderline oxacillin-resistant *S. aureus*

| 54 | g__Staphylococcus | 99.51 | 0.08 | 0 | 17 | 967,666 | 2,694,348 | 33 | 448,082 | 257,912 | 2 | 4 | 32.66 |
| 55 | g__Staphylococcus | 99.51 | 0.08 | 0 | 29 | 455,136 | 2,744,755 | 33 | 310,721 | 129,766 | 4 | 8 | 10.42 |
| 56 | g__Staphylococcus | 99.41 | 0.08 | 0 | 17 | 570,962 | 2,757,369 | 33 | 429,058 | 158,147 | 3 | 6 | 24.26 |
| 57 | g__Staphylococcus | 99.51 | 0.08 | 0 | 22 | 696,170 | 2,709,032 | 33 | 232,825 | 167,248 | 4 | 7 | 17.9 |
| 58 | g__Staphylococcus | 99.51 | 0.08 | 0 | 21 | 472,413 | 2,702,779 | 33 | 244,104 | 187,573 | 4 | 7 | 21.72 |
| 59 | g__Staphylococcus | 99.51 | 0.11 | 0 | 34 | 417,998 | 2,818,432 | 33 | 210,480 | 111,046 | 5 | 10 | 24.09 |
| 60 | g__Staphylococcus | 99.51 | 0.08 | 0 | 32 | 379,880 | 2,767,595 | 33 | 197,331 | 107,402 | 5 | 9 | 28.36 |
| 61 | g__Staphylococcus | 99.51 | 0.08 | 0 | 25 | 354,306 | 2,663,958 | 33 | 249,996 | 125,726 | 5 | 9 | 18.32 |
| 62 | g__Staphylococcus | 99.51 | 0.08 | 0 | 28 | 562,972 | 2,704,235 | 33 | 408,764 | 116,684 | 3 | 6 | 25.29 |
| 301 | g__Staphylococcus | 99.45 | 0.11 | 0 | 27 | 324,859 | 2,743,547 | 33 | 262,075 | 196,922 | 5 | 8 | 60.4 |
| 302 | g__Staphylococcus | 99.51 | 0.08 | 0 | 16 | 654,378 | 2,731,642 | 33 | 386,964 | 225,176 | 3 | 5 | 17.68 |
| 303 | g__Staphylococcus | 99.51 | 0.64 | 0 | 25 | 835,012 | 2,812,789 | 33 | 588,887 | 172,816 | 2 | 5 | 27.66 |
| 304 | g__Staphylococcus | 99.41 | 0.11 | 0 | 18 | 437,288 | 2,760,376 | 33 | 388,306 | 182,830 | 4 | 6 | 31.92 |
| 305 | g__Staphylococcus | 99.45 | 0.13 | 0 | 16 | 1,278,060 | 2,692,018 | 33 | 902,921 | 902,921 | 2 | 2 | 21.81 |
| 306 | g__Staphylococcus | 99.51 | 0.08 | 0 | 17 | 792,787 | 2,728,880 | 33 | 286,563 | 150,086 | 3 | 6 | 14.36 |
| 307 | g__Staphylococcus | 99.51 | 0.08 | 0 | 20 | 792,787 | 2,729,170 | 33 | 286,679 | 151,552 | 3 | 6 | 10.7 |
| 308 | g__Staphylococcus | 99.32 | 0.08 | 0 | 21 | 1,063,139 | 2,778,168 | 33 | 382,810 | 207,172 | 5 | 2 | 31.1 |
| 309 | g__Staphylococcus | 99.51 | 0.1 | 0 | 26 | 385,786 | 2,795,011 | 33 | 316,103 | 202,040 | 4 | 7 | 24.4 |
| 310 | g__Staphylococcus | 99.51 | 0.08 | 0 | 34 | 562,655 | 2,761,463 | 33 | 236,285 | 117,778 | 4 | 9 | 31.43 |
| 311 | g__Staphylococcus | 99.51 | 0.08 | 0 | 47 | 562,655 | 2,799,508 | 33 | 197,159 | 92,436 | 5 | 10 | 27.36 |
| 312 | g__Staphylococcus | 99.51 | 0.08 | 0 | 35 | 562,655 | 2,761,442 | 33 | 236,285 | 118,249 | 4 | 9 | 28.21 |
| 314 | g__Staphylococcus | 99.51 | 0.22 | 0 | 29 | 389,396 | 2,719,976 | 33 | 170,008 | 98,229 | 5 | 10 | 32.1 |
| 315 | g__Staphylococcus | 99.51 | 0.08 | 0 | 16 | 760,417 | 2,700,194 | 33 | 501,762 | 354,227 | 3 | 4 | 36.07 |
| 316 | g__Staphylococcus | 99.51 | 0.08 | 0 | 35 | 436,681 | 2,696,966 | 33 | 215,835 | 110,841 | 5 | 9 | 21.69 |
| 318 | g__Staphylococcus | 99.51 | 0.08 | 0 | 37 | 436,681 | 2,696,283 | 33 | 215,835 | 110,841 | 5 | 9 | 21.81 |
| 320 | g__Staphylococcus | 99.51 | 0.1 | 0 | 24 | 765,770 | 2,751,292 | 33 | 316,102 | 202,040 | 3 | 6 | 28.46 |
Chapter 3: Comparative genomics of borderline oxacillin-resistant *S. aureus*

<table>
<thead>
<tr>
<th></th>
<th>g__Staphylococcus</th>
<th>99.51</th>
<th>0.08</th>
<th>0</th>
<th>27</th>
<th>560,554</th>
<th>2,734,724</th>
<th>33</th>
<th>410,626</th>
<th>134,105</th>
<th>3</th>
<th>7</th>
<th>42.71</th>
</tr>
</thead>
<tbody>
<tr>
<td>322</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>26</td>
<td>469,413</td>
<td>2,744,806</td>
<td>33</td>
<td>263,568</td>
<td>151,219</td>
<td>5</td>
<td>8</td>
<td>14.14</td>
</tr>
<tr>
<td>323</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.13</td>
<td>0</td>
<td>31</td>
<td>383,281</td>
<td>2,771,397</td>
<td>33</td>
<td>200,699</td>
<td>131,233</td>
<td>5</td>
<td>9</td>
<td>24.07</td>
</tr>
<tr>
<td>324</td>
<td>g__Staphylococcus</td>
<td>99.23</td>
<td>0.64</td>
<td>0</td>
<td>22</td>
<td>980,743</td>
<td>2,855,233</td>
<td>33</td>
<td>834,636</td>
<td>345,459</td>
<td>2</td>
<td>3</td>
<td>16.95</td>
</tr>
<tr>
<td>325</td>
<td>g__Staphylococcus</td>
<td>99.41</td>
<td>0.08</td>
<td>0</td>
<td>26</td>
<td>333,918</td>
<td>2,785,978</td>
<td>33</td>
<td>256,509</td>
<td>170,214</td>
<td>5</td>
<td>8</td>
<td>20.93</td>
</tr>
<tr>
<td>327</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>18</td>
<td>599,396</td>
<td>2,694,783</td>
<td>33</td>
<td>284,797</td>
<td>159,882</td>
<td>4</td>
<td>7</td>
<td>14.47</td>
</tr>
<tr>
<td>328</td>
<td>g__Staphylococcus</td>
<td>99.42</td>
<td>0.08</td>
<td>0</td>
<td>25</td>
<td>560,550</td>
<td>2,676,172</td>
<td>33</td>
<td>410,628</td>
<td>134,528</td>
<td>3</td>
<td>7</td>
<td>40.17</td>
</tr>
<tr>
<td>329</td>
<td>g__Staphylococcus</td>
<td>99.4</td>
<td>0.27</td>
<td>0</td>
<td>66</td>
<td>417,935</td>
<td>2,890,944</td>
<td>33</td>
<td>108,969</td>
<td>60,679</td>
<td>7</td>
<td>16</td>
<td>10</td>
</tr>
<tr>
<td>330</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.64</td>
<td>0</td>
<td>39</td>
<td>426,040</td>
<td>2,899,439</td>
<td>33</td>
<td>238,406</td>
<td>105,779</td>
<td>5</td>
<td>9</td>
<td>23.35</td>
</tr>
<tr>
<td>331</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>22</td>
<td>570,246</td>
<td>2,735,906</td>
<td>33</td>
<td>249,432</td>
<td>190,751</td>
<td>4</td>
<td>7</td>
<td>17.84</td>
</tr>
<tr>
<td>332</td>
<td>g__Staphylococcus</td>
<td>99.48</td>
<td>0.08</td>
<td>0</td>
<td>19</td>
<td>1,368,543</td>
<td>2,764,855</td>
<td>33</td>
<td>701,096</td>
<td>169,878</td>
<td>2</td>
<td>3</td>
<td>17.32</td>
</tr>
<tr>
<td>333</td>
<td>g__Staphylococcus</td>
<td>99.46</td>
<td>0.64</td>
<td>0</td>
<td>30</td>
<td>860,895</td>
<td>2,849,735</td>
<td>33</td>
<td>657,756</td>
<td>106,414</td>
<td>2</td>
<td>6</td>
<td>10.25</td>
</tr>
<tr>
<td>334</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>36</td>
<td>482,740</td>
<td>2,742,445</td>
<td>33</td>
<td>215,835</td>
<td>110,996</td>
<td>5</td>
<td>9</td>
<td>17.61</td>
</tr>
<tr>
<td>335</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>26</td>
<td>322,032</td>
<td>2,764,895</td>
<td>33</td>
<td>247,054</td>
<td>125,652</td>
<td>5</td>
<td>9</td>
<td>20.76</td>
</tr>
<tr>
<td>336</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.64</td>
<td>0</td>
<td>24</td>
<td>878,154</td>
<td>2,842,210</td>
<td>33</td>
<td>657,606</td>
<td>295,835</td>
<td>2</td>
<td>4</td>
<td>10.2</td>
</tr>
<tr>
<td>337</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.11</td>
<td>0</td>
<td>16</td>
<td>1,306,960</td>
<td>2,753,907</td>
<td>33</td>
<td>325,151</td>
<td>170,745</td>
<td>2</td>
<td>4</td>
<td>17.83</td>
</tr>
<tr>
<td>338</td>
<td>g__Staphylococcus</td>
<td>99.48</td>
<td>0.08</td>
<td>0</td>
<td>31</td>
<td>540,669</td>
<td>2,763,637</td>
<td>33</td>
<td>222,575</td>
<td>108,432</td>
<td>4</td>
<td>9</td>
<td>17.51</td>
</tr>
<tr>
<td>339</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>21</td>
<td>584,480</td>
<td>2,716,553</td>
<td>33</td>
<td>456,298</td>
<td>136,983</td>
<td>3</td>
<td>5</td>
<td>17.56</td>
</tr>
<tr>
<td>340</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.11</td>
<td>0</td>
<td>13</td>
<td>1,307,888</td>
<td>2,754,593</td>
<td>33</td>
<td>616,827</td>
<td>170,584</td>
<td>2</td>
<td>3</td>
<td>17.75</td>
</tr>
<tr>
<td>341</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>18</td>
<td>795,635</td>
<td>2,729,049</td>
<td>33</td>
<td>396,720</td>
<td>167,208</td>
<td>3</td>
<td>6</td>
<td>7.26</td>
</tr>
<tr>
<td>342</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>20</td>
<td>795,635</td>
<td>2,729,273</td>
<td>33</td>
<td>229,732</td>
<td>151,588</td>
<td>3</td>
<td>7</td>
<td>7.14</td>
</tr>
<tr>
<td>343</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>28</td>
<td>549,371</td>
<td>2,677,669</td>
<td>33</td>
<td>410,570</td>
<td>107,093</td>
<td>3</td>
<td>7</td>
<td>32.53</td>
</tr>
<tr>
<td>344</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.08</td>
<td>0</td>
<td>26</td>
<td>557,137</td>
<td>2,677,476</td>
<td>33</td>
<td>410,570</td>
<td>118,845</td>
<td>3</td>
<td>6</td>
<td>28.76</td>
</tr>
<tr>
<td>345</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.11</td>
<td>0</td>
<td>19</td>
<td>570,986</td>
<td>2,710,451</td>
<td>33</td>
<td>399,244</td>
<td>249,935</td>
<td>3</td>
<td>6</td>
<td>17.75</td>
</tr>
<tr>
<td>346</td>
<td>g__Staphylococcus</td>
<td>99.51</td>
<td>0.15</td>
<td>0</td>
<td>34</td>
<td>562,655</td>
<td>2,760,536</td>
<td>33</td>
<td>236,285</td>
<td>117,778</td>
<td>4</td>
<td>9</td>
<td>24.67</td>
</tr>
</tbody>
</table>
Table 3.6: List of amino acid (AA) substitutions by BORSA isolate.

<table>
<thead>
<tr>
<th>MLST type</th>
<th>BORSA isolate</th>
<th>PBP1</th>
<th>PBP2</th>
<th>PBP3</th>
<th>PBP4</th>
<th>GdpP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>D118N, V617M</td>
<td>-</td>
<td>G167R</td>
<td>T189S</td>
<td></td>
<td>I52V</td>
</tr>
<tr>
<td>37</td>
<td>K67R, D118N</td>
<td>-</td>
<td>-</td>
<td>T189S</td>
<td></td>
<td>T104I, Y168F</td>
</tr>
<tr>
<td>323</td>
<td>D118N, V617M</td>
<td>-</td>
<td>G167R</td>
<td>T189S</td>
<td></td>
<td>I52V</td>
</tr>
<tr>
<td>332</td>
<td>D118N, V617M</td>
<td>-</td>
<td>G167R</td>
<td>T189S</td>
<td></td>
<td>I52V</td>
</tr>
<tr>
<td>338</td>
<td>D118N, V617M</td>
<td>-</td>
<td>G167R</td>
<td>T189S</td>
<td></td>
<td>I52V</td>
</tr>
<tr>
<td>ST5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ST15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>D118N</td>
<td>-</td>
<td>S438T</td>
<td>T189S</td>
<td></td>
<td></td>
</tr>
<tr>
<td>316</td>
<td>D118N</td>
<td>-</td>
<td>S438T</td>
<td>T189S</td>
<td></td>
<td></td>
</tr>
<tr>
<td>318</td>
<td>D118N</td>
<td>-</td>
<td>S438T</td>
<td>T189S</td>
<td></td>
<td></td>
</tr>
<tr>
<td>334</td>
<td>D118N</td>
<td>-</td>
<td>S438T</td>
<td>T189S</td>
<td></td>
<td></td>
</tr>
<tr>
<td>345</td>
<td>D118N</td>
<td>-</td>
<td>S438T</td>
<td>T189S</td>
<td></td>
<td>S196W, H558Y</td>
</tr>
<tr>
<td>ST45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ST97</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Chapter 3: Comparative genomics of borderline oxacillin-resistant *S. aureus*

<table>
<thead>
<tr>
<th></th>
<th>342</th>
<th>D118N</th>
<th>A315E, S576A</th>
<th>S438T</th>
<th>C12F, T101R, H214C, E398A</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12</td>
<td>D118N</td>
<td>A285P</td>
<td>V30I, S438T, D683N</td>
<td>A25T, A409T</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>D118N</td>
<td>A285P</td>
<td>V30I, S438T, D683N</td>
<td>A25T, A409T</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>336</td>
<td>D118N</td>
<td>A285P</td>
<td>V30I, S438T, D683N</td>
<td>A25T, A409T</td>
<td>-</td>
</tr>
<tr>
<td><strong>ST27</strong></td>
<td>49</td>
<td>-</td>
<td>K209R, A285P</td>
<td>S438T</td>
<td>A25T, T189S</td>
<td>I52V</td>
</tr>
<tr>
<td></td>
<td>315</td>
<td>-</td>
<td>A285P</td>
<td>K138Q</td>
<td>T189S, Q383K</td>
<td>-</td>
</tr>
<tr>
<td><strong>ST72</strong></td>
<td>327</td>
<td>D118N, T146A</td>
<td>A285P, A315E, R379L</td>
<td>S386F</td>
<td>V202E, P253T, E398A</td>
<td>-</td>
</tr>
</tbody>
</table>

AA substitutions in the transglycosylase and transpeptidase domain of the PBP proteins are in italics and bold, respectively. Stop codons and deletions are colored red.
Table 3.7: Bivariate association of the top Random Forest Classifier features with BORSA and MSSA status.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Present in BORSA</th>
<th>Absent in BORSA</th>
<th>Present in MSSA</th>
<th>Absent in MSSA</th>
<th>Frequency in BORSA</th>
<th>Frequency in MSSA</th>
<th>p-value</th>
<th>FDR-corrected p-value</th>
<th>BORSA correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A285P</td>
<td>12</td>
<td>49</td>
<td>27</td>
<td>32</td>
<td>19.7%</td>
<td>45.8%</td>
<td>3.27E-03</td>
<td>5.45E-03</td>
<td>-</td>
</tr>
<tr>
<td>PBP2 End Truncation</td>
<td>0</td>
<td>61</td>
<td>6</td>
<td>53</td>
<td>0.0%</td>
<td>10.2%</td>
<td>1.23E-02</td>
<td>1.23E-02</td>
<td>-</td>
</tr>
<tr>
<td>I52V</td>
<td>7</td>
<td>54</td>
<td>24</td>
<td>35</td>
<td>11.5%</td>
<td>40.7%</td>
<td>3.25E-04</td>
<td>8.13E-04</td>
<td>-</td>
</tr>
<tr>
<td>GdpP Premature STOP</td>
<td>16</td>
<td>45</td>
<td>0</td>
<td>59</td>
<td>26.2%</td>
<td>0.0%</td>
<td>1.01E-05</td>
<td>5.03E-05</td>
<td>+</td>
</tr>
<tr>
<td>T189S</td>
<td>14</td>
<td>47</td>
<td>28</td>
<td>31</td>
<td>23.0%</td>
<td>47.5%</td>
<td>7.11E-03</td>
<td>8.89E-03</td>
<td>-</td>
</tr>
</tbody>
</table>

Significance determined via Fisher’s Exact test with False Discovery Rate (FDR) correction. Feature correlation and anti-correlation with BORSA phenotype is noted.
Chapter 4: Comparative genomics of diagnostic and commensal Staphylococcus pseudintermedius reveals niche adaptation through parallel selection of defense mechanisms

The contents of this chapter are adapted from a manuscript under revision at *Nature Communications*:


* = equal contribution
4.1 Abstract

*Staphylococcus pseudintermedius* is historically understood as a prevalent commensal and pathogen of dogs, though modern clinical diagnostics reveal an expanded host-range that includes humans. It remains unclear whether differentiation across *S. pseudintermedius* populations is driven primarily by niche-type or host-species. We sequenced 572 diagnostic and commensal isolates from a hospital, veterinary diagnostic laboratory, and within households in the American Midwest, and performed a comparative genomics investigation contrasting human diagnostic, animal diagnostic, human colonizing, pet colonizing, and household-surface *S. pseudintermedius* isolates. Though indistinguishable by core and accessory genome architecture, diagnostic isolates harbor more encoded and phenotypic resistance, whereas colonizing and surface isolates harbor CRISPR defense systems likely reflective of common household phage exposures. Furthermore, household isolates that persist through anti-staphylococcal decolonization accrue mutations in defense genes associated with reduced oxacillin and trimethoprim-sulfamethoxazole susceptibility. Together we report parallel niche-specific bolstering of *S. pseudintermedius* defense mechanisms through gene acquisition or mutation.
4.2 Introduction

The *Staphylococcus intermedius* group (SIG) consists of closely related coagulase-positive staphylococcal (CoPS) species *S. pseudintermedius*, *S. intermedius*, *S. delphini*, *S. cornubiensis*, and *S. ursi*, which together have long been characterized for their wide host range of domesticated and wild animal species (1, 2). *S. pseudintermedius*, the most frequently recovered member from clinical specimens, exists commensally on most dogs yet is also the most common etiologic agent of canine pyoderma, ear, postoperative, and urinary tract infections (1-4). Biochemically, SIG species appear remarkably similar to the archetype human CoPS pathogen, *S. aureus*, likely contributing to a historic underreporting of SIG as a contributor to human infection (5-9). Indeed, recent advancements in laboratory typing methods, including repetitive sequence PCR, MALDI-TOF mass spectrometry, and whole genome sequencing (WGS), have revealed *S. pseudintermedius* as an opportunistic pathogen found in human wounds, as well as respiratory, skin and soft tissue, and bloodstream infections (9-12). Though the earliest reports of human *S. pseudintermedius* infections were linked with exposure to companion animals (8, 13), larger retrospective studies have reported a history of animal contact in just 6-10% of human *S. pseudintermedius* infections (9, 11), indicating the clinical context of *S. pseudintermedius* disease in humans extends beyond direct zoonotic transmission. Prevalence of methicillin resistance – which is very closely tied to multidrug resistance – among *S. pseudintermedius* canine infections has increased dramatically from < 5% in 2001 to 20-45% by 2017 (14-17). Though methicillin resistant *S. pseudintermedius* (MRSP) prevalence among human infections remains largely unknown, early reports indicate a congruence with methicillin-resistant *Staphylococcus*
* aureus* (MRSA) rates among Western populations (2, 9, 12, 18, 19), signaling a burgeoning, serious, public health concern.

WGS has enabled comparative genomics studies that have provided foundational knowledge of *S. pseudintermedius* populations and pathologies (20, 21). Similar to *S. aureus*, *S. pseudintermedius* has an open pangenome; most accessory genes are lineage- or strain-specific (20-23), and antibiotic resistance gene (ARG) content and virulence factor repertoires vary by multilocus sequence type (MLST) (20, 24). Additional studies have characterized *S. pseudintermedius* genome content by geographic isolation (21), host-species for clinical isolates (22), niche-type for veterinary isolates (20, 25), and *mecA* presence (26). These reports have identified circulation of phylogenetically-diverse lineages across companion animals (21) as well as differences in genome size between MRSP and MSSP (25, 26). However, these efforts were unable to identify genes that tie diagnostic *S. pseudintermedius* populations to canine or human hosts (22), nor did they observe differences in virulence factor repertoire between clinical and colonizing veterinary isolates, blurring preconceptions of host-species boundaries for diagnostic isolates and suggesting that commensal veterinary *S. pseudintermedius* have pathogenic potential (25, 27). As there have been no studies contrasting isolates across both host-species (human vs. companion animals) and niche-type (diagnostic vs. colonizing), the relative influence of these two modalities on *S. pseudintermedius* population genomics remains unknown. Incorporation of cohorts from multiple host-species and niche-types is needed to determine the major contributing factors towards differentiation among *S. pseudintermedius* populations, and how such differentiation impacts both spread between companion animals and humans, as well as transitions between opportunistic pathogen and commensal lifestyles.
Households present individualized case studies within which *S. pseudintermedius* strain persistence, transmission, and replacement within a microenvironment can be observed. Prior studies reveal that humans and companion animals in a shared household have the potential to introduce *S. pseudintermedius* from outside sources, circulate strains amongst household members, and contribute to host reinfection following clearance of prior MRSP infection (28-31). Additionally, a 2011 observational study investigating the prevalence of MRSP in households with an index MRSP-positive companion animal indicated that environmental contamination was also widespread, finding sleeping, feeding, and lounging areas to also harbor *S. pseudintermedius* (28). Further, a longitudinal study in the same year found that MRSP can be identified on these household surfaces without ongoing MRSP colonization or infection of a human or companion animal, indicating such abiotic surfaces may act as reservoirs that enable *S. pseudintermedius* persistence (32). Recent investigations into *S. pseudintermedius* in households have supported these findings but suggest sampling from multiple sites over time is needed to better understand within-host diversity and within-household transmission of *S. pseudintermedius* (29). Investigating the propagation of *S. pseudintermedius* in the shared biotic and abiotic environment is a key element to understanding *S. pseudintermedius* as an opportunistic, multi-host pathogen (33).

In this study, we unite three cohorts of SIG isolates captured at a major tertiary care medical center and veterinary diagnostic laboratory, both in the American Midwest, and the Household Observation of MRSA in the Environment 2 (HOME2) and Staph Household Intervention for Eradication (SHINE) studies (34, 35) that longitudinally sampled inhabitants and surfaces within households before, during, and after MRSA decolonization efforts in the St. Louis, Missouri, metropolitan area. From these collection sites we have acquired, sequenced,
and performed antibiotic susceptibility testing (AST) on over 500 human diagnostic, animal diagnostic, and household (human and pet colonizing, and environmental) isolates to investigate genotypic and phenotypic differences across host-species and niche-type. While whole genome architecture does not track with host-species or niche, we observe greater encoded and phenotypic resistance among diagnostic isolates. In contrast, we find household isolates harbor similar CRISPR systems which likely reflect evolutionary phage predation agnostic of overall genome similarity, and find that household isolates persisting through decolonization interventions accrue non-synonymous mutations in defense genes that may result in reductions in drug susceptibility. Together we report simultaneous niche-dependent bolstering of SIG defense mechanisms through gene acquisition or mutation.

4.3 Results

4.3.1 Isolate collection and characterization from three unique cohorts.

Given the history of misidentification of SIG genomes amidst coagulase-positive Staphylococci (8, 10), our first aim was to validate MALDI-TOF MS calls against whole-genome sequencing. Towards this, we collected all 572 isolates reported as "Staphylococcus intermedius/pseudintermedius" (n=565) or “S. delphini” (n=7) by MALDI-TOF MS at the Barnes-Jewish clinical microbiology laboratory between December 2011 and July 2019. These isolates were originally captured at the Barnes-Jewish Hospital in St. Louis, MO (n=181), Kansas State University Veterinary Diagnostic Laboratory in Manhattan, KS (n=100), and within households in the St. Louis metropolitan area (n=290) (Figure 4.1) (36). After Illumina sequencing and filtering by genome quality, we retained 501 isolate assemblies (n=163 human diagnostic isolates, n=94 veterinary diagnostic isolates, and n=244
Chapter 4: Comparative genomics of diagnostic and commensal *S. pseudintermedius*

household isolates) that were reported by CheckM and QUAST to be ≥ 98% complete, ≤ 2% contaminated, 0% strain heterogeneity, N50 > 40 Kbp, and coverage ≥ 15x (37, 38). We retained all isolates called “*S. delphini*” (n=7) by MALDI-TOF MS and most called “*S. intermedius/pseudintermedius*” (n=493) after filtering.

### 4.3.2 WGS validates MALDI-TOF MS as a precise tool for SIG identification.

For species identification of the 501 isolates, we queried NCBI and downloaded 40 *S. pseudintermedius* assemblies designated as “Complete Genomes,” along with all assemblies available at the time of analysis for *S. delphini* (n=22), *S. intermedius* (n=7), and *S. ursi* (n=1). With these reference assemblies, we performed an all-against-all whole genome alignment using FastANI (39), reporting average nucleotide identity (ANI) and percent total alignment between each isolate comparison (**Figure 4.2A**). As expected, all reference assemblies clustered with other reference assemblies of the same species. None of the 501 undesignated isolates reported ≥ 95% ANI with either the *S. intermedius* reference cluster or the *S. ursi* reference assembly. Eight of the 501 isolates were present within the *S. delphini* reference cluster (≥ 95% ANI) and the remaining (n=493) were found among the *S. pseudintermedius* reference cluster; all isolates only had ≥ 95% ANI with a single reference species cluster. All seven isolates called by MALDI TOF MS as “*S. delphini*” (**Figure 4.2B**) were among the eight isolates mixed into the *S. delphini* reference cluster. The eighth was the only isolate among the 501 assessed to have been erroneously typed, miscalled as “*S. intermedius/pseudintermedius*.” Thus, MALDI-TOF reports 99.80% and 100% precision when calling an isolate as “*S.
Chapter 4: Comparative genomics of diagnostic and commensal *S. pseudintermedius*

*intermedius/pseudintermedius* (492/493) and *"S. delphini"* (7/7), respectively. Given its predominance within our cohort, we centered the remainder of our investigation on *S. pseudintermedius*.

### 4.3.3 Niche adaptation is not apparent via whole genome architecture.

We first performed *in silico* multilocus sequence typing (MLST), finding both a surprising amount of diversity within our cohort and a surprising lack of representation of such diversity in the PubMLST database. Together we observe 315 unique allelic combinations among our *S. pseudintermedius* isolates, 237 of which are previously uncharacterized. We have consequently added ST2385-ST2621 to the PubMLST database, expanding the total number of reported sequence types by 9.9%. Within our cohort, the most well-represented sequence types include previously known ST759, ST181, and ST923, (n = 9, 6, and 6 isolates, respectively), and newly added ST2553, ST2550, ST2558, ST2414, ST2606, ST2544, and ST2545 (n = 11, 9, 9, 7, 7, 6, and 6 isolates, respectively). As expected, most of these STs exclusively represent isolates from the household arm of our cohort, reflecting the multiple sites of collection and multiple collection timepoints for each household. Exceptions include ST923 and ST2414, which represent a mix of human diagnostic, animal diagnostic, and household isolates, and ST181, which exclusively represents human diagnostic isolates. ST181 is of unique clinical relevance as a known MRSP lineage that has previously been reported in colonizing and diagnostic human and canine isolates from Israel, Thailand, Alberta, and New England; this is the first report of its identification in the American Midwest (13, 21, 40, 41). These data represent early indications of the uniqueness of our captured isolates relative to published *S. pseudintermedius* literature.
Chapter 4: Comparative genomics of diagnostic and commensal *S. pseudintermedius*

To determine the full population structure of *S. pseudintermedius* and assess for phylogenetic separation based on niche or species-host, we annotated all protein coding sequences of each isolate assembly using Prokka and PGAP, and generated a species-specific core genome alignment with Roary. From this, we generated a maximum-likelihood phylogenetic tree reflecting 1,673 genes present in ≥ 99% of isolates at ≥ 95% identity. Surprisingly, we did not observe clear clustering of isolates by niche-type (Figure 4.3A) or host-species (Figure 4.4A), though we do note two clades with high ARG burden that appear predominantly diagnostic in origin. Isolates do appear to cluster locally by MLST and by household, with the most common sequence types often representing multiple isolates from the same household, though a larger relationship between pairwise ANI and geographic distance did not materialize (R² = 0.019, Figure 4.5).

Upon expanding our analysis to accessory gene content present in 1-99% of isolates, we again did not observe any clustering by niche-type or host-species (Figure 4.3B, Figure 4.4B). An in-depth examination of Jaccard distances between the accessory genome content of isolates of the same cohort and niche, niche-only, and across cohorts and niches further confirmed a lack of differentiation by niche-type (Figure 4.3C); the same was observed when compared by host-species (Figure 4.4C). This is likely due to the magnitude of cloud genes (those present in ≤ 15% of isolates) within the *S. pseudintermedius* pangenome, implying a typical genome to be comprised of the 1,673 core genes, and a few hundred of the ~5,000 cloud genes that are unique to each isolate or a handful of isolates (Figure 4.3D). Finally, upon categorizing the whole genome of each *S. pseudintermedius* assembly into functional COG classifications, we again observe a nearly indistinguishable breakdown of gene
content across cohorts (Figure 4.3E). Collectively, we do not observe any niche differentiation among \textit{S. pseudintermedius} by whole genome analysis.

**4.3.4 Two clades of diagnostic isolates are distinguished by greater ARG burden and phenotypic resistance.**

We observed that two core genome clades of diagnostic isolates are associated with a substantial ARG burden (Figure 4.3A, green color strip). Comparing across niches, we find that diagnostic isolates have significantly more ARGs than isolates of household origin ($p < 0.0001$, Mann-Whitney; Figure 4.6A, 2C). Many diagnostic isolates, regardless of host-species, harbored antibiotic resistance genes (ARGs) in several antimicrobial classes (Figure 4.6A). This ARG overabundance among diagnostic isolates translates to greater rates of phenotypic resistance: human diagnostic isolates are significantly more resistant to ciprofloxacin, doxycycline, and oxacillin relative to other isolates, and animal diagnostic isolates are significantly more resistant to trimethoprim-sulfamethoxazole (TMX) ($q < 0.05$, FDR-corrected Chi-square; Figure 4.6B). Notably, isolates with greatest ARG burden ($\geq 9$ ARGs) were significantly more likely to also carry \textit{mecA} compared to those of lesser burden (2-8) ($p < 0.0001$, Chi-square; Figure 4.6D). A minority of isolates carrying the \textit{mecA} gene presented as oxacillin-susceptible MRSP (OS-MRSP) (Figure 4.6A).

**4.3.5 Syntenic sets of cellular defense genes distinguish specific clusters of diagnostic \textit{S. pseudintermedius} isolates**

In consideration of the repeat sampling of households, and the temporal proximity between some human and animal participants in the clinic, we assessed for persistence of \textit{S.
Chapter 4: Comparative genomics of diagnostic and commensal *S. pseudintermedius*

*pseudintermedius* strains in and across community and diagnostic settings. For this, we employed snp-sites and snippy to identify strain clusters, defined as isolates (1) within 25 whole-genome single nucleotide polymorphisms (SNPs) of each other (>99.999% ANI) and (2) sharing ≥ 98% assembly coverage. Isolates that did not meet these criteria but were within 600 core genome SNPs of another were considered to be of the same lineage (Figure 4.7A, Figure 4.8A). These efforts revealed widespread interrelatedness, capturing 230 of the 493 *S. pseudintermedius* isolates (n=186 within strain clusters and n=44 at lineage classification) (Figure 4.7B, 4.8B). The largest strain clusters (comprising 6-12 isolates per cluster) were composed entirely of isolates from the household cohort. Of significance, however, approximately half of the smallest strain clusters (capturing 2-3 isolates per cluster) comprised exclusively of human diagnostic isolates, implying some patients presenting at Barnes-Jewish Hospital are members of the same diagnostic *S. pseudintermedius* transmission networks. While our stringent strain cutoffs precluded appearances of any animal diagnostic isolates within strain clusters, or the observance of mixed human diagnostic-household strain clusters, expanding our analysis to lineage clusters found the appearance of both (n=7 animal diagnostic isolates, each a member of a different mixed cluster; n=10 mixed clusters total; Figure 4.7B). Appearances of lineage clusters encompassing mixed household and diagnostic isolates (2-5 isolates per cluster) further extends the transmission potential noted in human diagnostic-exclusive strain clusters and signifies that the two niches are not mutually exclusionary, suggesting that crossover events involving a recent common ancestor may have seeded isolates of the same lineage in colonizing and diagnostic settings.

Capturing isolate interrelatedness also enabled a phylogenetically-informed genome-wide association study (GWAS) between household and diagnostic isolates, in which isolates
of the same strain cluster were first collapsed into and represented by a single cluster-wide pangenome to avoid artificial signal from strain overrepresentation. These efforts identified two groups of genes that were entirely or nearly entirely absent in the genomes of household isolates (Figure 4.8C). First, we observed exclusively in diagnostic isolates the known pathogenicity island (ant (6)-Ia, sat4, aph (3’)-IIIa) (21, 26, 42-47) encoding resistance to aminoglycosides and streptothricin, flanked by either IS1182- or IS1216- family transposases (Figure 4.8D). The pathogenicity island was often found alongside a toxin-antitoxin system and was occasionally positioned directly downstream of _erm (B)_ and CAT, genes that traditionally encode resistance to macrolides, lincosamides, and streptogramin B (48), and chloramphenicol, respectively. Notably, when not in proximity to the pathogenicity island, _erm (B)_ was often found alongside trimethoprim-resistance gene _dfrG_. We also observed a variant of the type-IIIa CRISPR system containing _cas_ genes for Cas1-2, Csm2-6, and Cas10/Csm1 in 17 diagnostic isolates (Figure 4.8E). Of significance, all 17 isolates are also _mecA_-positive, and in at least three isolates we observe the CRISPR system to be adjacent to the _mecA-mecR1-IS6_ family transposon. While this type-IIIa system is found in an additional 48 isolates across the three cohorts, proteins in the systems of these 17 isolates cluster at ≥ 95% amino acid identity (AAI) with each other and ≤ 78% AAI with those in the other 48 isolates. Taken together, we observe a subset of diagnostic isolates of different lineages that are distinguishable from household _S. pseudintermedius_ through their accumulation of genes involved in cellular defense against antibiotics and bacteriophages.
4.3.6 Household isolates are more likely to share CRISPR spacers with other household isolates, agnostic of overall ANI.

In addition to the two variants of the type-IIIa CRISPR system, a type II CRISPR system harboring the cas9-cas1-cas2 genes is more commonly found among isolates across cohorts. Altogether, CRISPR-positive isolates (n = 238) harbored an average of 1.1 +/- 0.3 CRISPR systems, as some isolates harbored both type II and IIIa systems, and 15.4 +/- 5.6 spacers per CRISPR system (range 4-36 spacers) as identified by CRISPRCasFinder (49). These CRISPR spacers – segments of nucleic acid from past foreign invaders incorporated into bacterial genomes – convey an immunologic memory of prior bacteriophage infections and form the basis of an adaptive immune response engendered by the linked Cas proteins (50). Such prevalence of CRISPR systems and spacer content offers a window into the evolutionary histories of phage predation for each isolate, enabling a comparative analysis on shared or diverging exposures to natural predators within and across cohorts.

Towards this, we generated a co-occurrence matrix for all unique spacers shared between CRISPR-positive isolates at 100% ANI (Figure 4.9B). In consideration of phylogeny, all isolates of the same strain cluster were again condensed into a single cluster-representative pan-spacer collection prior to the spacer content comparisons. However, we noticed that in all cases, isolates that fell only into lineage clusters also shared all or nearly all of the spacers as their corresponding strain cluster-representative (Figure 4.9A), so these representatives were amended to be lineage cluster-representative for downstream analyses. We found high-identity spacer sharing to be infrequent between isolates, with most isolate-isolate comparisons sharing zero spacers between them. Nonetheless, 27.9% of pairwise comparisons shared at least one spacer; for these, an average of 3.7 spacers were shared, with up to
36 spacers shared between phylogenetically diverse isolates or lineage clusters. Despite accounting for lineage associations, we initially speculated that spacer sharing was a function of isolate relatedness, with phylogenetically similar isolates expected to share a more recent common ancestor and therefore a longer timeline of shared evolutionary and phage predatory histories; surprisingly, this did not appear to be the case, as there was only a very weak correlation ($R^2 = 0.044$) between ANI and shared spacer count (**Figure 4.9C**). However, when comparisons were binned based on cohort identity, we observed that household isolates were significantly more likely to share spacers with other household isolates ($q = 0.00120$, FDR-corrected Mann-Whitney) than human or animal diagnostic isolates (**Figure 4.9D**). Environmental exposure, rather than phylogenetic similarity, dictated spacer sharing, indicating niche separation is present among household *S. pseudintermedius* in the context of host defense.

**4.3.7 Persisting strains within households evolve defense genes amidst MRSA decolonization interventions.**

Core genome analysis revealed that some household isolates cluster together phylogenetically and constitute reoccurring strains across longitudinal samplings (**Figure 4.3A, Figure 4.10A**). Such households resemble microenvironments within which persistence, replacement, and/or reappearance of strain clusters are observable. 64% of households (28/44) in the HOME2 and SHINE sub-cohorts harbored strain clusters that were found across multiple collection events. Strain clusters were typically household-specific, though several were isolated in multiple households and are highlighted (**Figure 4.10A**, strain clus-
Chapter 4: Comparative genomics of diagnostic and commensal *S. pseudintermedius*

ters 01, 03, 04, 08). These data reveal extensive *S. pseudintermedius* diversity in the commu-
nity and the ability of strains to spread among inhabitants and surfaces once found within a household.

Each household underwent a period of attempted decolonization, in which mupiro-
cin, chlorhexidine/dilute bleach water, and/or Clorox were applied to human inhabitants or household surfaces (34, 35). Though intended to interrupt community-acquired MRSA transmission networks, these interventions present a real-world opportunity to observe the presence and extent of putative adaptive evolution within *S. pseudintermedius*. We hypothe-
sized that such decolonization efforts would result in evolved genomic responses in the form of an accrual of non-synonymous – base changing – single nucleotide substitutions (SNSes) in staphylococcal defense genes.

To this end, we leveraged high depth of sequencing to track SNS accumulation in ten household strain clusters that persisted through the decolonization period, together captur-
ing 45 isolates. These strain clusters comprise an average of 4.5 +/- 2.2 isolates collected over 6.8 +/- 3.3 months across 1.3 +/- 0.5 households (range 2-8 isolates, 3-12.4 months, and 1-2 households per persisting strain cluster). For each cluster, a high-quality isolate assem-
bly from the initial month of detection was chosen as the cluster reference, against which reads from all subsequent isolates within the strain cluster were mapped. We then ran in-
Strain to identify genes in subsequent isolates that accrued at least one SNS, together captur-
ting 182 genes across 270 reference-subsequent isolate comparisons (the same gene will appear in multiple comparisons if the mutation appeared early in the strain cluster and was maintained over time). The gene list was filtered to remove genes described as partial by
PGAP and those without clearly defined COG annotations; genes that remained were categorized based on SNS impact, with 91 harboring mutations that resulted in at least one non-synonymous SNS (NSS), and 34 whose mutations resulted in synonymous SNSes only (SSO). These categories represent adaptive and neutralizing/purifying selection, respectively (51).

5 of the 91 NSS genes and 0 of the 34 SSO genes were annotated by COG as defense mechanisms. Compared against the distribution of defense mechanism genes within a representative *S. pseudintermedius* genome, we observed a significant overrepresentation of defense mechanism genes within the NSS dataset ($p=0.023$, Binomial test) (Figure 4.10B). This relationship was unique to genes undergoing adaptive evolution, as the lack of SSO defense mechanism genes reflected the expected distribution ($p=1$, Binomial test). Importantly, there was no difference between the overall COG distribution of genes accruing SNSes (n=125) and the expected COG distribution of the full reference assembly ($p=1$, Multinominal goodness of fit), nor was there an overall difference between the COG distribution of NSS and SSO genes ($p=0.59$, Fisher’s Exact test) (Figure 4.11). Taken together, the collection of genes that have accrued mutations resulting in non-synonymous substitutions over time and through decolonization do not appear to collectively be different in annotated function from the genes collecting synonymous substitutions only, and both gene populations together resemble the overall distribution of functional annotation of genes seen in a complete *S. pseudintermedius* assembly. The NSS genes are distinguished from the expected distribution solely by their elevated presence of defense mechanism genes accruing base-changing mutations.

Of note, we observed that five of the 101 NSS genes were mutated in more than one strain cluster. A permutation test found that this gene count to be significantly outside the expected count, indicating the five genes collecting non-synonymous mutations to be chosen
non-randomly (Figure 4.10C). These genes include a multidrug efflux pump, *bshC*, *murG*, *ami_2*, and an ABC transporter permease (Figure 4.10D). Interestingly, while the efflux pump was the only gene to be formally annotated as a defense mechanism, *bshC* has a role in antifolate defense (52), *murG* catalyzes cell wall synthesis (53, 54), and *ami_2* is involved in endolysis (55). Most notably, mutations in *murG* in strain cluster 03 isolates were associated with a significant reduction in susceptibility to oxacillin (*p*=0.009, T-test) and trimethoprim-sulfamethoxazole (*p*<0.0001, T-test) (Figure 4.10E).

4.4 Discussion

We present a strain-resolved comparative genomics investigation of 493 *S. pseudintermedius* isolates captured across the American Midwest. These isolates were collected across multiple host-species and exist both commensally on human and pet skin and nares, pet fur, and on household fomites, as well as pathogenically within human and animal wounds, SSTIs, and respiratory and urinary tract infections. Contrary to what has been observed with *S. aureus* (56), we do not observe *S. pseudintermedius* clustering by core genome similarity or accessory genome content along host-species or niche delineations, or by geographic distance. This lack of correlation between pairwise ANI and geographic distance, coupled with our identification of over 300 unique multilocus sequence types, emphasizes a genomic landscape of *S. pseudintermedius* that is far more diverse than previously understood. Together, our findings suggest that the spread of *S. pseudintermedius* across multiple hosts and environments does not reflect longstanding lineage-based acclimation to host-species or niche.
Human and animal diagnostic isolates did, however, report significantly greater ARG carriage relative to those found in household settings, the first genomic indication of niche differentiation in this study. This is in part due to a transposon-mediated pathogenicity island identified by strain-corrected GWAS to be overrepresented among diagnostic isolates. Comprehensive testing of all isolates against 12 antibiotics similarly revealed significantly greater levels of phenotypic resistance against ciprofloxacin, doxycycline, oxacillin, and trimethoprim-sulfamethoxazole among human and animal diagnostic isolates. These antibiotics or antibiotic classes are first- and second-line therapeutics empirically prescribed for most clinical presentations that contextualize the human and animal diagnostic \textit{S. pseudintermedius} cohorts (57), exemplifying an alarming trend that the \textit{S. pseudintermedius} isolates found in clinical contexts are also the most resistant to the antibiotics frequently prescribed to resolve them. (58) Interestingly, isolates with the greatest ARG burden were significantly more likely to be MRSP. The finding that a minority of isolates that carried the \textit{mecA} gene presented an oxacillin-susceptible MRSP phenotype was likely due to the absence of transcriptional regulator \textit{mecR1, mecl,} or both (59), which was seen in the majority of \textit{mecA}-positive isolates.

238 of the 493 (48.3\%) \textit{S. pseudintermedius} isolates across our cohorts harbored at least one CRISPR system, far greater than the 8-20\% carriage rates observed by others (43, 60). We investigated sharing of CRISPR spacers between isolates, given their role as immunologic memory reflecting history of phage predation. Spacer-sharing was infrequent among isolates, an expected result given the range of bacteriophage invaders and the individualized trajectory of phage exposure and incorporation into immunologic memory. Still, 27.9\% of
pairwise comparisons shared at least one spacer. Remarkably, of these we observed that environmental exposure, rather than genomic relatedness, impacted the magnitude of shared spacers between pairwise comparisons. This trend was especially apparent among household isolates, which shared significantly more spacers with each other than with diagnostic isolates, reflecting for the first time a form of niche adaptation among commensal household S. pseudintermedius. Interestingly, like the ARG enrichment seen in diagnostic isolates, this niche response also appeared in the context of microbial host defense.

The household cohorts were originally designed to study decolonization protocols against community-associated MRSA, but also potentiated off-target collateral selection against coexisting S. pseudintermedius populations. Through our longitudinal sampling and sequencing, we were able to observe the genomic correlates of such interventions in real-world settings, which materialized through accrual of non-synonymous SNSes in genes specifically annotated as microbial host defense. We further observed five genes accruing these non-synonymous SNSes in multiple strain clusters in different households, four of which have direct or indirect roles in host defense. Ultimately, we observed during one decolonization intervention that strain cluster 03, which collected a non-synonymous SNS in cell wall formation gene murG, experienced a significant drop in susceptibility against oxacillin and trimethoprim-sulfamethoxazole. These results indicate that decolonization interventions may reduce drug susceptibility of some persisting S. pseudintermedius, making them more phenotypically akin to the diagnostic isolates collected in our hospital and veterinary clinic cohorts.

Our work is not without limitations. All isolates were captured in the American Midwest; though we observe considerable diversity by MLST, this has not been contextualized
Chapter 4: Comparative genomics of diagnostic and commensal *S. pseudintermedius*

against a global background. Repeat sampling of households within and across timepoints increases the risk of clonal overrepresentation when comparing against single isolate representatives from human and animal wounds and infections. While we have performed considerable mitigation efforts via generation of strain- and lineage-cluster representatives in our GWAS and CRISPR spacer screen, it is possible we were not completely successful. Conversely, not every *S. pseudintermedius* isolate captured in the household surveillance cohorts was sequenced, obfuscating our ability to differentiate strain persistence from disappearance and reintroduction within households. Lastly, while *murG* accrued a non-synonymous SNS in strain clusters 03 and 28, the isolate pair displaying a reduction in susceptibility in cluster 03 also accrued non-synonymous SNSes in several other genes, each of which warrant further investigation.

While others have conducted comparative genomics studies of *S. pseudintermedius* populations, all have so far been limited to one host-species and/or niche type (20-22, 25, 26, 31, 61, 62). Our study is the first of its kind to integrate whole genome sequencing of *S. pseudintermedius* isolated from both human and animal hosts as a commensal and as an opportunistic pathogen, as well as isolates captured on abiotic surfaces within household environments. Together we have compiled the largest multi-center genomics-based cohort study of *S. pseudintermedius* to date, and have added 237 new multilocus sequence types and 501 new SIG genome assemblies to publicly available databases (BioProject PRJNA908872). Within the context of existing literature, our work supports data by others that the *S. pseudintermedius* pangenome is overwhelmingly composed of shell genes present in < 15% of isolates (20, 21), likely influencing the lack of accessory genome signature by host-species or niche type. We extend reports of *S. pseudintermedius* captured on dogs with and without
Chapter 4: Comparative genomics of diagnostic and commensal *S. pseudintermedius*

pyoderma (25) that isolates collected in diagnostic conditions carry a significantly larger ARG burden than those found commensally, with our work showing this to be agnostic of host-species. Others have previously been unable to identify host-specific genes distinguishing human and canine clinical isolates (22); indeed, our GWAS indicates that *S. pseudintermedius* are distinguishable not by host species but instead by niche-type. There have been numerous reports of the (ant (6)-Ia, sat4, aph (3’)-IIIa) pathogenicity island and/or type-IIIa CRISPR system in some *S. pseudintermedius* isolate genomes (20-22, 26, 42-47, 60, 61), but our work is the first to show that these gene clusters are overrepresented among diagnostic relative to commensal isolates. With the inclusion of the type-IIc CRISPR system, our study is also the first to use spacer content to observe niche adaptation via putatively shared evolutionary exposures. Lastly, beyond *S. pseudintermedius*, we validate findings by others that MALDI-TOF MS has sufficient discriminatory ability to differentiate between SIG species (62), and add eight high-quality *S. delphini* genome assemblies to NCBI, a 23% expansion beyond existing resources. Altogether, this study is novel in that we have elucidated non-traditional niche differences within *S. pseudintermedius*, demonstrating parallel selection of defense mechanisms via gene acquisition in diagnostic isolates and gene mutation in household isolates, which would have not been uncovered by phylogenetic or phenotypic analyses alone.
Chapter 4: Comparative genomics of diagnostic and commensal *S. pseudintermedius*

4.5 Materials & Methods

4.5.1 Study Cohort and Identification.

Washington University Institutional review board (IRB) and Institutional Animal Care and Use Committee (IACUC) approval was obtained for this study. Human clinical specimens were cultured in the microbiology laboratory at Barnes-Jewish Hospital in St. Louis, MO according to laboratory standard operating procedures. Isolates from these specimens (n=181), primarily from human wounds, tissue, respiratory tract, and drainage, were collected as part of routine clinical care, with cultures submitted from patients with clinical symptoms suggestive of infection. Isolates collection occurred between December 2011 – July 2019. The diagnostic canine and feline SIG isolates were selected from a collection of isolates recovered from submissions to the Kansas State Veterinary Diagnostic laboratory. These isolates were recovered from specimens submitted for clinical culture, primarily from animal urinary tract and skin and soft tissue infections (n=100), with primary isolation occurring between January 2016 – June 2019. Colonizing isolates (from human nares, axillae, or inguinal folds and from dogs nares, mouth, or dorsal fur) and household environmental surface isolates (e.g., bed linens, kitchen table, refrigerator door handle, bathroom countertop, bathroom faucet handles, bathroom light switch, toilet seat, bathtub, television remote control, computer keyboard and mouse, and telephone) were collected through two *S. aureus* surveillance projects among households with children with recent *S. aureus* infections – the Staph Household Intervention for Eradication (SHINE) and Household Observation of MRSA in the Environment 2 (HOME2) as previously described (34, 35) (Pet = 108, Human = 31, Surface=151). All samples were selected from the first timepoint in which a SIG isolate was recovered from a specific individual, pet, or environment site. For human samples, the nares
were prioritized, followed by the axillae, and then the inguinal folds. In pets, the nares or mouth sample was prioritized, followed by the dorsal fur sample. In each of these studies, participants performed a decolonization regimen to eradicate *S. aureus* carriage from the skin and nares, consisting of intranasal mupirocin application and daily chlorhexidine gluconate body washes or dilute bleach-water baths.

**4.5.2 Identification and Resistance Characterization.**

The identification of all isolates was confirmed using the VITEK MS MALDI-TOF MS system (bioMerieux, Durham, NC). Susceptibility testing against oxacillin, delafloxacin, ciprofloxacin, enrofloxacin, cefoxitin, doxycycline, linezolid, trimethoprim-sulfamethoxazole (TMP-SXT), rifampin, clindamycin, erythromycin, and eravacycline was performed for each isolate via disk diffusion on Mueller Hinton agar. Methods followed the procedural guidelines outlined by the Clinical and Laboratory Standards Institute (CLSI) in the M02 and interpreted using M100 (30th ed) and VET01 (3rd ed) standards (63-65). Where CLSI breakpoints were not available, FDA breakpoints were used. Quality control was performed on each day of testing.

**4.5.3 Resistance Characterization.**

Pearson’s Chi-squared test was used to measure differences in AST interpretations across cohorts for 513 isolates using the R STATS package. Total susceptible and resistant counts for each cohort were compared against the sum for all other cohorts. For each antibiotic, isolates were excluded if results were unavailable for the specific antibiotic (n=30 for ciprofloxacin, n=30 for TMX) or if an intermediate antibiotic interpretation was reported
(n=27 for doxycycline, only reported for human diagnostic samples). P-values were corrected using the Benjamini-Hochberg false discovery rate method with the \texttt{p.adj} function in the R \texttt{STATS} package (method="fdr").

\textbf{4.5.4 Illumina Whole Genome Sequencing.}

Genomic DNA was isolated from cell-suspensions and sequenced as previously described (23, 66). Briefly, DNA was quantified with Quant-iT PicoGreen dsDNA assay (Thermo Fisher Scientific, Waltham, MA, USA), with 0.5 ng used as input for Illumina Nextera XT library preparation (Illumina, San Diego, CA, USA). Libraries were pooled at equal concentrations and sequenced on the Illumina NextSeq 500 HighOutput platform (Illumina, San Diego, CA, USA) to a depth of 1.5 million reads per sample (2 x 150 bp). FastQC was used to determine read quality (67). Trimmomatic (version 0.38; flags: leading, 10; trailing, 10; sliding window, 4:15; and minimum length (minlen), 60) was used to remove adapter sequences and low-quality reads from demultiplexed data (68).

\textbf{4.5.5 Core genome analysis.}

Contigs were assembled using Unicycler (version 0.4.8) (69). Deep-sequenced assemblies were downsampled to 100x coverage using seqtk (version 1.3) (70). Inclusion criteria of draft genomes for all downstream genomic analyses included >98% completeness, <2% contamination, 0% strain heterogeneity, N50 > 40 kbp, and >15x average coverage, as assessed by CheckM (version 1.0.13), QUAST (version 4.5), BBmap (BBtools version 1.0.1) and seqtk (version 1.3) (37, 38, 70, 71).
Chapter 4: Comparative genomics of diagnostic and commensal *S. pseudintermedius*

Assembled contigs were annotated using the NCBI Prokaryotic Genome Annotation Pipeline (PGAP, version 5.3) (72, 73). Core and accessory genome content for *S. pseudintermedius* and *S. delphini* was determined using Roary (version 3.12) (74). *In silico* multi-locus sequence typing (MLST) was performed for all isolates against the PubMLST database (75) using the mlst tool (version 2.19) (76). 237 new multilocus sequence types representing 376 *S. pseudintermedius* isolates were uploaded and accepted to PubMLST under the submission ID BIGSdb_20220927223226_035094_56619 and BIGSdb_20230109221715_2606963_61712.

A maximum likelihood tree was constructed with FastTree (version 2.1.10) and visualized using iTOL (77, 78). Pairwise average nucleotide identity (ANI) values were obtained using fastANI (version 1.1) (39). 70 SIG genome assemblies (40 *S. pseudintermedius*, 22 *S. delphini*, 7 *S. intermedius*, and 1 *S. ursi*) were downloaded from NCBI and used to determine species-level taxonomy of all isolates.

Zip codes of the residence of the hospital patient or household of each isolate were collected. Zip code metadata was unavailable for the animal diagnostic cohort, which was excluded from this analysis. Geographic distance between isolates was determined using the geocode_zip function (R zipcodeR package, version 0.3.4) (79). This was plotted against pairwise ANI after removal of self- and repetitive-comparisons using the ggscatter function (R ggpubr package, version 0.4.0; flags: add="reg.line", cor.method="pearson", conf.int=TRUE, core.coef=TRUE) and ggmargin (R ggExtra package version 0.10) (80, 81).
4.5.6 Accessory genome analysis.

The gene_presence_absence.Rtab output file from Roary was purged of core genes (those present in >99% of isolates), resulting in 5,910 accessory genes. Presence-absence of these were used to calculate Jaccard distance between all isolates using the `vegdist` function (R vegan package) (82). Clustering by accessory gene content similarity was visualized through principal coordinate analysis using the `pcoa` and `ggplot` functions (R ape and ggplot2 packages) (83, 84). Jaccard dissimilarity scores were plotted following isolate metadata associations, including site of isolation and host-type. COG categories were assigned to all genes within each isolate assembly using eggnog (version 2.0.1; flags: -m diamond --query-cover 0.9) (85, 86).

ARGs in the protein FASTA (.faa) output files from Prokka were annotated against the NCBI comprehensive database of acquired and intrinsic antimicrobial resistance proteins at >90% identity and >90% coverage using AMRfinder (version 3.9.8; flag: --organism Staphylococcus_pseudintermedius (87). A presence-absence heatmap was generated using the `pheatmap` function (R pheatmap; flag: clustering_method = ‘mcquitty’), with columns clustered based on shared ARG content and rows organized by antimicrobial class (88). Associated metadata, including cohort, AST interpretation, and antimicrobial class, are displayed as color strips. ARG count was compared between diagnostic and colonizing isolates (two-tailed Mann-Whitney) and plotted in Prism 9, and `mecA`-positivity was tested for association with 'high ARG burden' status (Chi square) and plotted using the `ggplot` function.
4.5.7 Strain Cluster identification.

Isolate-specific SNPs were called against the Roary-generated core genome alignment file using snp-sites (version 2.4.0) (89). Isolate assemblies were then compared in an all-versus-all manner based on their shared SNP content, with 600 core genome SNPs considered the threshold for lineage-based relatedness (Figure 4.7). Intra-lineage isolates were further compared using Snippy (version 4.4.3) for pairwise whole genome SNP calling (90). Isolates were considered members of the same strain cluster if their whole genome SNP distance (Snippy VariantTotal) was below 26 (≥99.999% ANI) and overall coverage breadth ≥98% (91, 92). Accessory gene presence-absence data for isolates of the same strain cluster were collapsed into single strain-representative accessory genomes to control for phylogenetic overrepresentation in downstream analyses.

Phylogenetically-controlled genome-wide association analysis between diagnostic and household isolates was performed using Scoary (version 1.6.16) on the strain-informed gene_presence_absence.csv (93). Presence-absence of genes with Benjamini-Hochberg-adjusted p-value <0.05 were hierarchically clustered and visualized using the pheatmap function (R pheatmap and dendsort packages); isolates lacking all significant genes were removed from the heatmap, as were genes present in >100 isolates (n=17, mostly tRNA alleles), or genes with nonspecific “domain-containing protein” annotations (n=6). Gene structures significantly associated with a cohort were visualized using EasyFig (version 2.2.2) (94). Towards this, GenBank files outputted by PGAP were first filtered to only the contig(s) carrying the genes of interest, and then preprocessed in ApE (version 3.1.2) (95).
4.5.8 CRISPR Spacer Investigation.

Contigs harboring CRISPR-Cas operons were extracted from isolate assemblies, concatenated, and inputted into CrisprCasFinder (49). Identified spacer sequences were captured, labeled with their isolate of origin and numeric position within the Cas operon, and reverse complements were generated using The Sequence Manipulation Suite (96). All spacers and their reverse complements were inputted into clustalo (version 1.2.4; flags: --dealign --full --percent-id) to generate an all-v-all ANI-based spacer distance matrix (97). Self-comparisons (spacers from a given Cas operon compared against themselves) were removed, and only the remaining spacer comparisons with ANI = 100% were retained for spacer-co-occurrence analysis. The subset dataframe was further reduced such that inverse entries were removed (i.e., if spacer 1 vs. spacer 2 is already present, spacer 2 vs. spacer 1 is removed). The cleaned dataframe was converted into an all-v-all isolate occurrence matrix plotting the number of spacers shared at 100% ANI between each isolate comparison, and visualized as a bubbleplot using the ggplot function. The number of shared spacers between isolates was plotted as a function of ANI and visualized as a scatterplot with trendline (geom_smooth (method=“lm”)) using the ggplot function. For each cohort type, isolate-isolate shared spacer counts were binned as intra-cohort or inter-cohort, compared (FDR corrected one-way Mann Whitney) (wilcox.test (paired = FALSE, exact = FALSE, conf.int = TRUE, alternative = “greater”); R STATS package), and plotted in Prism 9.

4.5.9 Within Household Tracking and Evolution.

To investigate strain dynamics within households, human colonizing, animal colonizing, and surface isolates were visualized within and between households and studies
(HOME2 and SHINE). Strain clusters were plotted by household and faceted by study. Households with 5 or more isolates or where strain clusters appeared in more than one household across time were longitudinally plotted.

Isolates of the same strain cluster that appeared in households over a minimum time span of 3 months, were assembled at ≥30x coverage, and were visualized in Figure 4.10A were evaluated for single nucleotide substitution (SNS) accumulation over time. For clusters that appeared in both SHINE and HOME2, only isolates from SHINE households were used to maintain consistency among decolonization exposures. If multiple isolates of the same strain cluster were captured in the index month, a reference isolate was chosen using the following formula: Completeness − (5 x Contamination ) + ( Contamination x (Strain Heterogeneity / 100)) + (0.5 x log(N50)) (98). Open reading frames (ORFs) within the reference assembly for each strain cluster were called via Prodigal (version 2.6.3) (99). Reads for all subsequent longitudinally-captured isolates within a strain cluster were aligned to the strain cluster reference assembly using Bowtie2 (version 2.4.2) (100). Each strain cluster’s reference assembly and Prodigal-called ORFs, as well as the Bowtie2 alignment file for each subsequent isolate against the strain cluster reference, were inputted into inStrain-profile (version 1.5.7) to track SNS accumulation over time, as well as the genes they reside in (92). The inStrain gene_info output file was reviewed to identify all instances of ORFs with at least one SNS (SNS_count ≥ 1; n=182 unique genes). These Prodigal-called ORFs were united with their corresponding pgap annotations via blastp (BLAST+ version 2.12.0) and assigned a COG category via eggnog (version 2.1.8; flags: -m diamond --pident 90 --query_cover 80) (85, 101). Genes with blank, “-”, or multi-letter COG categories, as well as genes with “partial” in
their pgap annotation, were removed from further analysis, resulting in 125 remaining genes with SNS_count ≥ 1.

Genes were divided into two groups, depending on whether all harbored substitution (s) were synonymous (synonymous SNSes only; SSO), or if at least one substitution was non-synonymous (non-synonymous SNS es; NSS). Observed COG distribution of NSS genes annotated as “defense mechanisms” was assessed for enrichment (one-sided Binomial test) (binom.test (alternative = “greater”); R STATS package). The distribution of COG categories among all genes accruing SNSes (SNS_count ≥ 1) was compared against the expected COG distribution of a full S. pseudintermedius assembly (Multinomial goodness of fit) (multinomial.test (useChisq = FALSE, MonteCarlo = TRUE); R STATS package). Overall distribution of COG categories was compared across the NSS and SSO groups (Fisher’s Exact Test of Independence) (fisher.test (simulate.p.value); R STATS package).

A permutation test was implemented to determine whether the number of observed NSS genes mutated in multiple strain clusters was significantly greater than the expected count. Gene positions (numbered 1 to 1,871, the total number of genes in the reference S. pseudintermedius assembly) were randomly sampled without replacement for each cluster, with the number of samplings equaling the number of genes that had undergone at least one non-synonymous substitution in that cluster (range 3-19 samplings, cluster-dependent). Samplings across clusters were combined and frequency of each gene position was determined. The total number of gene positions that were randomly sampled more than once were recorded. This process was iterated 10,000 times, and the frequency of gene positions randomly sampled more than once (0-8 total gene positions) was plotted in ascending order
numerically. Gene counts that were greater than the value reported by the 9,500th iteration was determined to be significantly outside the range expected from chance alone.

Two isolates in strain cluster 03 flanking decolonization reported drops in susceptibility against two antibiotics. Technical replicates of each isolate were retested against oxacillin and trimethoprim-sulfamethoxazole and results were compared using Student’s T test. Results were plotted in Prism 9.

### 4.5.10 Statistics and reproducibility.

Unless otherwise stated, comparisons of categorical data were performed by Pearson’s chi-squared test or Fisher’s exact test depending on sample size while comparisons of continuous variables were tested using the Mann-Whitney U test. All tests were two-tailed, and statistical significance was defined as $p < 0.05$. Multiple hypothesis correction was performed when appropriate using the FDR method. Multinomial goodness of fit was used to determine if the distribution of genes accruing SNSes by COG category differed from expected. Binomial test was used to determine if a specific COG category was enriched. A permutation test was employed to determine whether observed count of NSS genes appearing in multiple clusters was greater than expected. Statistical analyses were performed using the R STATS package or Prism 9.
Chapter 4: Comparative genomics of diagnostic and commensal S. pseudintermedius

4.6 Data availability

All isolate short reads and assemblies are available on NCBI SRA and GenBank under BioProject PRJNA908872. A full list of specimen metadata, antibiotic susceptibility testing data, assembly quality and genome metadata, and inStrain data is available in Tables 4.1 – 4.4, respectively.
4.7 References


Chapter 4: Comparative genomics of diagnostic and commensal *S. pseudintermedius*


Chapter 4: Comparative genomics of diagnostic and commensal *S. pseudintermedius*


Chapter 4: Comparative genomics of diagnostic and commensal S. pseudintermedius


Chapter 4: Comparative genomics of diagnostic and commensal *S. pseudintermedius*


Chapter 4: Comparative genomics of diagnostic and commensal *S. pseudintermedius*


Chapter 4: Comparative genomics of diagnostic and commensal *S. pseudintermedius*

4.8 Figures

![Workflow schematic](image)

**Figure 4.1:** Workflow schematic.
Figure 4.2: WGS identifies 493 isolates as *S. pseudintermedius* and 8 as *S. delphini*. (A) Pairwise ANI of 500 sequenced isolates with 70 NCBI reference assemblies. The large upper-left quadrant represents *S. pseudintermedius* assemblies. The next compact quadrant is composed of *S. intermedius* assemblies, all of which originated from NCBI. The bottom-right quadrant represents a mix of *S. delphini* isolates from our cohort and from NCBI. (B) Core genome phylogeny of the eight *S. delphini* isolates captured in our household cohort.
Chapter 4: Comparative genomics of diagnostic and commensal *S. pseudintermedius*
Chapter 4: Comparative genomics of diagnostic and commensal *S. pseudintermedius*

Figure 4.3: Niche adaptation is not apparent via whole genome architecture. (A) Core genome phylogeny of 492 *S. pseudintermedius* isolates. Cohort, MLST, household (if applicable), and ARG burden are indicated by concentric color strips. (B-C) Jaccard dissimilarity by accessory gene content represented by (B) principal coordinate analysis ordination and (C) beta diversity compositions of within-group and between-group Jaccard distance. White, light gray, and dark gray violins indicate within cohort and niche, niche-only, and across cohort and niche comparisons, respectively. (D) *S. pseudintermedius* pangenome composition. (E) Total gene content breakdown by COG category.
Chapter 4: Comparative genomics of diagnostic and commensal *S. pseudintermedius*

![Diagram of comparative genomics of diagnostic and commensal *S. pseudintermedius*]

**A.**

- **Ring 1:** Diagnostic v. Household
- **Ring 2:** Host Type
- **Ring 3:** Combined

**B.**

- **Accessory Gene Content**
  - **PCA1 (3.41%)**
  - **PCA2 (3.41%)**

**C.**

- **Jaccard dissimilarity score**
  - Animal
  - Human
  - Environment

- **Box plots** showing Jaccard dissimilarity scores for different host types.

---

[168]
Figure 4.4: Host-species adaptation is not apparent via whole genome architecture. (A) Core genome phylogeny of the 493 *S. pseudintermedius* isolates. Ring 1 represents niche-type consistent with Figure 1A. Rings 2-3 describe host-species or environment for each isolate. (B-C) Jaccard dissimilarity by accessory gene content similarity represented by (B) principal coordinate analysis ordination and (C) beta diversity compositions of between-group and within-group Jaccard distance. White, light gray, and dark gray violins indicate within host-species and niche, within host-species only, and across host-species and niches, respectively.
Figure 4.5: Relationship between ANI and geographic distance. Self-comparisons were removed. Animal diagnostic isolates were excluded based on availability of zip code data. The centroid of each zipcode was used to calculate the latitude and longitude for each isolate.
Chapter 4: Comparative genomics of diagnostic and commensal S. pseudintermedius

A.

B.

C. Whiskers: 10-90 Percentile

D. Number of Isolates

ARG Count
Figure 4.6: Diagnostic isolates are distinguished by greater ARG burden and phenotypic resistance. (A) Antibiotic resistance gene content and antibiotic susceptibility testing results, organized by antibiotic class. Only isolates with ≥ 2 ARGs are shown. (B) AST interpretation for commonly-prescribed antibiotics (Abbreviation TMX, trimethoprim-sulfamethoxazole). Asterisks indicate a cohort is significantly more resistant or susceptible to the respective antibiotic relative to all other isolates (q < 0.05, Chi-square). (C) ARG count for diagnostic and household isolates (p < 0.0001, Mann-Whitney). (D) Presence of mecA among isolates, distributed by total ARG count. MRSP is significantly overrepresented among high ARG burden isolates (≥ 9 ARGs) relative to low ARG burden isolates (2-8 ARGs) (p < 0.0001, Chi-square).
Figure 4.7: Determination of lineage cluster boundaries. (A) Pairwise core genome SNP distance. 600 core genome SNPs (inlet) was the designated cutoff for lineage relationships. (B) Cohort composition of lineage clusters, plotted by cluster size.
Chapter 4: Comparative genomics of diagnostic and commensal *S. pseudintermedius*

A. ANI

B. Number of Strains

C. Total Isolates in a Strain

D. IS1216 family transpose

E. Cochrane Cohort
Figure 4.8: Syntenic gene structures continue to distinguish select diagnostic isolates following strain correction. (A) Pairwise comparisons of ANI and coverage within lineage clusters. Most comparisons reside in the top-right quadrant (> 99.999% ANI, > 98% coverage), indicating most isolates of the same lineage are members of the same strain cluster. (B) Cohort composition of strain clusters, plotted by cluster size. (C) Genes found by GWAS to be overrepresented among diagnostic or household isolates and strain cluster representatives. Only isolates or strain clusters with at least one GWAS-identified gene are displayed. Blue and red brackets denote a CRISPR-Cas system and pathogenicity island, respectively. (D-E) Representative appearances of the (D) pathogenicity island and (E) type-IIIa CRISPR-Cas system among diagnostic isolates.
Figure 4.9: Household isolates share more spacers with isolates within their niche than with diagnostic isolates, agnostic of pairwise ANI. (A) Co-occurrence matrix of spacers shared at 100% ANI between only the 17 mecA-positive isolates with Type-IIIa CRISPR systems. Bubble size and saturation reflect total quantity of spacers and percent of spacer repertoire shared between two isolates, respectively. HI_0003-HI_0186 and HI_0035-HI_0177 are two examples of isolate pairs that fall into lineage but not strain clusters, yet have virtually identical spacer repertoires. (B) Co-occurrence matrix of spacers shared at 100% ANI between all CRISPR-positive isolates. Isolate names and intra-cohort pairwise comparison bubbles are colored by cohort. (C) Pairwise ANI by shared spacer count. (D) Pairwise shared spacer counts of household isolates with other household isolates (intra), and with diagnostic isolates (inter) (q = 0.00120, FDR-corrected Mann-Whitney).
### Chapter 4: Comparative genomics of diagnostic and commensal *S. pseudintermedius*

#### A. Households with an isolate count of 5 or greater

<table>
<thead>
<tr>
<th>Month</th>
<th>P</th>
<th>M</th>
<th>L</th>
<th>K</th>
<th>J</th>
<th>G</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>2017</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>21</td>
</tr>
</tbody>
</table>

#### B. Defense Mechanism Genes

<table>
<thead>
<tr>
<th>Key</th>
<th>Gene</th>
<th>COG</th>
<th>Cluster</th>
<th>Isolates</th>
<th>Mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>RND multidrug efflux pump</td>
<td>V</td>
<td>01</td>
<td>1</td>
<td>N:S2002F</td>
</tr>
<tr>
<td>2</td>
<td>bshC – bacillithiol biosynthesis</td>
<td>S</td>
<td>02</td>
<td>1</td>
<td>N:A1061T</td>
</tr>
<tr>
<td>3</td>
<td>murG – cell wall formation</td>
<td>M</td>
<td>03</td>
<td>6</td>
<td>N:S628L</td>
</tr>
<tr>
<td>4</td>
<td>Ami_2 amidase – cell division</td>
<td>M</td>
<td>04</td>
<td>1</td>
<td>N:N437K</td>
</tr>
<tr>
<td>5</td>
<td>ABC transporter permease</td>
<td>P</td>
<td>05</td>
<td>1</td>
<td>N:G137C</td>
</tr>
</tbody>
</table>

#### C. Count (10,000 simulations)

![Count graph](image)

#### D. Expected vs. SSO vs. NSS

- **Expected distribution (33 of 1,871 genes)**
- **Synonymous SNSes only (0 of 34 genes)**
- **Non-synonymous SNSes (5 of 91 genes)**

#### E. Zone of Inhibition (mm)

- **Oxacillin**: ![Graph](image)
- **Tri-sulfa**: ![Graph](image)
Figure 4.10: Persisting strains accrue mutations in defense mechanism genes amidst household MRSA decolonization. (A) Observance of strain clusters across households, faceted by sub-cohort. MRSA decolonization is highlighted and months without sample collections are noted in gray. Samples collected during the highlighted HOME2 intervention were collected immediately prior to MRSA decolonization and are considered baseline samples. Samples collected during the highlighted SHINE intervention were collected during MRSA decolonization. Bubble size corresponds to the number of isolates per strain cluster found in a collection month. (B) Defense mechanism genes as a percent of total genes accruing synonymous SNSes only (SSO) and non-synonymous SNSes (NSS), compared against the expected distribution of a reference S. pseudintermedius genome. (C) Number of genes accruing non-synonymous SNSes in multiple clusters, over 10,000 iterations. 9,500 iterations are captured to the left of the dashed vertical line. (D) Schematic representation of the five NSS genes mutated in multiple strain clusters. COG key: [M] Cell wall/membrane/envelope biogenesis, [P] Inorganic ion transport & metabolism, [S] Function unknown, [V] Defense mechanisms. (E) AST results for oxacillin and trimethoprim-sulfamethoxazole for Cluster 03 isolates before and during household MRSA decolonization.
Figure 4.11: COG distributions of genes accruing at least one SNS. (A) Distribution of COG categories across a representative *S. pseudintermedius* genome compared to that of all genes accruing SNSes. (B) Distribution of COG categories across all NSS and SSO genes.
Chapter 5: Conclusion and Outlook

Throughout this Thesis, and my time in the Dantas lab, I have explored a wide range of topics with seemingly little connection. The careful reconstruction of microbial genomes from dynamic communities in early life may appear unrelated to the whole genome sequencing of *Staphylococcal* isolates. Tracking within-host evolution of persisting microbes through massive structural remodeling could be considered peripheral to the search for the genomic underpinnings of drug resistance or the assessment of niche adaptation within a species. Yet, several themes unite my research and contextualize my Thesis.

None of this work would be possible without recent advances in microbial genomics and computational biology that have enabled cost-effective sequencing of isolates and metagenomes at scale. Five years ago, I would be profiling communities solely based on relative abundances of marker-genes – the intricacies of intraspecies variation would be completely lost on my analyses. Yet all work described in this Thesis, be it from isolates or metagenome-assembled genomes, utilize whole-genome resolution to make claims about encoded functional capacity, SNP tracking, and adaptation to environments and perturbations. These advancements have enabled me to study the interface of humans and microbes with targeted questions and at scale. From individual microbes to whole communities. From infants to all household members. In hospital settings to the real-world.

If there is one thing I have learned during my graduate studies, it is that microbes will find a way to continue. They are remarkably resilient and have demonstrated their ability to
Persist through or adapt to all perturbations or interventions that I have investigated. Microbes were here long before us, and will continue to thrive long after we are gone. In the meantime, it has been a joy to observe their evolution alongside humans, as perceived through the brief glimpses in time offered by the samples analyzed in this Thesis.
Addendum: The resistance within – antibiotic disruption of the gut microbiome and resistome dynamics in infancy

The contents of this section are adapted from a review published in Cell Host & Microbe:


* = equal contribution
Addendum: The resistance within

A.1 Summary

Intestinal host-microbiota interactions during the first year of life are critical for infant development. Early-life antibiotic exposures disrupt stereotypical gut microbiota maturation and adversely affect childhood health. Furthermore, antibiotics select for resistant bacteria and enrich the resistome – the compendium of antibiotic resistance genes – within the gut microbiota. Here, we discuss acute and persistent impacts of antibiotic exposure during infancy on pediatric health, the gut microbiome, and, particularly, the resistome. Reviewing our current understanding of antibiotic resistance acquisition and dissemination within and between microbiomes, we highlight open questions which are imperative to resolve in the face of rising bacterial resistance.
Addendum: The resistance within

A.2 Introduction

During the first weeks of life, neonates are colonized with a diverse set of bacteria, viruses, fungi, and archaea (1–4). The collection of these microbes and their functions – the microbiome – diversifies and matures under environmental selection (1, 2). During early life, critical host-microbe interactions at the intestinal interface promote healthy immune and metabolic development.

In high-income countries eighty percent of children receive antibiotics during the first 48 months of life (5, 6). In low-income and middle-income countries, the prescription rate is even higher, and children are estimated to receive an average of 11 antibiotic courses in the first two years of (7). In high-income countries the majority of exposures occur in the first year of life (infancy), during which 40-70% of all infants receive at least one course of antibiotics (8, 9). Although antibiotics are among the most important medical advances in the last hundred years and have substantially decreased global mortality due to infectious disease, their overuse has accelerated the spread of resistance among bacterial populations. Moreover, they can also disrupt microbiome structure and function, often resulting in dysbiotic microbiome states associated with adverse health outcomes (10).

Our growing understanding of the gut microbiome has revealed the collateral damage of antibiotic use on the complex ecosystem of bacteria colonizing the gastrointestinal tract. Compositional and functional changes of the microbiome following antibiotic exposure result from differences in bacterial drug sensitivity determined either by the expression of antibiotic resistance genes (ARGs) or antibiotic concentration-dependent effects in the gut (3, 11). The collection of ARGs – the resistome – is a core characteristic of the microbiome. Pathogens can acquire ARGs from less pathogenic members of the microbiota via horizontal gene
transfer (HGT), potentially gaining antibiotic resistance phenotypes of immediate clinical impact. While environment, maternal health, diet, birth mode, and gestational age can all affect maturation of the gut microbiome in infancy (12), here we focus on antibiotic exposure and its impact on the gut microbiome and resistome during early life. Antibiotic exposure is a major determinant of the diversity and composition of the infant gut resistome (2, 3). The effect of antibiotics is impacted both by host-intrinsic and extrinsic factors including, but not limited to, age, co-morbidities, genetics, and microbiome composition, as well as host diet and environment, and the spectrum, route, duration, and history of exposure to the antibiotic (3, 10, 11, 13).

While our knowledge about the dynamics that shape resistome establishment and the effects of antibiotics on the microbiome in early life has improved in recent years, critical questions about the inter-bacterial dynamics that contribute to the spread of resistance within bacterial communities, the genome-resolved functional impact of antibiotics in early life, and the clinical consequences of resistance within the microbiome remain unanswered. In this review we summarize knowledge on the impact of antibiotics on childhood health outcomes, the infant microbiome and resistome, and outline future avenues of research.

A.3 Antibiotic exposure and health outcomes in early childhood

In early childhood, respiratory, skin, and gastrointestinal infections are the most common indications for use of antibiotics (Figure A.1), but retrospectively, in 20-40% of cases prescriptions are found to be inappropriate according to medical guidelines corresponding to a patient’s diagnosis (14–16). While most pediatric antibiotic prescriptions occur in the outpatient setting (17), exposure is frequent and extensive in neonatal intensive care units
Addendum: The resistance within

(NICUs), where over 80% of preterm infants receive antibiotics within 72 hours of birth (18). Although the most common antibiotics administered in NICUs are narrow spectrum agents such as ampicillin, gentamicin, and vancomycin, up to 26% of antibiotic-exposed neonates receive broader spectrum 3rd and 4th generation cephalosporins and carbapenems (19). Empiric antibiotic therapy for suspected infections is the main indication in these cases (18, 20). In the majority of them, negative bacterial culture results suggest alternate etiologies, but neonates often continue to receive antibiotics, contributing to their overuse (20). Thus, antibiotic use is common in infancy regardless of the environment (Figure A.1).

Antibiotic use has declined in NICUs over the last 20 years (21). Where implemented, antibiotic stewardship programs have contributed to a more rapid decline in antibiotic prescriptions in the NICU compared to institutions without such programs (21). More broadly, these programs have led to declining antibiotic prescriptions in both inpatient and outpatient pediatric populations in the United States over the last ten years, continuing a trend that began in the 2000s (22, 23). While these are hopeful signs that antibiotic overuse can be curtailed (17), recent epidemiological studies have found sustained high prescription rates in outpatient settings, rural communities, low- and middle-income countries, and in NICUs worldwide (19, 24, 25).

Early-life antibiotic exposure has been associated with a multitude of long-term adverse health outcomes, including childhood asthma, obesity, inflammatory bowel disease, and impaired growth (Figure A.1) (10). Recent research, however, has called some of these associations into question. The lack of association between early-life antibiotics and autism or hyperactivity disorders and conflicting results for asthma in familial analyses stands in
Addendum: The resistance within

contrast to significant associations in case-control designs and implicates common household co-exposures as a potential confounder of previous analyses (26–28). Nevertheless, as studies using genetically-identical and environmentally-controlled animal models provide mechanistic support for a direct link between early-life antibiotics and adverse health outcomes (29–32), further epidemiological research is required to resolve apparent discrepancies in human populations. Though the mechanisms underlying the impacts of antibiotics on many chronic health sequelae remain elusive, their impact on the microbiome and host immune system are likely contributors.

A.4 Effects of antibiotics on the gut microbiome of healthy infants

At birth, all neonates have a low diversity gut microbiome (2, 33). In the first weeks of life, infants accrue microbial species, including key anaerobic symbionts from the Bifidobacteriaceae, Clostridiaceae, and Lachnospiraceae families, following individualized trajectories (2, 34). The early-life gut microbiome undergoes distinct developmental phases, during which weaning is an inflection point that accelerates microbiome maturation (35, 36). Multiple factors reviewed elsewhere (37), including delivery mode, sex, gestational age, antibiotics, diet, and environment shape the composition and trajectory of individual infant microbiomes. Antibiotics can stunt early life gut microbiome maturation, resulting in transiently delayed dynamics and potential regression relative to age matched controls (34, 38, 39).

Emerging evidence implicates the first year of life as a critical window during which cues from the intestinal microbiota educate the developing infant immune system (40–42). Infants with gut microbial dysbiosis in the first 40 days of life exhibited altered populations
Addendum: The resistance within

of circulating immune cells and increased T cell activity at 12 weeks (43). Recent work has further shown that individual microbial taxa, specifically bifidobacteria, and their human milk oligosaccharide utilization gene cluster are key to educating intestinal T helper cell populations and preventing inflammation during the first months of life (42). Microbiota disruption during this time can hamper immune development and is associated with adverse childhood health outcomes (40, 41). Critically, the impact of antibiotic exposure is largest in the first year of life (34), resulting in attenuated gut microbiome maturation and diversity (1). While these alterations are generally transient, some may persist long-term (2, 44), and the immunological importance of early life exacerbates the impact of even temporary perturbations (41).

Depletion of vulnerable taxa essential for host immune education may be at the core of observed lasting health impacts of early-life microbiota disruption (42, 43). Bifidobacteria, and other anaerobic commensals known to modulate host-microbiota interactions in infancy, are particularly vulnerable to antibiotic-induced replacement (13, 34). Data from animal models and observational studies suggest that transient early-life antibiotic-induced microbiota disruptions may indeed be at the root of long-term health consequences (29–31, 40). Further studies at the interface of mucosal immunology and the gut microbiome are critical to identify elusive mechanistic links between transient gut microbiome disruptions and long-term health.

Recently, strain-resolved metagenomic analysis has suggested that community-level characterization may not reveal the full impact of antibiotics on the developing infant gut microbiome. Healthy, term neonates often accrue multiple strains of the same species that
Addendum: The resistance within

can differ in encoded functional potential. In contrast, the antibiotic-exposed infant gut microbiota exhibits less strain diversity compared to antibiotic-naïve infants (1). For example, *Bacteroides spp.* in antibiotic-exposed infants were often dominated by a single strain, unlike antibiotic-naïve infants who harbor multiple strains of the same species, reflecting reduced within-species genomic variation and potentially significant functional effects following antibiotic exposure (1). This effect is strongest for species seeded in a single colonization event (1). Antibiotics can cause rapid shifts in the genomic composition of individual species that often occur without obvious changes in relative abundance, highlighting that these antibiotic-induced effects are missed by species-resolved analysis (45). It remains to be investigated whether antibiotic-induced strain depletion persistently impacts within-species genomic diversity, how taxonomic shifts impact the functional repertoire of keystone species, and whether strain shifts impact health outcomes.

**A.5 Effect of antibiotics during postnatal hospitalization**

Infants born prematurely frequently require hospitalization in NICUs where gut microbiome diversity remains low for months and is often comprised of only a few species (2–4). Antibiotic-resistant *E. coli, Klebsiella spp., Staphylococcus spp.*, and *Enterococcus spp.*, common etiologic agents of infant bloodstream infections (BSIs) (46), dominate the neonatal preterm gut microbiome and their environment in the first months of life (3, 4, 47). The preterm gut microbiome clusters into distinct community states dominated by one of these genera (4). Staphylococci are the dominant member in the first 7 days of life with high relative and absolute abundance (4). This initial stage can be followed by an influx of *Klebsiella spp.* exploiting physiological niches and repressing the abundance of Staphylococci. Alternatively,
Addendum: The resistance within

if Enterococci are present, they repress Klebsiella spp. abundance (4). Thus, the dynamics of hospitalized preterm infant gut microbiomes are characterized by opportunistic, often drug-resistant pathobionts, differing significantly from the dynamics observed in their non-hospitalized term counterparts (2).

In premature infants, increased antibiotic duration, spectrum, and frequency correlate with decreased overall diversity and increased abundance of potentially antibiotic-resistant gram-negative organisms such as E. coli, Enterobacter spp., and Klebsiella spp. (2, 3, 48). The impact of antibiotics on preterm gut microbiome diversity and richness is, however, not uniform, as most antibiotic classes exhibit specific inhibition spectra (11). Generally, broader spectrum antibiotics such as carbapenems and 3\textsuperscript{rd} generation cephalosporins decrease intestinal species richness more than narrower agents such as ampicillin or vancomycin (3). However, the impact of each antibiotic on each neonate is not uniform, leading to variability in post-antibiotic microbiome compositions (3). Many antibiotics have anti-anaerobic activity (11), preventing anaerobes critical for infant immune education from gaining a foothold in the preterm gut (13). Each additional day of ampicillin, cefotaxime, tobramycin, and/or metronidazole exposure in the NICU is associated with a 16% decrease in anaerobic species (13). Taken together, the specific antibiotic, duration, and patient gut microbiota composition interact, resulting in individualized responses to antibiotic interventions in the NICU. While the acute impact of antibiotics in the NICU have been studied, there remains, with few exceptions (2), a paucity of long-term follow up studies investigating health outcomes and lasting microbiota effects to early-life antibiotic exposures. Such studies would inform risk-benefit assessment of empiric therapies by pediatricians and potentially impact overall health outcomes in this vulnerable population.
A.6 The resistome in early life

Early life reflects a period of elevated resistome burden (Figure A.2), demonstrated by greater relative ARG abundance in the gut microbiomes of infants relative to those of their adult mothers (49, 50). At this age, efflux pumps often conferring multidrug resistance (MDR) comprise the majority of the resistome and – along with aminoglycosides, bacitracin, beta-lactams, colistin, quinolones, and sulphonamide resistance classes – are found at higher relative abundance compared to their mothers (49, 51, 52). The elevated ARG burden is most pronounced in the first weeks after birth and abates throughout the first year of life, primarily through decreasing abundance of efflux pumps (53).

The expanded early-life resistome is linked to the taxonomic composition of the gut microbiota (1–3, 49, 51, 53). Thus, the relative abundance of specific ARGs often rises rapidly during antibiotic treatment and falls immediately upon cessation, and such spikes are closely correlated with commensurate changes in the abundance of specific taxa (1, 3). *Escherichia coli* is thought to be a main contributor to the increased ARG burden in infancy, as its abundance strongly correlates with that of an infant’s resistome (2, 49, 51, 53). Indeed, a recent study found that *E. coli* is the putative bacterial host of 36 of the 50 most abundant ARGs in the infant microbiome(53). When infant gut microbiomes are categorized by either low or high ARG richness and diversity, *E. coli* abundance is the main variable distinguishing the two (51). 94% of the total abundance of the 58 most variable ARGs between the two groups were found on *E. coli* metagenome-assembled genomes (MAGs) (51), substantiating the disproportionate impact of *E. coli* on the infant resistome. Notably, a recent meta-analysis has demonstrated that microbiomes dominated by Staphylococci can have similarly high relative ARG abundances (54), highlighting that *E. coli* is not the sole correlate of elevated resistome...
Addendum: The resistance within burdens in early life. Moreover, it is important to note that the relative lack of genomic investigation of hard-to-culture gut microbiota may cause their contribution to the early-life resistome to be underestimated.

Determinants of infant gut resistome seeding and development are understudied (Figure A.2). Evidence for intergenerational transfer of antibiotic resistant strains, reviewed recently (55), remains scarce. There is observational evidence that antibiotic-resistant bacteria colonizing the maternal vagina, including extended-spectrum beta-lactamase producing and MDR E. coli, can be transferred to neonates during vaginal birth (55). In the first weeks of life, cesarean birth is associated with an increased resistome burden, specifically enriching glycopeptide, phenicol, pleuromutilin, bacitracin, sulfonamide, and diaminopyrimidine ARG abundance (53, 56). However, this association subsides in the first year of life (51, 56, 57), indicating that the impact of cesarean birth may be related to the altered dynamics of intergenerational microbiota transfer and the prolonged exposure to the hospital environment compared to vaginal-born neonates. Intrapartum antibiotic therapy, often suggested as another effector of resistome development, has not been shown to impact overall resistome abundance or composition (49, 53).

In early life, infants are either fed a diet of breastmilk, formula, or a combination of the two. During this time, formula feeding has been associated with a significantly greater relative resistome burden (54). Breastmilk has been found to contain antibiotic-resistant Bifidobacteria, Lactobacilli, Staphylococci, and Enterococci that can be transferred to infants (55). In fact, exposure to breastmilk has a sustained effect on the resistome composition throughout infancy (52, 57). Consequently, infant gut resistomes share ARGs with maternal
Addendum: The resistance within

the gut and breastmilk resistomes (49). However, studies differ in their assessment of the magnitude to which the maternal microbiome seeds the infant resistome (49, 50, 57), potentially due to contrasting methodological approaches. Future genome resolved-studies at the strain level that link ARGs to their bacterial host – for example using Hi-C technology (58) – can resolve this open question.

Early-life antibiotic exposure is not necessary for resistome development. Even antibiotic-naïve infants harbor ARGs and MDR bacterial strains in their gut microbiome in the first months of life (1, 2, 49). This may be a consequence of transgenerational transmission, colonization from environmental sources, or a survival-based response of the gut microbiota to antimicrobials produced by normal gut inhabitants (1, 49). Notably, ARGs conferring resistance to antibiotics not commonly used in neonatal populations are found in most infant gut microbiomes. Functional metagenomics identified ARGs against chloramphenicol and tetracyclines in a cohort of antibiotic-exposed preterm infants that were never exposed to these drugs during their NICU stay (3). A follow-up study also found chloramphenicol ARGs in the gut microbiome of fully antibiotic-naïve infants during the first two years of life, as well as resistance against ‘last line of defense’ drugs colistin and tigecycline in both antibiotic-exposed and -naïve infants (2). These observations have been confirmed in independent work (49, 51), highlighting that the resistome is a core characteristic of the infant gut microbiome rather than a feature that arises solely in response to antibiotic selection.

Nevertheless, ARG richness and resistome burden correlate with cumulative antibiotic exposure in the first weeks of life (2). Preterm infants who receive multiple courses of antibiotics during their NICU stay have a significantly higher resistome burden and richness relative to age-matched antibiotic-naïve near-term infants (2, 3). Further, their resistome
Addendum: The resistance within

composition is distinct from that of antibiotic-naïve infants (2), mostly driven by abundance changes of specific taxa in response to antibiotic exposure. Studies have reported *Klebsiella*-driven surges of beta-lactamase genes during penicillin exposure (1), as well as significant increases in ARGs as a consequence of *S. epidermidis* and *K. pneumoniae* blooms following meropenem and ticarcillin-clavulanate exposure, respectively (3), indicating that hospitalization and early-life antibiotics directly alter resistome composition. Importantly, if not repeatedly prescribed the effect of antibiotics on the pediatric resistome is likely to be transient, as the impact of even broad spectrum agents given in the first week after birth subsides by one year of life (59).

A.7 Multidrug resistant organisms in the infant gut

Early-life antibiotic exposure has been linked to persistent gut colonization by MDR organisms (Figure A.2), defined as bacteria resistant to more than three antibiotic classes (2, 48). As highlighted above, these may be seeded intergenerationally or acquired environmentally (47, 55). In preterm infants, prolonged hospitalization has been linked to the acquisition of Enterococci and *Enterobacteriaceae* strains from the hospital environment (47), many of which are MDR and can persist in the gut following hospital discharge (2). Rising antibiotic resistance among neonatal pathogens may thus further exacerbate the long term health impact of early life hospitalization (60). Notably, persistence of MDR bacteria throughout infancy was similarly observed in at least one antibiotic-naïve term infant (2), suggesting that hospitalization and early-life antibiotics may increase the likelihood of the initial acquisition but are not a prerequisite for MDR strain persistence.
Addendum: The resistance within

The presence of MDR pathobionts in the gut microbiome is associated with adverse health outcomes. Intestinal colonization with Enterococci and *Enterobacteriaceae* in the NICU and after discharge is associated with bacterial BSIs (46, 48), and similar strains of *Enterobacteriaceae* inhabit the gut prior to BSI (46). In pediatric short bowel syndrome, strains of BSI-causing Enterococci and *Enterobacteriaceae* can persist in the intestine late into childhood (48). During years of colonization, these strains can cause recurrent BSIs while persisting as minor members of the intestinal flora (48). While the unique intestinal characteristics and repeated antibiotic exposures accompanying short bowel syndrome likely facilitate pathobiont persistence, such findings illustrate the risk associated with intestinal pathobiont persistence in pediatric populations. However, while it can be assumed that MDR in intestinal pathobionts can have significant impacts on treatment outcomes, studies investigating the specific impact of MDR pathobionts acquired in early life on childhood health remain elusive. Furthermore, comprehensive identification of factors governing early-life acquisition and persistence of MDR organisms are necessary to decrease the risk of serious adverse health outcomes in infancy and throughout childhood.

A.8 Mobilization of resistance in the infant gut

Mobile genetic elements (MGEs) – including plasmids, transposons, integrons, and insertion sequences – are responsible for the spread of ARGs intracellularly and between bacterial cells in the gut microbiome (*Figure A.2*) (61), and may underlie the sustained increase in abundance of certain ARGs following early-life antibiotic exposure (1). Conflicting reports exist as to the extent to which maternal GM contribute to infant mobilome seeding. Mobilome composition is more similar between unrelated infants than mother-infant dyads and MGE
abundance and diversity in infants is increased relative to their mothers despite significantly reduced taxonomic diversity (49, 57), implying that taxa overrepresented in the infant gut disproportionately harbor MGEs. Indeed, Enterobacteriaceae, the predominating family and oversized contributor to resistome burden in the first months of life (3), are also frequent MGEs carriers, illustrating the intimate connection between the resistome and the mobilized genetic content, or ‘mobilome’, and shared bacterial hosts (2, 3, 49, 51, 53).

Spatial association between MGEs and ARGs in the infant gut is infrequent but can have broad clinical consequences. Functional metagenomic selection of infant stool metagenomes against 16 clinically relevant antibiotics found that 6.4% of ARGs were syntenic or located in close proximity to an MGE, commonly a transposase or integrase (2). This observation is supported by reference-based and predictive computational analysis of 20 infant gut metagenomes that found that 6-8% of ARGs are located on plasmids or bacteriophages (56). ARG-MGE associations allow for the mobilization of resistance that contribute to the rise of drug resistant infections, yet to date there is a dearth of studies comprehensively describing the relationship between specific ARG and MGE classes within drug-resistant strains in the infant gut microbiome.

Though tetracyclines are rarely prescribed during infancy, tetracycline resistance genes are frequently transferred intergenerationally and are potentially mobilizable (57). Tetracycline resistance genes, including tetX, are the most commonly shared ARG class between related mother-infant dyads (50, 57) and co-localize with MGEs in the infant gut microbiome (2, 49), highlighting their elevated potential for vertical and horizontal transfer. Inter-species transfer in the gut microbiome has been reported for multiple resistance clas-
Addendum: The resistance within

The resistance within 197 ses (56), and may be triggered by antibiotic selection (58). Transfer of ARGs in human microbiomes may in fact follow a predictable framework governed by the presence of key functions, including toxin-antitoxin and restriction modification systems (62). Recent research has shown frequent ARG transfer in the gut microbiome of adults (58), though similar analyses in pediatric cohorts have yet to be performed. Given its enrichment of ARGs, MGEs, and pathobionts the infant gut microbiome is a potential hotspot for the inter-species spread of antibiotic resistance, highlighting the need for and clinical importance of further investigation.

A.9 Future directions

In the face of surging resistance, even against antibiotics of last-resort in pathogenic bacteria inhabiting the infant gut, there is an urgent need to further our understanding of the early-life resistome. Here, we highlight aspects of resistome research with immediate clinical importance in pediatric populations.

Seminal research leveraging the power of high-throughput metagenomic sequencing coupled to large human cohorts have significantly advanced our understanding of the factors that govern microbiome assembly in infancy. Comparably, less is known about the resistome in early life. While multiple factors impact resistome seeding and composition in early life, including gestational age, birth mode, breastmilk, environment, and antibiotic exposures (2, 47, 49, 51, 53, 55), less is known about how they affect seeding and persistence of high-risk MDR pathogens in the pediatric gut (2). Recent work found that *E. coli* is the largest contributor to the gut resistome in healthy infants (51, 53), but lacked investigation of its initial sourcing or persistence of specific drug-resistant strains through infancy. Strain-resolved
Addendum: The resistance within

studies identifying sources as well as factors impacting persistence of high-risk MDR pathogens throughout childhood could enable development of novel interference strategies aimed at mitigating the risk associated with MDR pathogen gut colonization. Microbiota-humanized animal models may serve as a promising avenue to interrogate the impact of strain-resolved inter-bacterial interactions on resistome seeding and ARG spread in the gut microbiome.

The risk of individual ARGs for human health is closely tied to the pathogenic potential of its bacterial host, as well as its potential to facilitate resistance against antibiotics commonly used in empiric therapy. Therefore, we propose the need for comprehensive assessment of the risk associated with the presence and host-association of ARGs within the infant gut microbiome. Assigning risk scores based on frequency of observed treatment failures to host-ARG pairings could identify associations of high clinical relevance in diverse pediatric populations. Further, coupling this risk to predictions of ARG dissemination across bacterial species within patient-hospital transmission networks would allow the development of strategies limiting the clinical impact of high-risk ARGs. Recent work has highlighted that CRISPR-Cas9 based targeted genome editing using diverse delivery systems as a potentially revolutionary approach to microbiome manipulation (63). This system can facilitate strain-specific depletions and genomic deletions in vivo and thus has great potential for the targeted depletion of both MDR strains as well as high-risk ARGs and MGEs from the infant microbiome. Similarly, dietary interventions or fecal microbiota transplants could be used to alleviate infants of high resistome burdens and persistent low diversity microbiome states.

Finally, recent technological advances have opened new horizons in the field of rapid bacterial diagnostics. Groundbreaking work used the MinION sequencing platform for anal-
Addendum: The resistance within

ysis of fecal metagenomic DNA from preterm infants, identifying intestinal carriage of pathogens and their resistance profile in near real-time (64). This technology has the potential to inform treatment decisions specifically in the context of hospitalized preterm infants at risk of hospital-acquisition of MDR pathogens (3, 47). Together these studies and technological advances have the potential to revolutionize our understanding of the early life microbiome and resistome, its clinical importance, and equip us with tools to protect infant health in the face of rising resistance.
Addendum: The resistance within

A.10 References


Addendum: The resistance within


Addendum: The resistance within


Addendum: The resistance within


Addendum: The resistance within


Addendum: The resistance within


A.11 Figures

Figure A.1: Indications, acute and long-term effects of early-life antibiotic exposure on gut microbiome trajectory and health outcomes in term and preterm neonates. 40-70% of all term infants in high-income countries receive an antibiotic course in the first year of life, typically following community-acquired infection. Exposure results in a temporary decrease in gut microbiome diversity compared to antibiotic-naive infants. Most preterm infants receive empiric antibiotic therapy in the NICU, resulting in persistent low diversity microbiomes in the first year of life and sustained colonization with MDR pathobionts capable of causing infection. For all infants, antibiotic exposure in infancy is associated with adverse childhood health outcomes. Created with BioRender.com.
Addendum: The resistance within

Figure A.2: Resistome dynamics in infancy. The infant gut resistome, the collection of antibiotic resistance genes (ARGs) encoded by the gut microbiota, is seeded from maternal and environmental sources. Resistome burden is elevated in early life and compositionally different to infant mothers. Early-life antibiotics open niches for the colonization by pathobionts that are often multi-drug resistant (MDR) and associated with childhood infections, as well as increase resistome burden and abundance of gut endemic MDR bacteria. Further, antibiotic resistance is mobilized on plasmids, bacteriophages, or via transposons, integrons, and insertion sequences. These contribute to the spread of ARGs within gut communities, potentially triggered by antibiotic exposure. Created with BioRender.com.
Sanjam Sawhney

sanjsawhney@gmail.com
(925) 997-4353

1397 Whispering Oaks Way, Pleasanton, CA 94566

EDUCATION

Washington University in St. Louis, Molecular Microbiology & Microbial Pathogenesis GPA: 3.81
Ph.D. Candidate, September 01, 2018 – May 12, 2023
Research advisor: Gautam Dantas, PhD

University of California, Davis, Microbiology and Microbial Genetics GPA: 3.83
Bachelor of Science, received June 2017, magna cum laude

SKILLS

• Wet lab: Nucleic acid extraction; DNA library preparation and Illumina and Nanopore sequencing; aerobic and anaerobic bacterial culture; antibiotic susceptibility testing
• Computation: Genome/Metagenome assembly, annotation, and analysis on Linux/Unix; Bash scripting; supervised machine learning models; Rstudio; Prism; Microsoft Office suite

RESEARCH & PROFESSIONAL EXPERIENCE

Graduate Student Researcher – Dantas Lab, St. Louis, MO 63110 05/2019 – present
Constructed thousands of high-quality strain-resolved metagenome-assembled genomes from deep hybrid and culture-enhanced sequencing of infant twin pairs and their mothers to interrogate bacterial within-host evolution in the developing gut microbiome. Built a random forest classifier to identify genomic correlates of non-mec-mediated borderline oxacillin-resistance within Staphylococcus aureus isolates from the St. Louis Children’s hospital. Performed comparative genomics analyses of diagnostic and commensal Staphylococcus pseudintermedius isolates to identify niche adaptation through parallel selection of defense mechanisms. Co-wrote NIH R01 and CDC BAA grant applications, resulting in funding awards of $4.5M over 5 years for the Dantas Lab.
PI: Gautam Dantas, Ph.D., Conran Professor of Laboratory and Genomic Medicine, Departments of Pathology & Immunology, Pediatrics, Biomedical Engineering, and Molecular Microbiology

Freelance Consultant – BioGenerator Ventures & Missouri Technology Corporation 04/2021 – present
BioGenerator Ventures is the startup creation & investment arm of innovation hub, BioSTL. Its portfolio has raised $2.2B in follow-on funding. Missouri Technology Corporation is a public-private partnership with the state of MO to promote entrepreneurship and foster growth of emerging high-tech life science companies.

Worked intimately with Director Tom Cohen and EIR David Smoller (former CSO, Sigma Aldrich), using strong background in microbiome science to guide multi-million dollar investment decisions in next-generation sequencing CROs. Vetted scientific claims of start-up CEOs in real time. Presented internally and with external venture capital. Guided investment decisions ($100K - $2M / applicant) as an external reviewer for state-sponsored VC (IDEA Fund). 130+ paid hours.
Engagements: Due diligence, customer insights, mergers & acquisition, commercialization assessment

Research Associate – 10x Genomics, Pleasanton, CA 94566 09/2017 – 08/2018
Performed R&D experiments to optimize our Single Cell CNV solution investigating genome heterogeneity and clonal evolution prior to market release. Contributed to Verification & Validation efforts on the Single Cell Immune Profiling (Mouse) solution, affirming product’s ability to enrich for paired, full length V(D)J sequences from individual B and T cells, and simultaneously gather gene expression data.
PI: Joe Shuga, Ph.D., Senior Scientist

Lab Assistant – Baumgarth Lab, Center for Comparative Medicine, Davis, CA 95616 05/2016 – 08/2017
Investigated natural antibody production of B-1 cells sans antigen-specific stimulation. Analyzed the role of transcription factors on initiation of morphological changes leading to the antibody-secreting cell conformation. Examined CD5 downregulation in LPS-stimulated B-1a cells to characterize B-1 cell subsets and provided data to support hypotheses disputing current literature on the topic. Genotyped mice for future experiments.
PI: Nicole Baumgarth, DVM, Ph.D., Department of Pathology, Microbiology, and Immunology
ACADEMIC AND COMMUNITY SERVICE

Advisor, Director of Finance – Biology & Life Science Advising (BALSA) Group 10/2018 – 03/2022

BALSA is a student-led nonprofit organization that provides biotechnology consulting services to innovators, offices of technology transfer, biotech incubators, and small- and mid-size biotech companies. In 2021, BALSA did $100K in revenue.

Oversaw all organizational revenue and expenses, charitable donations, financial accounts, and 501(c)(3) tax filings. Generated the 2022 fiscal year Budget and 2021 Q1-Q4 Reports documenting quarterly revenue and spending. Negotiated contracts with clients prior to project initiation, defining feasibility and expected deliverables. Led teams of graduate student and postdoctoral consultants on targeted, six-week engagements. Engagements: Customer outreach and insights (x6), due diligence and strategy (x6) in healthcare, infectious disease, medical devices, microbiome, molecular biology, neuroimaging, & operations.

Teaching Assistant – Introduction to Phage Annotation, St. Louis, MO 01/2020 – 05/2020

Led two phage bioinformatics computer lab sections per week, teaching viral biology and computational tools for genome analysis, annotation, and gene expression. Assisted students with annotation of their phage genomes and taught key science communication strategies. Graded weekly assignments, presentations, and final reports, and proctored exams for 550 students. Adapted course curriculum and educational format during the COVID-19 school shutdown.

Tutor – UC Davis Student Academic Success Center, Davis, CA 09/2016 – 06/2017

Tutored Calculus and Biostatistics to UC Davis students. Practiced time-efficient approaches to effectively assist students in drop-in settings of up to 40 individuals. Developed strong listening and interpersonal skills to reach individuals of diverse student populations. Tailored tutoring strategies to effectively serve the largely differing needs of tutees.

PUBLICATIONS / MANUSCRIPTS IN PREPARATION

*co-first author

HONORS

- Vagelos Division of Biology & Biomedical Sciences Inaugural Leadership Award, Spring 2023
- BALSA Outstanding Consultant of the Quarter, Summer 2019
- 2017 Department Citation – Outstanding academic achievement in the Dept. of Microb. & Molec. Genetics
- Dean’s Honors List Recipient (Fall 2013, Fall 2014, Winter 2015, Spring 2015, Winter 2017)
- Shifa Community Clinic Outstanding Service Award, Spring 2017
- Phi Beta Kappa Honor Society inductee, Spring 2017