Spring 5-15-2023

The Microbes Within: Pathogen In-Host Adaptation and the Gut Microbiome During Persistent Colonization and Recurrent Infection

Joohee Choi

WASHINGTON UNIVERSITY IN ST. LOUIS
Division of Biology and Biomedical Sciences
Human and Statistical Genetics

Dissertation Examination Committee:
Gautam Dantas, Chair
Megan T. Baldridge
Andrew Kau
Jennie H. Kwon
Christina Stallings

The Microbes Within: Pathogen In-Host Adaptation and the Gut Microbiome During Persistent Colonization and Recurrent Infection

by
JooHee Choi

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
May 2023

Contents

List of Figures

List of Tables

Acknowledgements

This Thesis would certainly not have been possible without the help and encouragement of many wonderful individuals with whom I have had the exceptional luck of crossing paths with:

Thank you to my Evogengingym-mates Sudikchya Shrestha and Leonore Wünsche, and my genetics partner-in-crime Tina Kim for encouraging my scientific pursuits and believing in my capabilities when I didn't believe them myself.

Thank you to my early scientific mentors Michael Purugganan, Jae Young Choi, Stephane Boissinot, Robert Ruggiero, and Youssef Idaghdour. You set me forward in my path.

Thank you to my HSG program coordinators Jeanne Silvestrini and Sara Holmes, for handling all the logistical paperwork involved in a PhD. Jeanne was the first person who welcomed me to St. Louis and made me feel like it could become my home.

Thank you past and present HSG program directors John Rice, Patrick Jay, and Nathan Stitziel, as well as HSG faculty Nancy Saccone and Ting Wang for admitting me into the program. Ting, I will never forget receiving that phone call!

Thank you to my thesis committee: Christina Stallings, Jennie Kwon, Megan Baldridge, and Andrew Kau. Your feedback and encouragement kept me going. To the Pathogenomics group, especially Carey-Ann Burnham, Jennie Kwon, Erik Dubberke, Meghan Wallace, Kimberly Reske, Tiffany Hink, and Katelin Nickel: I am grateful for the opportunity to explore Pathogenomics with you. Our meetings inspired me and

showed me the type of science I wanted to pursue. Thank you for providing valuable insight and pushing me to see things from a clinical, as well as genomic, perspective.

And of course, thank you to Gautam Dantas for being an excellent mentor. You are *really* good at your job. Somehow, I ended up in your lab, even though microbiology wasn't even on my mind when I moved to St. Louis. Thanks for trusting in my potential, listening to me (every single week at mentoring meeting), having my back (except when playing competitive boardgames), and overall just being a reliable source of honest feedback and guidance. We are lucky to have you.

Thank you to everyone in the Dantas Lab for collectively being the best lab I could have asked for. The sense of community and belonging helped me carry on when I was down. Kimberley Sukhum, Alaric D'Souza, and Robert Potter took the time to show me the reigns when I was starting out. It was their kind mentorship, as well as watching how they balanced hard work with play, that weighed in on my decision to join the lab. My closest collaborators in the lab, Robert Thänert and Suryang Kwak, taught me everything I know about what it takes to publish a manuscript from start to finish. Thank you for your patience and diligence. Skye Fishbein, Miranda Wallace, Drew Schwartz, and Aura Ferreiro are four exceptional scientists in the lab who answered my random questions 24/7. Bin Wang and Jie Ning keep the lab running smoothly. Without you I imagine the lab might spontaneously combust (just kidding). James Liao lent a hand whenever I needed help processing samples. I hope we can both keep our Duolingo streaks going *ad infinitum*.

Thank you to members of the Edison Family Center for Genome Sciences and Systems Biology: Bonnie Dee and Kathleen Matheny, for your logistical help and your warm presence. Jess Hoisington-Lopez and MariaLynn Crosby helped me balance every pool of sequencing I ever ran. They took the time to help me troubleshoot issues whenever they occurred- I would have no data without them. Eric Martin and Brian Koebbe maintain our cluster and have helped me fix countless error messages that have popped up over the years.

To my lab buddies past and present: Rhiannon Vargas, Sanjam Sawhney, Winston Anthony, Olivia Gorushi, Galen Wong, Kevin Blake, Erin Newcomer, Bejan Mahmud, Anna DeVeaux, Luke Diorio-Toth, Kailun Zhang, and Esse Aigbokhan- I cherish our shenanigans in and out of lab. I wish you the best of luck in your future endeavors and look forward to hearing all about them. I will miss you!

Thank you to my graduate school friends Julia Wang, Jeongmin Lee, and Mandy Chan (aka The Cat Ladies) for being there for me through all the trials and tribulations that go into a PhD and inspiring me to become a better biologist. My HSG friends, especially Kara Quaid and Sharon Freshour, for going through this endeavor with me (and being the best neighbors). My NYUAD friends Yoonhoo Chang, Jian Ryou, and Ariya Chaloemtoem, for trusting my decision to come here enough to follow suit. I am proud to be a "great-grandmother". And my Korean community, who gave me moments of home in a foreign land.

Thank you to Mom and Dad for your endless support, I am blessed to be your kid. Thank you to my brother, Jae Hyuk, for staying true to who you are.

Thank you to Jason for being with me all this time. From long distance from Chicago to St. Louis, to even longer distance (near-antipodal!). It's incredible what we've been through, and I am excited for what lies ahead.

Finally, I dedicate this Thesis to my cats: Miu Miu, Gustav, and Sammy.

JooHee Choi
St. Louis, MO
May 2023

ABSTRACT OF THE DISSERTATION

The Microbes Within: Pathogen In-Host Adaptation and the Gut Microbiome During Persistent Colonization and Recurrent Infection

by
JooHee Choi

Doctor of Philosophy in Biology and Biomedical Sciences
Human and Statistical Genetics
Washington University in St. Louis, 2023
Professor Gautam Dantas, Chair

Microbes not just surround us; they are inside of us. The gut microbiome has emerged in recent years as an important modulator of health and is thought to have co-evolved with us throughout evolutionary history. On the other hand, our immune systems are constantly surveilling and battling infection by external pathogens— some of which still manage to evade our immune response and colonize our bodies long-term. In this Thesis, I investigated microbes inhabiting our bodies in various contexts to understand their impact on human health.

In Chapter 2, I discuss the effects of fecal microbiome transplant (FMT) study drug RBX2660 on the gut microbiome of recipients with *Clostridium difficile*. I parsed apart post-antibiotic microbiome recovery from true RBX2660 effects and assessed what characteristics deemed some recipients more permissive for microbiome transplantation over others. I found that RBX2660 administration transplanted healthy microbiota in the recipients in a dose-dependent manner. *Veillonella atypica* and intrinsic vancomycin resistant species were discriminative features of patients showing long-lasting microbiota transplantation and resisting microbiota transplantation, respectively. RBX2660 more efficiently decolonized antimicrobial resistant organisms (AROs) than

placebo but simultaneously introduced new AROs. This study demonstrated the potential benefits of FMT and highlighted the importance of the design and quality control of microbiota-based drugs.

In Chapter 3, I report evidence for in-host adaptation in a cohort of longitudinally collected *Mycobacterium abscessus* isolates. Through comparative genomics, I demonstrated the presence of clusters of highly related isolates from multiple hospital centers despite lack of evidence for transmission. I also identified within-lineage polymorphisms occurring in parallel across multiple patients, suggestive of shared adaptative behavior to survive in the lung milieu. Through drug susceptibility assays I show that the genomic changes have phenotypic consequences, potentially providing opportunistic windows for effective treatment.

In Chapters 4 and 5, I share the findings from a multi-center cohort of participants with urinary tract infection (UTI). Some of these patients experienced recurrent UTI (rUTI) throughout the study period. Through regular stool and urine sampling, I obtained longitudinal gut microbiome and isolate WGS data. By tracing lineages of UPEC, I demonstrated four patterns of asymptomatic colonization: in the urinary tract, in the gastrointestinal tract, in both habitats, and no colonization. I then utilized comparative genomics to show niche-specific adaptive patterns that were putatively facilitated by mobile genetic elements.

Finally, in the last chapter I discuss the results of a clinical model predicting risk factors for rUTI. Recent antimicrobials and steroids elevated rUTI risk, while change of antimicrobials and TMP-SMX were associated with decreased risk. I also found

significant differences in gut microbiome composition between urinary tract colonized and non-colonized patients at post-antimicrobial days 7-14, marked by elevated *E. coli* abundance in urinary tract colonized patients. Together these studies explored in-host behavior of UPEC across two distinct habitats and point towards the gut as an important reservoir facilitating rUTI.

Chapter 1

1.1 Introduction

The number of microbes on our planet is said to outnumber the stars in our galaxy[1]. Even in the human body, microbes outnumber us: the estimated total number of microbial cells is 39 trillion compared to 30 trillion human cells[2]. These figures are difficult to imagine— even more so considering most microbes are indiscernible to the naked eye. Yet our lives are intricately intertwined with microbiota, from the yeast used to make our beers to the bacteria that ferment yogurt, cheese, and kimchi. Microbes also play an essential role in our health, though this relationship is anything but straightforward.

Throughout human history, civilizations have been plagued with various well-documented infectious outbreaks such as smallpox, leprosy, and cholera. With the understanding that these diseases were spread through contagion, medical scientists developed and utilized tools to isolate and study the causative pathogens. The invention of microscopes, culture plates, vaccines, and pasteurization throughout the centuries paved the way for microbiology as we know it today[3]. However, it was Paul Ehrlich's synthesis of what is broadly considered the first antimicrobial, salvarsan, which kickstarted antimicrobial discovery in the West in 1910[4]. In 1928, Alexander Fleming discovered penicillin, and penicillin was purified and developed as a drug in widespread clinical use by the 1940s[4]. The following two decades are often referred to

as the Golden Age of antimicrobials, and antimicrobials are credited with having added 23 years to the average human lifespan[4].

By the 1970s, infectious disease was considered by some to be largely conquered[5]. Humans had figured it out. Looking back, the hubris is amusing. We were overlooking a huge keystone in biology: evolution. In fact, antimicrobial resistance (AMR) was discovered soon after introduction of antimicrobials in the clinic[6], and rates of AMR rose quickly while rampant antimicrobial use went unchecked[4,6]. Coupled with a decline in the rate of novel drug discovery, the battle against "superbugs" with multidrug resistance has become one of the most urgent global health concerns today[7,8].

Yet microbes are not always at odds with human health. Microbes play important roles in the healthy functioning of our bodies, from assisting food digestion[9] to modulating the immune system[10]. A disruption to the gastrointestinal community of microbes (the gut microbiome) has been found to be associated with an increasing number of health problems, ranging from gut dysfunctions such as inflammatory bowel disease[11], to recurrent infections such as by *Clostridium difficile (*rCDI)[12] or uropathogenic *Escherichia coli* (UPEC)[13], to even neurological disorders such as Alzheimer's Disease[14]. Novel therapeutics such as prebiotics, probiotics, and as fecal microbiome transplant (FMT) are attempting to address these health problems from the root— or rather, the gut.

So how do we parse apart the "good" microbes from the "bad"? What distinguishes benign coexistence from infection? One key development that has advanced our ability to answer these questions is the advent of cheap shotgun

sequencing. Whole genome sequencing (WGS) now allows us to compare isolate genomes, identify antimicrobial resistance genes and virulence factors, as well as track the transmission routes of specific strains of pathogens[15]. WGS of longitudinally collected isolates can provide information on how the strain is adapting in response to its surroundings, and thus illuminate within-host adaptation in persistent pathogens[15]. On the other hand, shotgun metagenomic sequencing allows entire microbial communities to be assessed: by taxonomic diversity (alpha diversity), comparisons between communities (beta diversity), interrelatedness of taxa within communities (network analysis), and predicting metabolic pathways. Defining microbiome features of specific disease states may also be characterized, thus providing the means for a cheap biomarker in place of expensive diagnostics, and microbiome-targeting interventions.

The ongoing COVID-19 pandemic vividly illustrates the utility of sequencing in epidemiology. The first complete genome of 2019-nCoV was announced in January 2020 from metagenomic RNA sequencing of bronchoalveolar lavage fluid from a patient, and was quickly identified to belong to the virus family *Coronaviridae* via comparative genomics[16]. As the virus spread globally, subsequent variants of the virus were also swiftly annotated, and surveillance of wastewater became established an effective means of monitoring case numbers[17]. As climate change progresses and more infectious diseases emerge[18], sequencing will continue to be an indispensable tool in tackling future outbreaks.

In summary, microbial genomics has the potential to elucidate the complex and nuanced relationship between human health and microbes. To this end, I leveraged bacterial genomics and metagenomics in four longitudinal studies throughout my Thesis. In the first study, I investigated the dose-specific effects of an FMT study drug RBX2660 on recipient rCDI microbiomes. In the second study, I explored how the environmental saprophyte *Mycobacterium abscessus* can opportunistically infect and colonize the lungs of predisposed individuals via in-host adaptation. The third and fourth study both tracked a multi-center cohort of individuals with urinary tract infection (UTI) to observe which individuals experienced recurrence (rUTI). I traced lineages of UPEC throughout episodes of rUTI and annotated the habitat-specific in-host adaptation that occurred. Finally, I examined the gut microbiome and clinical metadata to determine what characteristics may elevate risk of rUTI. Collectively, this Thesis shares representative work from my PhD, and explores the fields of infectious disease, microbial genomics, and evolution.

1.2 References

1. Suttle CA. Viruses: unlocking the greatest biodiversity on Earth. *Genome*. 2013;56(10):542-544. doi:10.1139/gen-2013-0152

2. Sender R, Fuchs S, Milo R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLOS Biology*. 2016;14(8):e1002533. doi:10.1371/journal.pbio.1002533

3. Brachman PS. Infectious diseases—past, present, and future. *International Journal of Epidemiology*. 2003;32(5):684-686. doi:10.1093/ije/dyg282

4. Hutchings MI, Truman AW, Wilkinson B. Antibiotics: past, present and future. *Current Opinion in Microbiology*. 2019;51:72-80. doi:10.1016/j.mib.2019.10.008

5. Lederberg J. Infectious History. *Science*. 2000;288(5464):287-293. doi:10.1126/science.288.5464.287

6. Alekshun MN, Levy SB. Molecular Mechanisms of Antibacterial Multidrug Resistance. *Cell*. 2007;128(6):1037-1050. doi:10.1016/j.cell.2007.03.004

7. Antibiotic resistance. Accessed December 7, 2022. https://www.who.int/news-room/fact-sheets/detail/antibiotic-resistance

8. CDC. The biggest antibiotic-resistant threats in the U.S. Centers for Disease Control and Prevention. Published July 15, 2022. Accessed December 7, 2022. https://www.cdc.gov/drugresistance/biggest-threats.html

9. Oliphant K, Allen-Vercoe E. Macronutrient metabolism by the human gut microbiome: major fermentation by-products and their impact on host health. *Microbiome*. 2019;7(1):91. doi:10.1186/s40168-019-0704-8

10. Hooper LV, Littman DR, Macpherson AJ. Interactions between the microbiota and the immune system. *Science*. 2012;336(6086):1268-1273. doi:10.1126/science.1223490

11. Halfvarson J, Brislawn CJ, Lamendella R, et al. Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat Microbiol*. 2017;2(5):1-7. doi:10.1038/nmicrobiol.2017.4

12. Gilbert JA. Microbiome therapy for recurrent Clostridioides difficile. *The Lancet Microbe*. 2022;3(5):e334. doi:10.1016/S2666-5247(22)00096-9

13. Worby CJ, Schreiber HL, Straub TJ, et al. Longitudinal multi-omics analyses link gut microbiome dysbiosis with recurrent urinary tract infections in women. *Nat Microbiol*. 2022;7(5):630-639. doi:10.1038/s41564-022-01107-x

14. Sochocka M, Donskow-Łysoniewska K, Diniz BS, Kurpas D, Brzozowska E, Leszek J. The Gut Microbiome Alterations and Inflammation-Driven Pathogenesis of Alzheimer's Disease—a Critical Review. *Mol Neurobiol*. 2019;56(3):1841-1851. doi:10.1007/s12035-018-1188-4

15. Blake KS, Choi J, Dantas G. Approaches for characterizing and tracking hospital-associated multidrug-resistant bacteria. *Cell Mol Life Sci*. 2021;78(6):2585-2606. doi:10.1007/s00018-020-03717-2

16. Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. *Nature*. 2020;579(7798):265-269. doi:10.1038/s41586-020-2008-3

17. CDC. National Wastewater Surveillance System. Centers for Disease Control and Prevention. Published May 16, 2022. Accessed December 7, 2022. https://www.cdc.gov/healthywater/surveillance/wastewater-surveillance/wastewater-surveillance.html

18. Climate change and infectious diseases | What We Do | NCEZID | CDC. Published August 2, 2022. Accessed December 7, 2022. https://www.cdc.gov/ncezid/what-we-do/climate-change-and-infectious-diseases/index.html

Chapter 2

Impact of investigational microbiota therapeutic RBX2660 on the gut microbiome

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

Kwak S*, **Choi J***, Hink T, et al. Impact of investigational microbiota therapeutic RBX2660 on the gut microbiome and resistome revealed by a placebo-controlled clinical trial. *Microbiome*. 2020;8(1):125. doi:10.1186/s40168-020-00907-9

* = equal contribution

2.1 Abstract

Intestinal microbiota restoration can be achieved by replacing a subject's perturbed microbiota with that of a healthy donor. In this study, we investigated fecal specimens from a multicenter, randomized, double-blind, placebo-controlled phase 2b study of microbiota-based investigational drug RBX2660. Patients were administered either placebo, 1 dose of RBX2660 and 1 placebo, or 2 doses of RBX2660 via enema and longitudinally tracked for changes in their microbiome and antibiotic resistome. Antibiotic discontinuation alone resulted in significant recovery of gut microbial diversity and reduced ARG abundance, but RBX2660 administration more rapidly and completely changed recipient microbiomes. We identified 18 taxa and 21 metabolic functions distinguishing the baseline microbiome of non-transplanted patients. Most features were correlated to intrinsic vancomycin resistance. We also identified 7 patient-specific and 3 RBX2660-specific ARGs and tracked their dynamics post-treatment. Whole genome sequencing of AROs cultured from RBX2660 product and patient samples indicate ARO eradication in patients via RBX2660 administration, but also, to a lesser extent, introduction of RBX2660-derived AROs.

2.2 Introduction

Intestinal microbiota restoration by microbiota-based therapy, such as fecal microbiota transplantation (FMT) from healthy donors to patients, has been applied as a treatment for disorders caused by intestinal dysbiosis[1]. As the contributions of the gut microbiota to the host immune system, energy metabolism, and central nervous system have been

uncovered, the range of potential applications of intestinal microbiota restoration therapy is expanding to various disorders, such as inflammatory bowel disease[2], functional gastrointestinal disorders[3], metabolic syndrome[4,5], and neuropsychiatric disorders[6,7]. Accordingly, studies for understanding and refining the action of intestinal microbiota restoration therapies are being actively conducted[8].

Clostridioides difficile infection (CDI) is one area where intestinal microbiota restoration therapy has been applied successfully. Although oral administration of antibiotics is the standard first-line therapy for CDI, antibiotics perturb the commensal gut microbiota and decrease colonization resistance against other pathogens[9,10]. Approximately 15% to 30% of CDI patients therefore experience recurrent CDI (rCDI) resulting from either a relapse of the previous CDI or reinfection[11]. Moreover, antibiotic therapies during CDI treatment may promote the expansion of antibiotic resistant organisms (AROs) such as vancomycin resistant *Enterococci* (VRE)[12,13]. On the other hand, intestinal microbiota restoration has shown to be effective for CDI treatment as well as the restoration of colonization resistance against *C. difficile* and AROs[14,15]. Indeed, intestinal microbiota restoration has become a commonly performed investigational therapy for rCDI with decent success rates[8,16–19].

However, due to the transmissive nature of the treatment, microbiota restoration therapy may communicate not only desirable but also undesirable factors derived from donors. For instance, the transmission of antibiotic resistant genes (ARGs) and AROs derived from donor samples is a potential risk of fecal transplantation[20,21]. AROs are responsible for increasing infection cases each year, and more than 35,000 patients died

as a result of ARO infections in the United States in 2017[22]. Recently, two cases of bacteremia caused by extended-spectrum beta-lactamase (ESBL)-producing *Escherichia coli* in patients after FMT from the same donor sample have been reported, resulting in the death of one of the patients[21]. Moreover, the dissemination of ARGs and pathogenic AROs in patients hampers effective medical care of infections and results in longer hospitalization and higher medical expenditures[23]. Still, multiple studies report efficient reduction of ARGs and decolonization of AROs through microbiota transplantation[24,25].

In this study, we explored the effect of a microbiota-based investigational drug RBX2660, a suspension of healthy donor microbiota[26–29], on the intestinal microbiome and resistome of recipients treated for rCDI. In an international, multicenter, randomized, and blinded phase 2b study, rCDI patients received either placebo (control group), one dose, or two doses of RBX2660 (Figure 2.1), with more patients being recurrence-free after either RBX2660 regimen than placebo[26]. Through shotgun metagenomic sequencing, we demonstrate considerable shifts of taxonomic and resistome structures common to both placebo- and RBX2660-treated patients likely from discontinuation of antibiotics, particularly during the first week after treatment. By controlling for placebo effects, we could also distinguish taxonomic and resistome changes specific to RBX2660 treatment. Furthermore, we identified discriminative features strongly correlated with microbiota transplant and demonstrated an overall decrease in AROs as well as introduction of a few AROs by RBX2660.

2.3 Results

2.3.1 Study cohorts and sample collection

All donors of RBX2660 microbiota completed a comprehensive initial health and lifestyle questionnaire. Their blood and fecal samples were tested for immunodeficiency viruses, *C. difficile* toxin, and pathogens including AROs such as VRE and methicillin-resistant *Staphylococcus aureus* before enrollment into the donor program[27,28]. Fecal specimens from a total of 66 patients and their corresponding RBX2660 products were collected during a multicenter, randomized, blinded, and placebo-controlled phase 2b study for the treatment of rCDI (Figure 2.1)[26]. 94% of all patients (62/66) had received vancomycin, with the remainder receiving metronidazole or fidaxomicin prior to study drug (Figure 2.1). 21 patients received 2 doses of placebo (14 females, 9 CDI recurrence, median age 63 years), 22 patients received 1 dose of RBX2660 and 1 dose of placebo sequentially (15 females, 5 CDI recurrence, median age 63 years), and 23 patients received 2 doses of RBX2660 (15 females, 8 CDI recurrence, median age 68 years)[26]. Each RBX2660 dose derives from a single donor, and RBX2660 dose selection was not constrained to ensure a single donor was represented in patients that received two RBX2660 doses (Table 2.1). The first dose of study drug (RBX2660 or placebo) was administered 24–48 hours following completion of antibiotic treatment for CDI, and the second treatment was administered 7 ± 3 days later (Figure 2.1). Patients who experienced a new rCDI episode within 60 days after the first dose (9 placebo recipients, 5 single RBX2660 recipients, 8 double RBX2660 recipients) were moved to open-label treatment and received two additional doses of randomized RBX2660 (Figure 2.1). Patient fecal specimens were collected at selected time points from baseline (day 0)

through 365 days after the first dose. AROs from each fecal sample were isolated on selective media plates.

2.3.2 RBX2660 shifted taxonomic structures of patients' intestinal microbiome in a dose-dependent manner

rCDI patients had significantly lower alpha diversity (Shannon diversity) than RBX2660 products before the treatment (Fig 2a), as previously described with 16S sequencing[29]. Following study drug administration, the alpha diversity of all rCDI patients' microbiota increased to near-RBX2660 levels regardless of the treatment group, with the steepest increase during the first week (Figure 2.2b). The largest taxonomic structural shift also occurred during the first week in all treatment groups (Figure 2.3 and Figure 2.4).

Bray-Curtis dissimilarities between recipient and corresponding RBX2660 product were calculated to assess the level of taxonomic transformation towards that of RBX2660. For placebo recipients, the dissimilarity was measured from a pseudo-donor (DS00) profile calculated from the average species-level taxonomic profile of all RBX2660 products in this study (Figure 2.2c). The mean Bray-Curtis dissimilarity of DS00 from RBX2660 products was 0.4926, which was lower than the inter-RBX2660 Bray-Curtis distance of 0.6274. Considering the thorough inspection criteria for donors of RBX2660 products, we defined RBX2660 microbiomes as "unperturbed" gut microbiomes. Bray-Curtis dissimilarities between patients and RBX2660 demonstrate that RBX2660 administration effectively changed recipients' microbiome structure

towards unperturbed configurations at a larger magnitude and for a longer duration as compared to placebo (Kruskal-Wallis test, P=0.043 at day 30, P=0.028 at day 60, Figure 2.2d). These microbiome shifts by RBX2660 were not sensitive to the kind of antibiotic administered prior to RBX2660 (Figure 2.5).

We further compared the original Bray-Curtis dissimilarities between patients and respective RBX2660 (D_R) to dissimilarities between patients and other random RBX2660 (D_O). RBX2660 recipients still exhibited lower D_Os than those of placebo recipients in dose-dependent manner (Figure 2.6), indicating that RBX2660 shifted patients' gut microbiomes toward an unperturbed microbiome more actively than placebo. In addition, significantly lower D_Rs than D_Os of double dose recipients after the RBX2660 administration demonstrated dose-dependent and specific shifts toward corresponding RBX2660 (Figure 2.6). Principal coordinates analysis (PCoA) and PERMANOVA for patients and RBX2660 also indicated that placebo recipients did exhibit taxonomic structural shifts toward RBX2660, but they were not as dramatic as those of double RBX2660 dose recipients toward the first dose RBX2660 (Figure 2.2e).

When comparing groups based on rCDI treatment success, treatment-failure patients (who experienced a new rCDI episode within 60 days post-treatmemt) and treatment-success patients did not exhibit significant differences (Figure 2.7a−c). This is likely due to limited number of treatment-failure samples after baseline, as patients were omitted from the current blinded study for the standard-of-care treatment at failure determination. Thus, we performed general linear model-based multivariate statistical analyses of patient baseline metagenomes using MaAsLin2[30] to identify

baseline features correlated to rCDI prevention success or failure. *Klebsiella pneumoniae* was the only species whose relative abundance was significantly associated with treatment failure in all patients (Figure 2.7d). When patients were grouped by RBX2660 dose, the model identified *K. pneumoniae* as the only potential failure-associated feature again from placebo recipients (Figure 2.7e) but did not from RBX2660 recipients.

2.3.3 RBX2660 transplanted taxonomic structures to patients

To quantify and compare patients' levels of change in microbiome composition, we calculated a transplantation index quantifying the extent of microbiome convergence towards corresponding RBX2660 product. This index was defined as the change in Bray-Curtis distances between baseline ($Distance_{BL}$) and selected time point ($Distance_{T}$), scaled by the distance from RBX2660 at baseline: ($Distance_{BL}$-$Distance_{T}$)/$Distance_{BL}$. DS00 was used for placebo recipients, who were then used to determine taxonomic transplantation success. To validate the transplantation index as a metric for quantifying microbiome shifts by RBX2660, we also calculated pseudo transplantation indices using dissimilarities between patients and random, non-corresponding RBX2660 products and compared them with the original transplantation indices. The dose-dependent increase in pseudo indices (Figure 2.8) is additional evidence that RBX2660 shifted patients' intestinal microbiomes toward the unperturbed microbiome of RBX2660. Some of the pseudo indices were lower than zero, indicating that the transplantation index well reflects individual directionality of recipient's microbiome shift toward respective RBX2660 (Figure 2.8). Statistically significant differences

between the original and pseudo transplantation indices of double dose recipients, but not single dose (Figure 2.8), connoted that double dose administration allows more RBX2660-specific microbiome shift than single dose.

RBX2660 recipients were categorized as transplanted or non-transplanted based on whether their transplant index was higher (transplanted) or lower (non-transplanted) than the maximum value of the placebo group (Fig 3a). The transplantation ratio trended higher in double dose recipients versus single dose recipients; this categorization showed 33.3% and 70.6% transplantation for single and double dose recipients, respectively, by day 7 (Chi-square test, P=0.02752), and 29.4% and 58.3% by day 60 (Chi-square test, P=0.1212). Non-transplanted patients at day 7 maintained non-transplanted status until day 60, regardless of dose. On the other hand, 1 single dose recipient (R1-21) and 3 double dose recipients (R2-01, R2-03, and R2-14) failed to maintain their transplanted state at day 7 until day 60 and eventually reverted to below the transplantation threshold. *Veillonella atypica* was the only baseline taxonomic feature determined by Linear discriminant analysis Effect Size (LEfSe)[31] that distinguished patients with successful microbiome transplantation by day 60 from non-transplanted patients in both single and double RBX2660 treatment arms (Figure 2.9b).

Although double RBX2660 dosage led to more effective transplantation of RBX2660 microbiome structure, there were 4 double-dose recipients (R2-01, R2-02, R2-03, R2-14) who showed lower transplantation indices than placebo recipients at day 60 (Figure 2.9a and Figure 2.10a). All 4 patients received vancomycin prior to RBX2660 administration (Figure 2.1). We determined 18 taxa (Figure 2.9c) and 21 functions

(Figure 2.10b) as features specifically explaining the baseline microbiome of these 4 patients by comparing with other double-dose recipients that showed durable taxonomic transplantation by day 60 using LEfSe [31]. Of these, 4 taxonomic features were fungi, which are intrinsically vancomycin insensitive, and 7 functional features of eukaryote-specific metabolic pathways (Figure 2.9c and Figure 2.8b). We further investigated the predicted vancomycin insensitivity of other taxonomic features and found 8 additional intrinsically vancomycin resistant bacteria including *Pediococcus* strains[32–34], *Lactobacillus* and *Leuconostoc* strains[35–37] as well as gram-negative and fungal strains. *Enterococcus casseliflavus*, which has low level resistance to vancomycin, was also identified[38]. Four taxa (*Clostridium glycolicum*[39], *Gemella haemolysans*[40], *E. faecalis*[41], and *C. difficile*[42]) are predicted to be vancomycin susceptible. Compared to the transplanted patients, the 4 non-transplanted patients did not exhibit any other distinctive taxonomic characteristics in terms of alpha diversity and composition of *Bacteroidetes*, *Firmicutes*, and *Proteobacteria* phyla (Figure 2.10c−g).

Beyond baseline features, we further investigated which taxa were enriched during the process of transplantation. Through a two-part zero-inflated Beta regression model with random effects (ZIBR)[43], we investigated a subset of 12 patients (R1-02, R1-03, R1-09, R1-14, R1-21, R2-05, R2-06, R2-10, R2-11, R2-12, R2-13, and R2-20) matched for 4 different timepoints: baseline, day 7, 30, and 60. ZIBR models a taxon's presence and absence (logistic component) as well as its non-zero abundance (Beta component), while incorporating patient and time as random variables (random intercepts). Only two genera, *Barnesiella* and *Coprobacillus*, were significantly correlated with the taxonomic

transplantation. *Barnesiella* was significantly overrepresented in the transplanted patients as early on as day 7, while *Coprobacillus* was overrepresented in non-transplanted patients at days 30 and 60 (Figure 2.9d). At the species level, ZIBR models identified *Barnesiella intestinihominis*, *Coprobacillus* (unclassified), *Bacteroides ovatus*, *Bacteroides uniformis*, *Ruminococcus obeum*, and *Akkermansia muciniphila* (Figure 2.9e, *A. muciniphila* was omitted because its time point comparisons were not statistically significant in the actual data). *Barnesiella intestinihominis* and unclassified *Coprobacillus* species followed near-identical patterns from the genus-level analysis due to single species being identified from each genus.

2.3.4 Resistome regression significantly correlated with transplantation index

Prior to treatment, rCDI patients showed a similar resistome alpha diversity (Wilcoxon signed-rank test, P=0.18, Figure 2.11a) when ARGs were grouped into ARG families based on the organizational structure in CARD[44]. However, the relative abundance of total ARGs was significantly higher in the patients than RBX2660 (Wilcoxon signed-rank test, $P < 0.0001$, Figure 2.11b). It decreased over time in all treatment arms including the placebo group (Figure 2.11c). Patients' resistome composition was distinct from RBX2660 products, but the antibiotic treatment prior to study drug administration did not lead to noticeable difference in resistome (Figure 2.12a−c). Specifically, major facilitator superfamily (MFS) and resistance-nodulation-cell division (RND) efflux pumps were the major ARG families present in rCDI patients before the treatment, whereas CfxA beta-lactamase, tetracycline-resistant ribosomal protection proteins, and

Erm 23S rRNA methyltransferases were representative of the RBX2660 resistome (Figure 2.12d).

We tracked individual changes in resistome composition of each patient for 60 days using t-Distributed Stochastic Neighbor Embedding (t-SNE) analysis[45] and resistome transplantation indices defined analogously to the microbiome transplantation index. rCDI patients showed distinctive resistome compositions as compared to those of RBX2660 prior to the treatment, but over time their resistome compositions converged to become similar to RBX2660 (Figure 2.11d). The speed of resistome transformation toward RBX2660-like structures varied by patient. The convergence toward RBX2660 resistome structure showed strong correlation to the taxonomic transplantation irrespective of treatment arm ($R^2 = 0.406$, $P < 0.0001$, Figure 2.11e). RBX2660 administration led to higher taxonomic and resistome transplantation indices than the placebo (Figure 2.11e).

To identify features distinguishing patient and RBX2660 resistomes, we used a Random Forest classifier (Figure 2.13a−b). Of the top 10 features of importance, 7 ARGs, namely MFS efflux pump, RND efflux pump, OXY β-lactamase, Pmr phosphoethanolamine transferase, undecaprenyl pyrophosphate related proteins, ATP-binding cassette (ABC) efflux pump, small multidrug resistance (SMR) efflux pump, and tetracycline resistant ribosomal protein, were specific to patient baseline resistomes. Class A β-lactamases (CfxA and CblA) and a tetracycline resistance protein, which are frequently identified in healthy populations or donor stools in FMT trials[20,46–49], were classified as RBX2660-specific ARGs (Figure 2.14a). Relative abundances of all selected

ARGs were significantly altered in recipients one week after study drug administration (Figure 2.14b−k). The regression of patient-origin ARGs occurred in all patients without statistically significant differences among placebo and RBX2660 recipients (Figure 2.14b−h and Figure 2.13c−i). Administration of RBX2660 increased relative abundances of RBX2660-origin β-lactamases in a dose-dependent manner (Figure 2.14i and 5j), while the relative abundance of tetracycline resistant ribosomal protection protein increased in all patients irrespective of treatment (Figure 2.14k).

2.3.5 RBX2660 effectively cleared AROs compared to placebo but introduced new AROs

We identified both persisting and newly introduced AROs based on whole genome sequence analyses of isolates from both blind and open-label treatment patients. ARO isolates were *Escherichia coli* (n = 104), vancomycin-resistant *Enterococcus* (VRE) (n = 25) and other species (n = 135). The majority of RBX2660-derived AROs were *E. coli* (Figure 2.15). We selected *E. coli* and VRE, the plurality of screened AROs, for further analyses based on availability of donor-recipient matched pairs and longitudinal samples. Pairwise average nucleotide identity (ANI) was above 97% for all *E. coli* isolates (Figure 2.16), with more than 99.43% identity for all VRE (Figure 2.17). Core genome phylogeny indicated the *E. coli* were mostly of the B2 and D phylogroups. Isolates not only clustered together based on the patient of origin, but also with their corresponding RBX2660 (Figure 2.16).

In general, RBX2660 recipients demonstrated faster clearance of AROs as compared to placebo recipients (Figure 2.15). Simultaneously, new AROs from

RBX2660, mostly *E. coli*, were introduced to corresponding patients. Calculation of single nucleotide polymorphism (SNP) distances revealed many of these AROs were likely clonal, with a median of 6 SNPs for all pairwise distances indicating near-identical genomes (Table 2.2). We sorted post-treatmemt ARO *E. coli* into RBX2660-origin or patient-origin strains and determined clonal persistence following RBX2660 intervention. The introduced AROs were found in patients longitudinally for up to one year post-treatment (Figure 2.15). In some cases, we observed clonal persistence of patient AROs (*e.g.*, patients R1-05 and R2-18), while in some we observed strain replacement by RBX2660-derived AROs (*e.g.*, patient R2-16). Interestingly, patients receiving the same RBX2660 product did not display identical trends. Patient R2-21 received the same RBX2660 product as R2-18 yet only R2-21 engrafted the RBX2660 ARO (Figure 2.15). Persisting AROs derived from patients R1-05 and R2-18 showed higher phenotypic resistance than their corresponding RBX2660-derived AROs, which failed to engraft. On the other hand, patient R2-21 lacked baseline AROs and perhaps provided a "clean slate" for the ARO engraftment.

Isolate ARGs did not indicate a changing resistance profile for these ARO lineages over time. For instance, *E. coli* isolates exhibited an average of 60 predicted ARGs, and these numbers remained stable throughout the time frame of this investigation. The 15 RBX2660-origin AROs which were engrafted to corresponding recipients harbored beta-lactamase genes such as AmpC (12 AROs), TEM-1 (8), CARB (3, one each of CARB-17, 19, and 20) or CTX-M-14 (1). Antibiotic susceptibility testing (AST) corroborated these findings on the phenotypic level with all introduced AROs

being resistant to ciprofloxacin and levofloxacin, and 60% (9/15) resistant to ampicillin (Figure 2.18). Approximately half were resistant or intermediate to trimethoprim-sulfamethoxazole (7) and doxycycline (7), and a few were resistant to ampicillin-sulbactam (3) and cefazolin (4), while all were susceptible to cefotetan, ceftazidime, meropenem, imipenem, piperacillin-tazobactam, ceftazidime-avibactam, amikacin, aztreonam, tigecycline, and nitrofurantoin. The introduced AROs were *Enterobacteriaceae* and resistant to a median of 4 antibiotics, which was less than that of the patient-origin *Enterobacteriaceae* AROs (median resistance to 7 antibiotics). The most resistant isolate introduced from RBX2660 was an *E. coli* strain which was engrafted into patient R1-09. It was retrieved at 5 subsequent time points (final fecal sample collected at 12 months, all < 20 SNPs, Fig 6). This isolate, DI11, was resistant to ceftriaxone and cefepime and classified as an ESBL-producing *E. coli*. We further validated ESBL production of DI11 and the corresponding patient isolates using double-disk diffusion tests (Table 2.4).

2.4 Discussion

We investigated factors underlying changes in the microbiome derived from RBX2660 in a randomized, double-blind, placebo-controlled clinical trial [26]. Consistent with a previous evaluation [29] but in higher resolution using shotgun metagenomic sequencing, we demonstrated RBX2660 dose-dependent changes in the microbiome. Still, all patients initially increased alpha diversity and shifted taxonomic structure regardless of treatment, which could be accredited to the natural trajectory of recovery

after antibiotic discontinuation[10,50]. We hypothesized that it would be possible to distinguish RBX2660-derived effects from the microbiome recovery after antibiotic discontinuation by assessing both extent and direction of microbiome shifts of placebo recipients as thresholds. To test the hypothesis, we developed a simple yet novel metric, the transplantation index. The transplantation index accounts for long-term changes in the microbiome toward corresponding RBX2660 while controlling for individual variation in baseline composition. With the highest transplantation index among placebo recipients as threshold, we demonstrated that RBX2660 recipients exhibited stronger and longer-lasting microbiome changes toward corresponding RBX2660 than placebo recipients.

To predict transplantation success, we identified baseline taxonomic features that had strong correlations with taxonomic non-transplantation. Species with intrinsic vancomycin resistance were discriminative baseline features of the 4 patients who failed to acquire or maintain transplantation by double RBX2660 administration by day 60 (R2-01, R2-02, R2-03, and R2-14). Previously reported microbiome signatures of vancomycin administration including lower diversity, lower *Firmicutes* and higher *Proteobacteria* levels[10,51,52] could not distinguish the 4 non-transplanted patients from transplanted patients. The specific enrichment of intrinsically vancomycin-resistant species therefore could be an indicator of more severe microbiome disturbance by vancomycin. Interestingly, the baseline relative abundance of *V. atypica* was significantly and positively correlated with durable taxonomic transplantation of RBX2660 microbiome in both the single and double dose arms. *V. atypica* has long been

known as an oral bacteria that communicates and develops oral plaque biofilm with lactic acid bacteria[53,54], but a recent study has highlighted its capacity to build metabolomic networks via a peculiar metabolic function—converting lactate to propionate—in the host gut[55]. Further studies combining both metagenomic and metabolomic analyses are required to uncover the mechanism underlying the positive role of *V. atypica* in durable microbiota transplantation. Relative abundances of *Barnesiella* and *Coprobacillus* genera are significantly correlated with taxonomic transplantation status. *Barnesiella*, which exhibited positive correlation with taxonomic transplantation, also has been linked to clearance of VRE colonization in mice[56]. Two *Bacteroides* species, *B. ovatus* and *B. uniformis*, were overrepresented in transplanted patients, reflecting the previous report on their correlation with the unperturbed gut microbiome[57,58].

We also hypothesized that microbiome features of patients are also associated with the prevention of CDI recurrence during the RBX2660 clinical trial. General linear model-based multivariate statistical analyses identified *K. pneumoniae* as a species associated with treatment failure from all patients or only placebo recipients but not from RBX2660 recipients. Baseline *K. pneumoniae* might indeed be a rCDI-associated feature, such as a biomarker of the imbalanced microbiome[59] that underlies CDI, but not correlate with the outcomes of RBX2660 recipients whose microbiomes were affected by RBX2660. Together with the higher efficacy for RBX2660 on rCDI prevention than placebo[26], the model outputs suggest that RBX2660 transplantation restored the disturbed intestinal microbiota to outcompete *C. difficile*. We reckoned that both dose

levels provide enough unperturbed microbiota to exceed a minimum threshold to achieve clinical efficacy, and the second dose provides additional microbiota from which the taxonomic transplantation may arise. Despite their apparent difference between transplantation indices of single and double dose recipients, the two treatment arms showed equivalent clinical efficacy[26]. Likewise, although early-stage transplantation by day 7 appeared to be an important factor determining durable transplant by day 60, it did not always secure successful prevention of rCDI and vice versa.

The differences between rCDI patients and RBX2660 in both ARG relative abundance and resistome architecture became narrowed in all the three treatment arms over time. These outcomes suggest that antibiotic discontinuation could be the drivers of the changes in resistome during this clinical trial. Despite the natural recovery after antibiotic discontinuation, we hypothesized that transplantation of RBX2660 microbiota shaped patient resistomes. RBX2660 indeed simultaneously introduced and eradicated both ARGs and AROs in patients during the process of transplantation. Previous studies have also demonstrated the efficacy of FMT for eradicating AROs[60], but to our knowledge this is the first to comprehensively track clonality for both RBX2660- and patient-derived ARO isolates. Most introduced AROs were antibiotic resistant *E. coli* that are commonly present in a healthy population[61,62].

We identified one ESBL-producing *E. coli* strain from a RBX2660 product carrying AmpC and CTX-M-14, whose RBX2660 product was administered to one patient, R1-09. The patient was a single-dose recipient, with recorded treatment success

(*i.e.* no recurrence of CDI and absence of diarrhea for 8 weeks post-treatment) and no known clinical disease resulted from the trial. ESBL-producing *E. coli* are not inherently more virulent than other strains but can pose a therapeutic challenge if infection occurs[63]. Of note, this trial enrolled patients from December 2014 to November 2015, prior to recognition of ESBL as an important aspect of donor screening. At that time, donor stools were screened for carbapenem-resistant *Enterobacteriaceae* (CRE) but not ESBL, whereas Rebiotix now screens all donor stools for both CRE and ESBL. In light of a recent death caused by ESBL-producing *E. coli* bacteremia in an immunocompromised patient after FMT[21], our findings highlight the importance of a controlled and regulated donor screening program as well as mandatory, monitored safety reporting. Likewise, our findings prompt a general consideration of risk factors for infections from intestinal microorganisms in any life biotherapeutic investigational product.

2.5 Methods

2.5.1 Study cohort, drug, and specimen

Subjects were recruited from among 17 centers in the United States and Canada from 10 December 2014 through 13 November 2015. Subjects were adults with recurrent CDI who have had either i) at least two recurrences after a primary episode (total three CDI episodes) and had completed at least two rounds of oral antibiotic therapy or ii) had at least two episodes of severe CDI resulting in hospitalization. They were randomly assigned to one of three treatment groups: placebo, single dose, or double dose of RBX2660. All treatments were blinded and delivered by enema[26]. The second dose was

administered approximately 7 days after the first dose. For patients that received two RBX2660 doses, donor selection was random and not constrained to provide a single representative donor per patient.

The selection and screening of donors for RBX2660 was performed as previously described[27,28]. The placebo composed of normal saline and formulation solution including cryoprotectant in the same proportions used for RBX2660 preparation. RBX2660 and placebo were stored frozen after preparation until administration. They were thawed for 24 hours in a refrigerator and administered within 48 hours after thawing. AROs were isolated from patient fecal samples and RBX2660 products on selective agar media plates, chromID VRE (bioMerieux, Marcy-l'Etoile, France), MacConkey with Cefotaxime (Hardy Diagnostics, Santa Maria, CA), MacConkey with Ciprofloxacin, (Hardy Diagnostics), and HardyCHROM™ ESBL (Hardy Diagnostics), at 35˚C in air. The remaining fecal samples were stored frozen at -80˚C until metagenomic DNA extraction. Isolate colonies were sub-cultured to trypticase soy agar with 5% sheep blood (Becton Dickinson, Franklin Lakes, NJ) and identified using VITEK MS matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) system[64,65]. Each isolate was frozen in tryptic soy broth with glycerol at -80˚C.

2.5.2 Antibiotic susceptibility testing

Antibiotic susceptibility testing was performed through Kirby Bauer disk diffusion, and the resulting zone sizes were interpreted according to the M100 document from the Clinical and Laboratory Standards Institute[66].

2.5.3 DNA extraction and sequencing

Metagenomic DNA was extracted from approximately 100 mg of fecal samples using DNeasy PowerSoil Kit (Qiagen) following the manufacturer's protocol excepting the lysis step: fecal samples were lysed by 2 rounds of bead beating for 2 min (total 4 min) at 2,500 oscillations/min using a Mini-Beadbeater-24 (Biospec Products). Samples were chilled on ice for 2 minutes between the two bead beating rounds. Extracted DNA was quantified using a Qubit fluorometer dsDNA HS Assay (Invitrogen) and stored at −20°C until the library preparation. Metagenomic DNA was diluted to 0.5 ng/µL before preparing the sequencing library. Libraries were prepared using the Nextera DNA Library Prep Kit (Illumina) as previously described[67]. The libraries then were purified through the Agencourt AMPure XP system (Beckman Coulter) and quantified by Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen) before sequencing. Approximately 70 library samples were pooled in an equimolar manner at the final concentration of 5 nM for each sequencing lane. Prepared pools were submitted for 2 × 150 bp paired-end sequencing on an Illumina NextSeq High-Output platform at the Center for Genome Sciences and Systems Biology at Washington University in St. Louis with a target sequencing depth of approximately 5.5 million reads per sample.

Isolate genomic DNA was extracted using QIAmp BiOstic Bacteremia DNA Kit (Qiagen). Libraries for whole genome sequencing of isolates were prepared from diluted genomic DNA (0.5 ng/µL) as described above. About 180 libraries were pooled together in an equimolar manner at the final concentration of 5 nM for each sequencing lane. Prepared pools were submitted for 2 × 150 bp paired-end sequencing on an

Illumina NextSeq High-Output platform at the Center for Genome Sciences and Systems Biology at Washington University in St. Louis with a target sequencing depth of approximately 2 million reads per sample.

2.5.4 Data processing and genome assembly

Sequence reads were binned by index sequence. Adapter and index sequences were trimmed using Trimmomatic v.0.38[68] using the following parameters: java -Xms2048m -Xmx2048m -jar trimmomatic-0.38.jar PE -phred33 ILLUMINACLIP: NexteraPE-PE.fa:2:30:10:1:true SLIDINGWINDOW:4:15 LEADING:10 TRAILING:10 MINLEN:60. Human sequence contamination was eliminated using Deconseq[69], and the qualities of resulting reads were verified by FastQC (https://github.com/s-andrews/FastQC).

Isolate genomes were assembled, assessed, and annotated using SPAdes[70], QUAST[71], and Prokka[72], respectively. Average nucleotide identity between *E. coli* and VRE isolate pairs were calculated using dnadiff[73]. Within-species pan genomes and core genome alignments were obtained with Roary[74] with default parameters, using 24 and 4 NCBI reference strains (Table 2.3) for *E. coli* and VRE, respectively, with additional *Escherichia fergusonii* and general *Enterobacter faecalis* as outgroups. Alignments were converted via FastTree[75] and visualized on iTOL v4[76].

2.5.5 Microbiome analyses

Microbiome taxonomic composition was predicted by MetaPhlAn v2.0[77] and controlled for relative abundance. Genus-level composition plots were obtained by grouping

together genus present in less than 50% of samples as "Other." The DS00 pseudo-donor microbiome was obtained by averaging species-level taxonomic profiles of all RBX2660 microbiomes. Bray-Curtis distances were calculated using the vegan package[78] and visualized as PCoA plots via the ape package[79] in R 3.5.3. LEfSe[31] identified baseline taxonomic and metabolic features distinguishing transplanted and non-transplanted patients (alpha value for the factorial Kruskal-Wallis test = 0.05, threshold on the logarithmic LDA score = 2). HUMAnN2[80] was employed for metabolic pathway prediction. Longitudinal changes distinguishing transplanted and non-transplanted patients were identified using the ZIBR[43] package in R. Taxa were filtered for non-zero presence in at least 40% samples, and >0.01 relative abundance in the 90th percentile. Each taxon's relative abundance was modeled as both the logistic (X) and beta (Z) components (alpha value for Benjamini-Hochberg-adjusted P=0.05) with transplantation outcome as a fixed effect. Baseline features distinguishing patients with and without rCDI were detected using MaAsLin2. MaAsLin2 is a general linear model-based association detector for microbiome associations with metadata, in this case associations with treatment outcome (success or failure). Taxa were filtered with a minimum prevalence of 0.1 and a minimum relative abundance of 0.0001. Five different models were fitted: one for all patients (total n=63), one for each treatment arm separately (placebo, n = 21; single dose, n = 22; double dose, n=21), as well as one for RBX2660 recipients (n=43) (alpha value for Benjamini-Hochberg-adjusted P=0.05).

2.5.6 Resistome identification and Random Forest classifier

ARGs in the microbiome were identified using ShortBRED[81] with CARD[44]. Isolate ARGs were identified with RGI and CARD[44,82]. The resulting genes were manually curated into more general ARG families (n = 64). A subset of 70% of available resistomes were then used to train a Random Forest classifier distinguishing patient baseline and RBX2660 resistomes (training set n=103), which was then tested on the remaining samples (test set n=45). The Random Forest classifier was built with the package scikit-learn (https://scikit-learn.org/stable/index.html) on Python 3.7.3, with trees averaging 12 nodes and a maximum depth of 4.

2.5.7 ARO tracking and SNP calling

SNPs were called using Bowtie2[83], SAMtools, and BCFtools[84], with the first isolate from the patient or corresponding RBX2660 product used as the reference genome. Reads from subsequent isolates of the same species were aligned against the reference with Bowtie2 (-X 2000 --no-mixed --very-sensitive --n-ceil 0,0.01). BAM files were obtained and sorted with SAMtools (view and sort), which were then converted to pileup files (mpileup). BCFtools view generated VCF files, and variants were called, with the following criteria: minimum coverage of 10 reads per SNP, major allele frequency above 95%, and FQ-score of -85 or less. Indels were excluded. VCF files for each patient were compiled with BCFtools merge, after which SNPs were parsed and counted using custom python and R scripts.

2.6 Data availability

The metagenomic sequencing data are uploaded to NCBI under BioProject PRJNA606075. The isolate genome sequences and assemblies are uploaded to NCBI under BioProject PRJNA606074.

2.7 References

1. Smits WK, Lyras D, Lacy DB, Wilcox MH, Kuijper EJ. Clostridium difficile infection. *Nat Rev Dis Primers*. 2016;2:16020. doi:10.1038/nrdp.2016.20

2. Colman RJ, Rubin DT. Fecal microbiota transplantation as therapy for inflammatory bowel disease: a systematic review and meta-analysis. *J Crohns Colitis*. 2014;8(12):1569-1581. doi:10.1016/j.crohns.2014.08.006

3. Pinn DM, Aroniadis OC, Brandt LJ. Is fecal microbiota transplantation (FMT) an effective treatment for patients with functional gastrointestinal disorders (FGID)? *Neurogastroenterol Motil*. 2015;27(1):19-29. doi:10.1111/nmo.12479

4. Leshem A, Horesh N, Elinav E. Fecal Microbial Transplantation and Its Potential Application in Cardiometabolic Syndrome. *Front Immunol*. 2019;10:1341. doi:10.3389/fimmu.2019.01341

5. de Groot PF, Frissen MN, de Clercq NC, Nieuwdorp M. Fecal microbiota transplantation in metabolic syndrome: History, present and future. *Gut Microbes*. 2017;8(3):253-267. doi:10.1080/19490976.2017.1293224

6. Evrensel A, Ceylan ME. Fecal Microbiota Transplantation and Its Usage in Neuropsychiatric Disorders. *Clin Psychopharmacol Neurosci*. 2016;14(3):231-237. doi:10.9758/cpn.2016.14.3.231

7. Cerovic M, Forloni G, Balducci C. Neuroinflammation and the Gut Microbiota: Possible Alternative Therapeutic Targets to Counteract Alzheimer's Disease? *Front Aging Neurosci*. 2019;11:284. doi:10.3389/fnagi.2019.00284

8. Ooijevaar RE, Terveer EM, Verspaget HW, Kuijper EJ, Keller JJ. Clinical Application and Potential of Fecal Microbiota Transplantation. *Annu Rev Med*. 2019;70:335-351. doi:10.1146/annurev-med-111717-122956

9. Castro I, Tasias M, Calabuig E, Salavert M. Doctor, my patient has CDI and should continue to receive antibiotics. The (unresolved) risk of recurrent CDI. *Rev Esp Quimioter*. 2019;32 Suppl 2:47-54.

10. Isaac S, Scher JU, Djukovic A, et al. Short- and long-term effects of oral vancomycin on the human intestinal microbiota. *J Antimicrob Chemother*. 2017;72(1):128-136. doi:10.1093/jac/dkw383

11. Song JH, Kim YS. Recurrent Clostridium difficile Infection: Risk Factors, Treatment, and Prevention. *Gut Liver*. 2019;13(1):16-24. doi:10.5009/gnl18071

12. Deshpande A, Hurless K, Cadnum JL, et al. Effect of Fidaxomicin versus Vancomycin on Susceptibility to Intestinal Colonization with Vancomycin-Resistant Enterococci and Klebsiella pneumoniae in Mice. *Antimicrob Agents Chemother*. 2016;60(7):3988-3993. doi:10.1128/AAC.02590-15

13. Al-Nassir WN, Sethi AK, Li Y, Pultz MJ, Riggs MM, Donskey CJ. Both oral metronidazole and oral vancomycin promote persistent overgrowth of vancomycin-resistant enterococci during treatment of Clostridium difficile-associated disease. *Antimicrob Agents Chemother*. 2008;52(7):2403-2406. doi:10.1128/AAC.00090-08

14. Laffin M, Millan B, Madsen KL. Fecal microbial transplantation as a therapeutic option in patients colonized with antibiotic resistant organisms. *Gut Microbes*. 2017;8(3):221-224. doi:10.1080/19490976.2016.1278105

15. Woodworth MH, Hayden MK, Young VB, Kwon JH. The Role of Fecal Microbiota Transplantation in Reducing Intestinal Colonization With Antibiotic-Resistant Organisms: The Current Landscape and Future Directions. *Open Forum Infect Dis*. 2019;6(7). doi:10.1093/ofid/ofz288

16. Youngster I, Sauk J, Pindar C, et al. Fecal microbiota transplant for relapsing Clostridium difficile infection using a frozen inoculum from unrelated donors: a randomized, open-label, controlled pilot study. *Clin Infect Dis*. 2014;58(11):1515-1522. doi:10.1093/cid/ciu135

17. Quraishi MN, Widlak M, Bhala N, et al. Systematic review with meta-analysis: the efficacy of faecal microbiota transplantation for the treatment of recurrent and refractory Clostridium difficile infection. *Aliment Pharmacol Ther*. 2017;46(5):479-493. doi:10.1111/apt.14201

18. Iqbal U, Anwar H, Karim MA. Safety and efficacy of encapsulated fecal microbiota transplantation for recurrent Clostridium difficile infection: a systematic review. *Eur J Gastroenterol Hepatol*. 2018;30(7):730-734. doi:10.1097/MEG.0000000000001147

19. Hocquart M, Lagier JC, Cassir N, et al. Early Fecal Microbiota Transplantation Improves Survival in Severe Clostridium difficile Infections. *Clin Infect Dis*. 2018;66(5):645-650. doi:10.1093/cid/cix762

20. Leung V, Vincent C, Edens TJ, Miller M, Manges AR. Antimicrobial Resistance Gene Acquisition and Depletion Following Fecal Microbiota Transplantation for Recurrent Clostridium difficile Infection. *Clin Infect Dis*. 2018;66(3):456-457. doi:10.1093/cid/cix821

21. DeFilipp Z, Bloom PP, Torres Soto M, et al. Drug-Resistant E. coli Bacteremia Transmitted by Fecal Microbiota Transplant. *N Engl J Med*. 2019;381(21):2043-2050. doi:10.1056/NEJMoa1910437

22. Antibiotic Resistance Threats in the United States 2019. Published online 2019. https://www.cdc.gov/drugresistance/biggest-threats.html

23. Johnston KJ, Thorpe KE, Jacob JT, Murphy DJ. The incremental cost of infections associated with multidrug-resistant organisms in the inpatient hospital setting-A national estimate. *Health Serv Res*. 2019;54(4):782-792. doi:10.1111/1475-6773.13135

24. Millan B, Park H, Hotte N, et al. Fecal Microbial Transplants Reduce Antibiotic-resistant Genes in Patients With Recurrent Clostridium difficile Infection. *Clin Infect Dis*. 2016;62(12):1479-1486. doi:10.1093/cid/ciw185

25. Singh R, de Groot PF, Geerlings SE, et al. Fecal microbiota transplantation against intestinal colonization by extended spectrum beta-lactamase producing Enterobacteriaceae: a proof of principle study. *BMC Res Notes*. 2018;11(1):190. doi:10.1186/s13104-018-3293-x

26. Dubberke ER, Lee CH, Robert Orenstein, Khanna S, Hecht G, Gerding DN. Results From a Randomized, Placebo-Controlled Clinical Trial of a RBX2660-A Microbiota-Based Drug for the Prevention of Recurrent Clostridium difficile Infection. *Clin Infect Dis*. 2018;67(8):1198-1204. doi:10.1093/cid/ciy259

27. Orenstein R, Dubberke E, Hardi R, et al. Safety and Durability of RBX2660 (Microbiota Suspension) for Recurrent Clostridium difficile Infection: Results of the PUNCH CD Study. *Clin Infect Dis*. 2016;62(5):596-602. doi:10.1093/cid/civ938

28. Ray A, Jones C. Does the donor matter? Donor vs patient effects in the outcome of a next-generation microbiota-based drug trial for recurrent Clostridium difficile infection. *Future Microbiol*. 2016;11:611-616. doi:10.2217/fmb.16.10

29. Blount KF, Shannon WD, Deych E, Jones C. Restoration of Bacterial Microbiome Composition and Diversity Among Treatment Responders in a Phase 2 Trial of RBX2660: An Investigational Microbiome Restoration Therapeutic. *Open Forum Infect Dis*. 2019;6(4):ofz095. doi:10.1093/ofid/ofz095

30. Mallick H, McIver L, Rahnavard A, et al. Multivariable Association in Population-scale Meta-omics Studies.

31. Segata N, Izard J, Waldron L, et al. Metagenomic biomarker discovery and explanation. *Genome Biol*. 2011;12(6):R60. doi:10.1186/gb-2011-12-6-r60

32. Tankovic J, Leclercq R, Duval J. Antimicrobial susceptibility of Pediococcus spp. and genetic basis of macrolide resistance in Pediococcus acidilactici HM3020. *Antimicrob Agents Chemother*. 1993;37(4):789-792. doi:10.1128/aac.37.4.789

33. Mastro TD, Spika JS, Lozano P, Appel J, Facklam RR. Vancomycin-resistant Pediococcus acidilactici: nine cases of bacteremia. *J Infect Dis*. 1990;161(5):956-960. doi:10.1093/infdis/161.5.956

34. Barton LL, Rider ED, Coen RW. Bacteremic infection with Pediococcus: vancomycin-resistant opportunist. *Pediatrics*. 2001;107(4):775-776. doi:10.1542/peds.107.4.775

35. Campedelli I, Mathur H, Salvetti E, et al. Genus-Wide Assessment of Antibiotic Resistance in Lactobacillus spp. *Appl Environ Microbiol*. 2019;85(1). doi:10.1128/AEM.01738-18

36. Ammor MS, Flórez AB, van Hoek AHAM, et al. Molecular characterization of intrinsic and acquired antibiotic resistance in lactic acid bacteria and bifidobacteria. *J Mol Microbiol Biotechnol*. 2008;14(1-3):6-15. doi:10.1159/000106077

37. Zarazaga M, Sáenz Y, Portillo A, et al. In vitro activities of ketolide HMR3647, macrolides, and other antibiotics against Lactobacillus, Leuconostoc, and Pediococcus isolates. *Antimicrob Agents Chemother*. 1999;43(12):3039-3041.

38. Britt NS, Potter EM. Clinical epidemiology of vancomycin-resistant Enterococcus gallinarum and Enterococcus casseliflavus bloodstream infections. *J Glob Antimicrob Resist*. 2016;5:57-61. doi:10.1016/j.jgar.2015.12.002

39. Cai D, Sorokin V, Lutwick L, et al. C. glycolicum as the sole cause of bacteremia in a patient with acute cholecystitis. *Ann Clin Lab Sci*. 2012;42(2):162-164.

40. Buu-Hoï A, Sapoetra A, Branger C, Acar JF. Antimicrobial susceptibility of Gemella haemolysans isolated from patients with subacute endocarditis. *Eur J Clin Microbiol*. 1982;1(2):102-106. doi:10.1007/bf02014200

41. Lucas GM, Lechtzin N, Puryear DW, Yau LL, Flexner CW, Moore RD. Vancomycin-resistant and vancomycin-susceptible enterococcal bacteremia: comparison of clinical features and outcomes. *Clin Infect Dis*. 1998;26(5):1127-1133. doi:10.1086/520311

42. Tyrrell KL, Citron DM, Warren YA, Fernandez HT, Merriam CV, Goldstein EJC. In vitro activities of daptomycin, vancomycin, and penicillin against Clostridium difficile, C. perfringens, Finegoldia magna, and Propionibacterium acnes. *Antimicrob Agents Chemother*. 2006;50(8):2728-2731. doi:10.1128/AAC.00357-06

43. Chen EZ, Li H. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics*. 2016;32(17):2611-2617. doi:10.1093/bioinformatics/btw308

44. Jia B, Raphenya AR, Alcock B, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res*. 2017;45(D1):D566-D573. doi:10.1093/nar/gkw1004

45. Maaten L van der, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9:2579-2605.

46. Gibson MK, Forsberg KJ, Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J*. 2015;9(1):207-216. doi:10.1038/ismej.2014.106

47. Pehrsson EC, Tsukayama P, Patel S, et al. Interconnected microbiomes and resistomes in low-income human habitats. *Nature*. 2016;533(7602):212-216. doi:10.1038/nature17672

48. Aminov RI, Garrigues-Jeanjean N, Mackie RI. Molecular ecology of tetracycline resistance: development and validation of primers for detection of tetracycline resistance genes encoding ribosomal protection proteins. *Appl Environ Microbiol*. 2001;67(1):22-32. doi:10.1128/AEM.67.1.22-32.2001

49. Bryce A, Costelloe C, Hawcroft C, Wootton M, Hay AD. Faecal carriage of antibiotic resistant Escherichia coli in asymptomatic children and associations with primary care antibiotic prescribing: a systematic review and meta-analysis. *BMC Infect Dis*. 2016;16:359. doi:10.1186/s12879-016-1697-6

50. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. Diversity, stability and resilience of the human gut microbiota. *Nature*. 2012;489(7415):220-230. doi:10.1038/nature11550

51. Vrieze A, Out C, Fuentes S, et al. Impact of oral vancomycin on gut microbiota, bile acid metabolism, and insulin sensitivity. *J Hepatol*. 2014;60(4):824-831. doi:10.1016/j.jhep.2013.11.034

52. Tomas ME, Mana TSC, Wilson BM, et al. Tapering Courses of Oral Vancomycin Induce Persistent Disruption of the Microbiota That Provide Colonization Resistance to Clostridium difficile and Vancomycin-Resistant Enterococci in Mice. *Antimicrob Agents Chemother*. 2018;62(5). doi:10.1128/AAC.02237-17

53. Egland PG, Palmer RJ, Kolenbrander PE. Interspecies communication in Streptococcus gordonii-Veillonella atypica biofilms: signaling in flow conditions requires juxtaposition. *Proc Natl Acad Sci USA*. 2004;101(48):16917-16922. doi:10.1073/pnas.0407457101

54. Johnson BP, Jensen BJ, Ransom EM, et al. Interspecies signaling between Veillonella atypica and Streptococcus gordonii requires the transcription factor CcpA. *J Bacteriol*. 2009;191(17):5563-5565. doi:10.1128/JB.01226-08

55. Scheiman J, Luber JM, Chavkin TA, et al. Meta-omics analysis of elite athletes identifies a performance-enhancing microbe that functions via lactate metabolism. *Nat Med*. 2019;25(7):1104-1109. doi:10.1038/s41591-019-0485-4

56. Ubeda C, Bucci V, Caballero S, et al. Intestinal microbiota containing Barnesiella species cures vancomycin-resistant Enterococcus faecium colonization. *Infect Immun*. 2013;81(3):965-973. doi:10.1128/IAI.01197-12

57. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*. 2012;486(7402):207-214. doi:10.1038/nature11234

58. Qin J, Li R, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*. 2010;464(7285):59-65. doi:10.1038/nature08821

59. Ganji L, Alebouyeh M, Shirazi MH, et al. Dysbiosis of fecal microbiota and high frequency of Citrobacter, Klebsiella spp., and Actinomycetes in patients with irritable bowel syndrome and gastroenteritis. *Gastroenterol Hepatol Bed Bench*. 2016;9(4):325-330.

60. Saïdani N, Lagier JC, Cassir N, et al. Faecal microbiota transplantation shortens the colonisation period and allows re-entry of patients carrying carbapenamase-producing bacteria into medical care facilities. *Int J Antimicrob Agents*. 2019;53(4):355-361. doi:10.1016/j.ijantimicag.2018.11.014

61. Tadesse DA, Zhao S, Tong E, et al. Antimicrobial drug resistance in Escherichia coli from humans and food animals, United States, 1950-2002. *Emerging Infect Dis*. 2012;18(5):741-749. doi:10.3201/eid1805.111153

62. Bailey JK, Pinyon JL, Anantham S, Hall RM. Commensal Escherichia coli of healthy humans: a reservoir for antibiotic-resistance determinants. *J Med Microbiol*. 2010;59(Pt 11):1331-1339. doi:10.1099/jmm.0.022475-0

63. Lavigne JP, Blanc-Potard AB, Bourg G, et al. Virulence genotype and nematode-killing properties of extra-intestinal Escherichia coli producing CTX-M beta-lactamases. *Clin Microbiol Infect*. 2006;12(12):1199-1206. doi:10.1111/j.1469-0691.2006.01536.x

64. McElvania TeKippe E, Burnham C a. D. Evaluation of the Bruker Biotyper and VITEK MS MALDI-TOF MS systems for the identification of unusual and/or difficult-to-identify microorganisms isolated from clinical specimens. *Eur J Clin Microbiol Infect Dis*. 2014;33(12):2163-2171. doi:10.1007/s10096-014-2183-y

65. Westblade LF, Garner OB, MacDonald K, et al. Assessment of Reproducibility of Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry for Bacterial and Yeast Identification. *J Clin Microbiol*. 2015;53(7):2349-2352. doi:10.1128/JCM.00187-15

66. Clinical & Laboratory Standards Institute. M100 - Performance Standards for Antimicrobial Susceptibility Testing. Published online 2019.

67. Baym M, Kryazhimskiy S, Lieberman TD, Chung H, Desai MM, Kishony R. Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS ONE*. 2015;10(5):e0128036. doi:10.1371/journal.pone.0128036

68. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-2120. doi:10.1093/bioinformatics/btu170

69. Schmieder R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS ONE*. 2011;6(3):e17288. doi:10.1371/journal.pone.0017288

70. Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012;19(5):455-477. doi:10.1089/cmb.2012.0021

71. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29(8):1072-1075. doi:10.1093/bioinformatics/btt086

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

73. Kurtz S, Phillippy A, Delcher AL, et al. Versatile and open software for comparing large genomes. *Genome Biol*. 2004;5(2):R12. doi:10.1186/gb-2004-5-2-r12

74. Page AJ, Cummins CA, Hunt M, et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*. 2015;31(22):3691-3693. doi:10.1093/bioinformatics/btv421

75. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS ONE*. 2010;5(3):e9490. doi:10.1371/journal.pone.0009490

76. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res*. 2019;47(W1):W256-W259. doi:10.1093/nar/gkz239

77. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods*. 2012;9(8):811-814. doi:10.1038/nmeth.2066

78. Oksanen J, Blanchet FG, Friendly M, et al. *Vegan: Community Ecology Package*.; 2019. https://CRAN.R-project.org/package=vegan

79. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*. 2004;20(2):289-290. doi:10.1093/bioinformatics/btg412

80. Franzosa EA, McIver LJ, Rahnavard G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods*. 2018;15(11):962-968. doi:10.1038/s41592-018-0176-y

81. Kaminski J, Gibson MK, Franzosa EA, Segata N, Dantas G, Huttenhower C. High-Specificity Targeted Functional Profiling in Microbial Communities with ShortBRED. *PLoS Comput Biol*. 2015;11(12):e1004557. doi:10.1371/journal.pcbi.1004557

82. McArthur AG, Waglechner N, Nizam F, et al. The comprehensive antibiotic resistance database. *Antimicrob Agents Chemother*. 2013;57(7):3348-3357. doi:10.1128/AAC.00419-13

83. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357-359. doi:10.1038/nmeth.1923

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

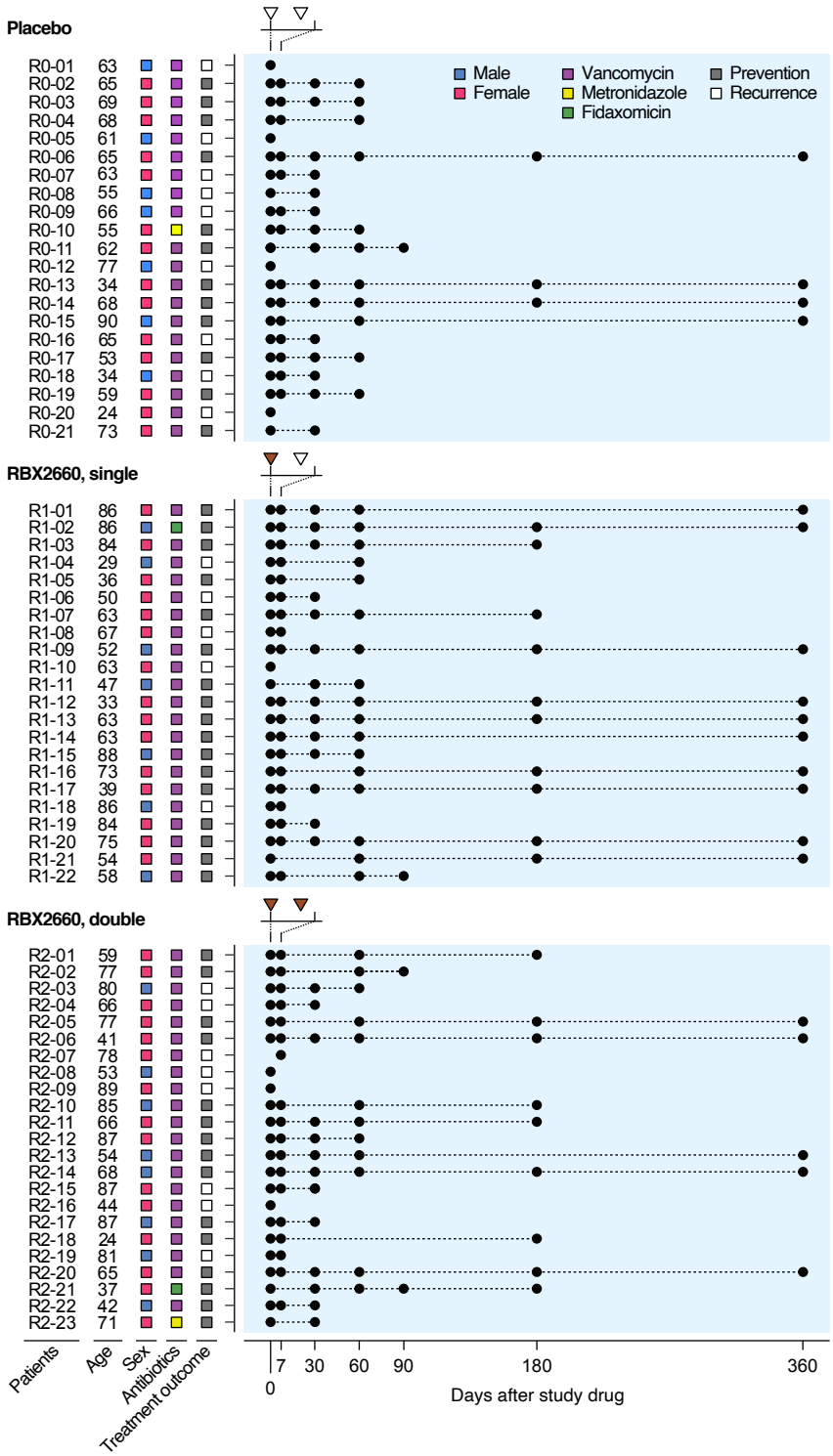

Figure 2.1 Study design for the use of RBX2660 to prevent recurrent *Clostridioides difficile* infection (rCDI). Total of 66 patients with a history of rCDI were treated with RBX2660 in a randomized and blinded manner. Placebo (white triangle) and RBX2660 (brown triangle) were administered and fecal samples (black circle) were collected at the indicated time points. Patients who were declared a new episode of rCDI within 60 days (white square) were moved to open-label treatment.

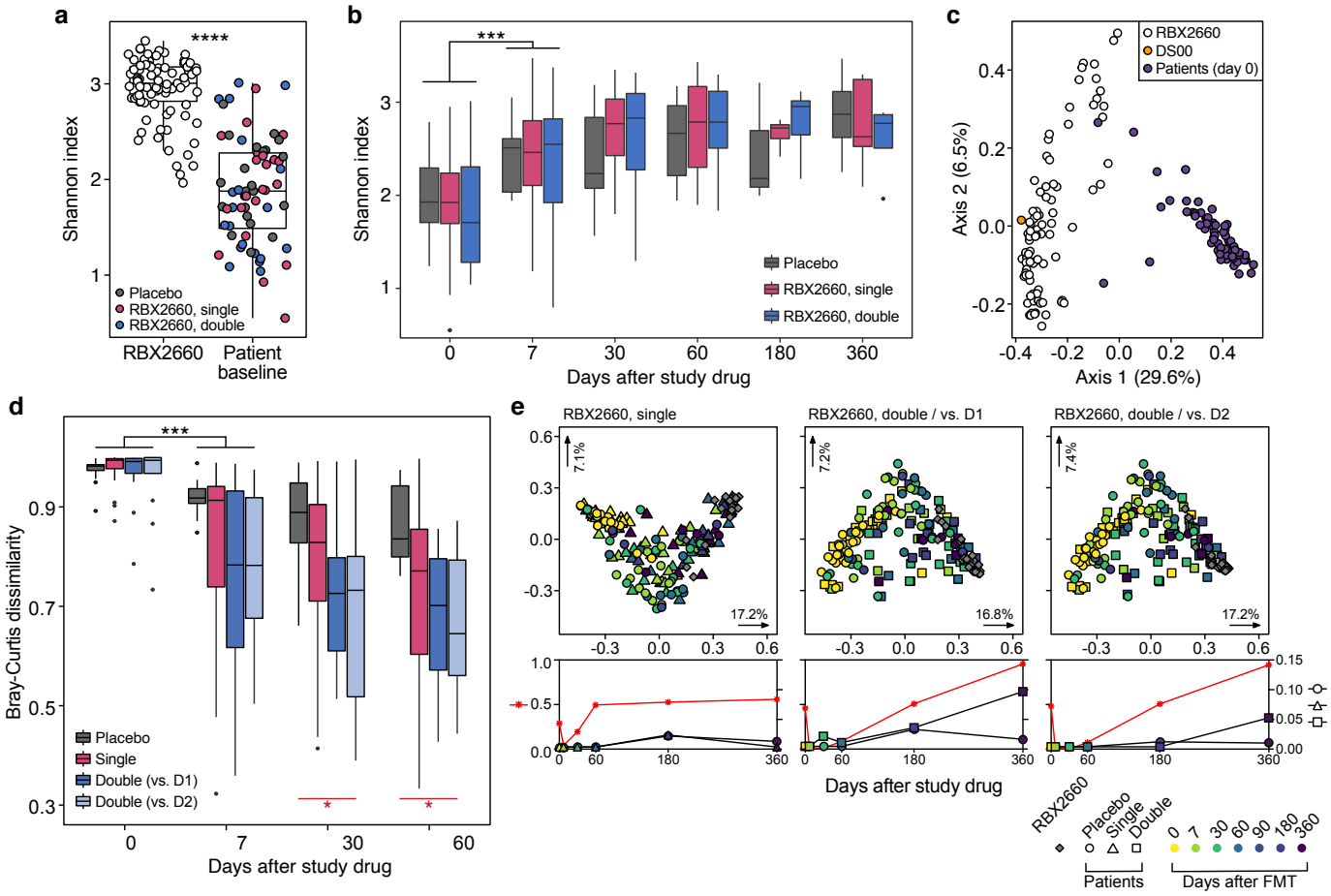

Figure 2.2 RBX2660 shifted taxonomic structures of the gut microbiome of recipients towards a healthy state. (a) RBX2660 products exhibited significantly higher alpha diversity than patient samples before treatment (Wilcoxon signed-rank test) based on the metagenomic taxonomic profiling data. **(b)** Alpha diversity of all patients including placebo recipients increased similarly after treatment. Changes in alpha diversity were significant for the first week after treatment, but there was no statistically significant difference among treatment groups (Kruskal-Wallis test). **(c)** Principal coordinates analysis (PCoA) showed a species level clustering of RBX2660 (white) and pseudo-donor sample DS00 (yellow) distinct from patient baseline samples (violet). **(d)** Bray-Curtis distance between taxonomic structures of patients and corresponding RBX2660. D1 and D2 indicate the first dose and the second dose, respectively. DS00 was used for calculating the Bray-Curtis distance of placebo recipients. The decrease in Bray-Curtis distances was steepest during the first week after treatment (black, Wilcoxon signed-rank test). RBX2660 recipients showed a more dynamic decrease in Bray-Curtis distances than placebo recipients by day 60 (red, Kruskal-Wallis test). *$P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$, **** $P \leq 0.0001$. **(e)**Upper panel s: PCoA describing the direction of changes in taxonomic structures of RBX2660 recipients. Corresponding RBX2660 products and all placebo recipients were included. Lower panels: adjusted *P*-values of PERMANOVA and relevant pairwise comparisons (Pillai-Bartlett non-parametric trace and Benjamini-Hochberg FDR correction). *P*-values of comparisons between placebo

and RBX2660 recipients (red asterisks, left y-axis), placebo recipients and RBX2660 (circle, right y-axis), single dose recipients and RBX2660 (triangle, right y-axis), and double dose recipients and RBX2660 (square, right y-axis) of PCoA plots were presented in corresponding lower panels.

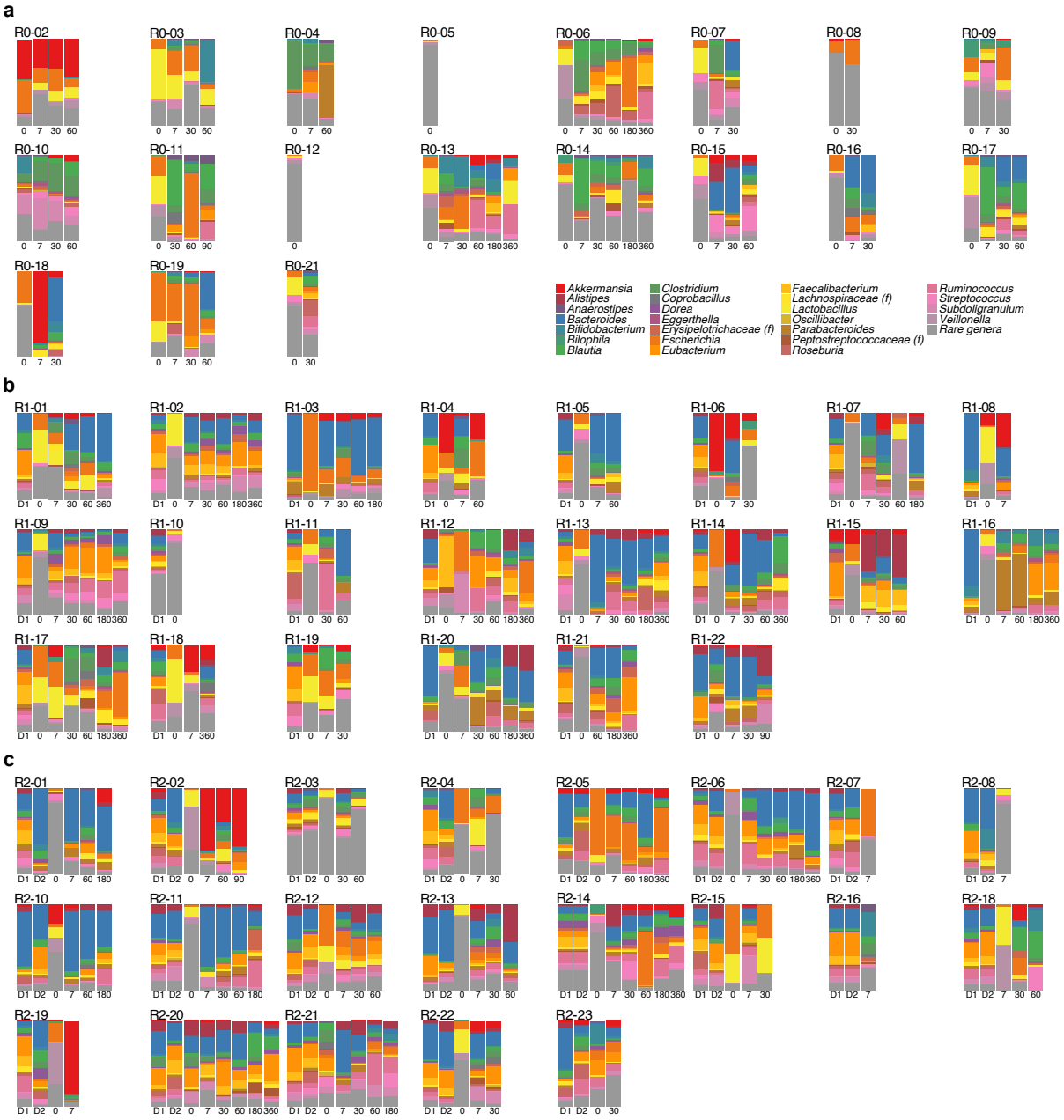

Figure 2.3 Taxonomic overview of patient stool samples at the genus level. Genus composition of RBX2660 products was added next to the corresponding recipient. **(a)** Patients who received 2 doses of placebo. **(b)** Patients who received 1 dose of RBX2660 and 1 dose of placebo. **(c)** Patients who received 2 doses of RBX2660. Patients R0-01, R0-20, and R2-09, who had only the baseline specimen (due to early rCDI) that exhibited insufficient sequencing depth for the analysis of taxonomic structure after decontamination of human reads, were omitted. Patient R2-17 was also omitted from this analysis due to incomplete donor RBX2660 information.

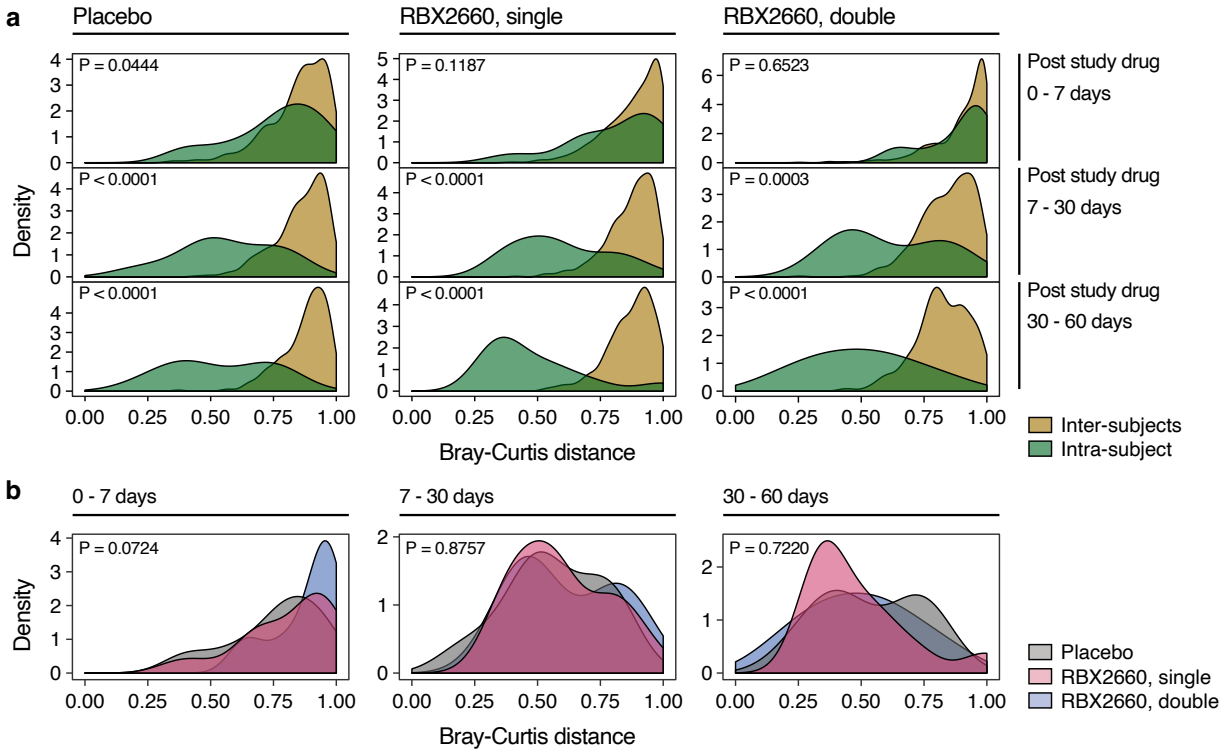

Figure 2.4 Taxonomic shift by treatments. (a) Distribution of inter-subject (yellow) and intra-subject (green) Bray-Curtis dissimilarities of taxonomic structures by treatment groups (column) and by time frames (rows). The similar distribution of inter-subject and intra-subject dissimilarities during the first week in all conditions suggests significant taxonomic shifts in early stage regardless of the dose of RBX2660. Wilcoxon rank sum tests were performed to compare inter-subject and intra-subject dissimilarities of each treatment group during each time frame. **(b)** Comparison of intra-subject dissimilarities of placebo (grey), single (red), and double RBX2660 recipients (blue) during each time frame with Kruskal-Wallis rank sum tests.

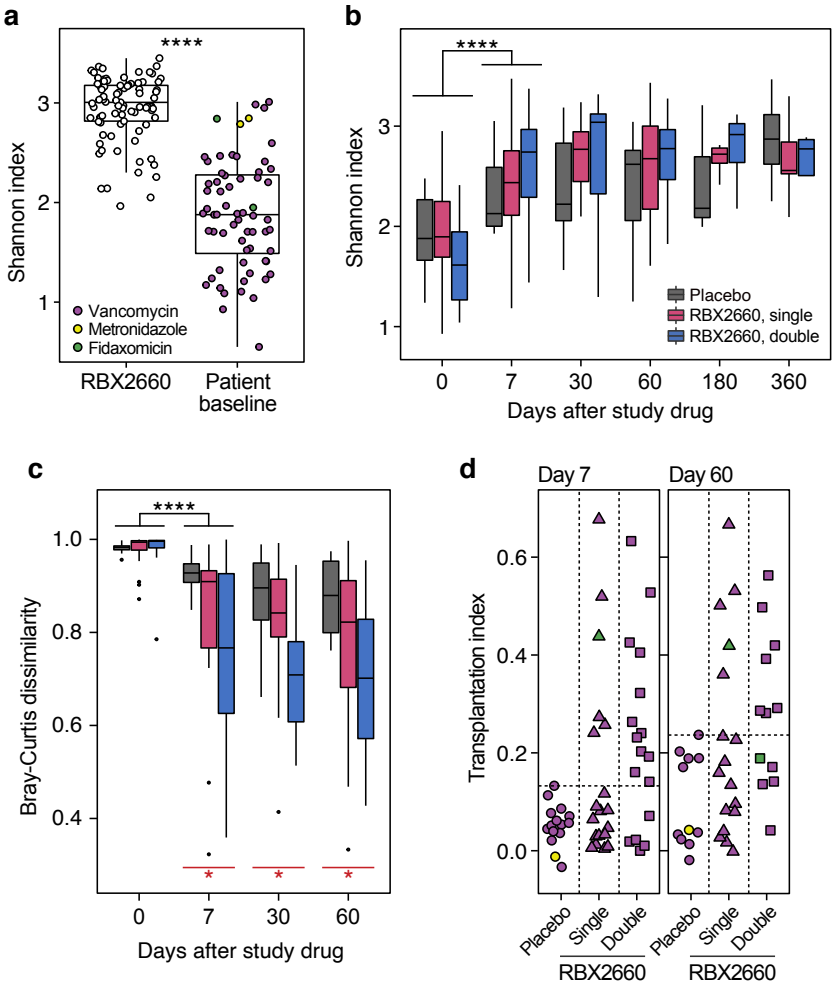

Figure 2.5 The effect of antibiotics prior to study drug on taxonomic shift by RBX2660. (a) RBX2660 products exhibited significantly higher alpha diversity than patient baseline samples (Wilcoxon signed-rank test). Changes in alpha diversity **(b)** and Bray-Curtis dissimilarity **(c)** to corresponding RBX2660 of vancomycin recipients. Changes in the diversity and dissimilarity were still statistically significant for the first week after study drug (black, Wilcoxon signed-rank test) without the metronidazole and fidaxomicin recipients, and RBX2660 recipients showed a more dynamic decrease in Bray-Curtis dissimilarity than placebo recipients after the first week (red, Kruskal-Wallis test). **(d)** Transplantation index of patients on day 7 and 60. Horizontal dash lines indicate the threshold of taxonomic transplantation (Figure 2.2a). Violet, vancomycin; yellow, metronidazole; green, fidaxomicin. *$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$, ****$P \leq 0.0001$.

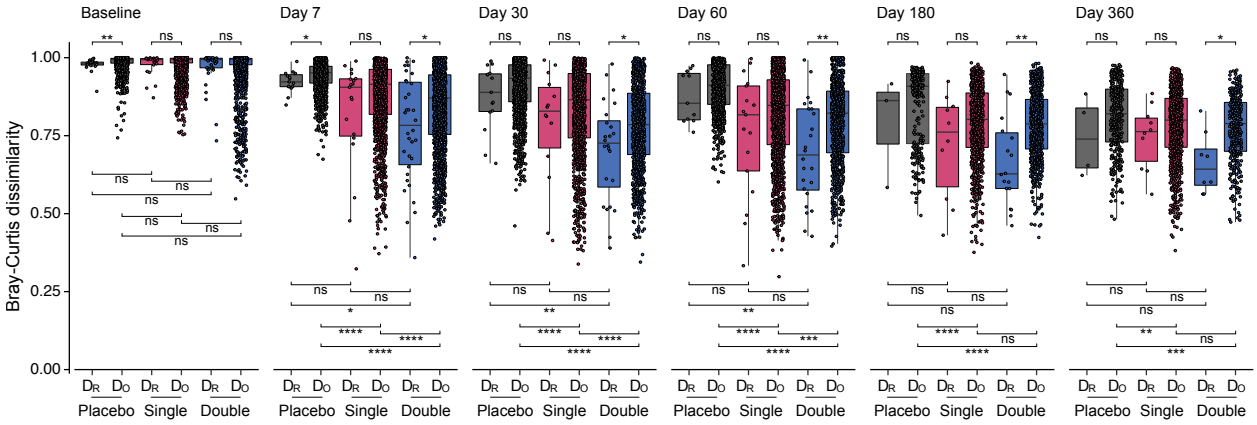

Figure 2.6 Bray-Curtis dissimilarities between patients and respective RBX2660 (D_R) or other random RBX2660 (D_O). Pairwise comparisons of all D_Rs and D_Os of placebo (gray), single dose (red), and double dose recipients (blue) in each time point were simultaneously performed (Wilcoxon signed-rank test with Benjamini-Hochberg FDR correction, FDR < 0.05). Dissimilarities of double dose recipients include both dissimilarities to the first and second RBX2660 doses. *$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$, ****$P \leq 0.0001$.

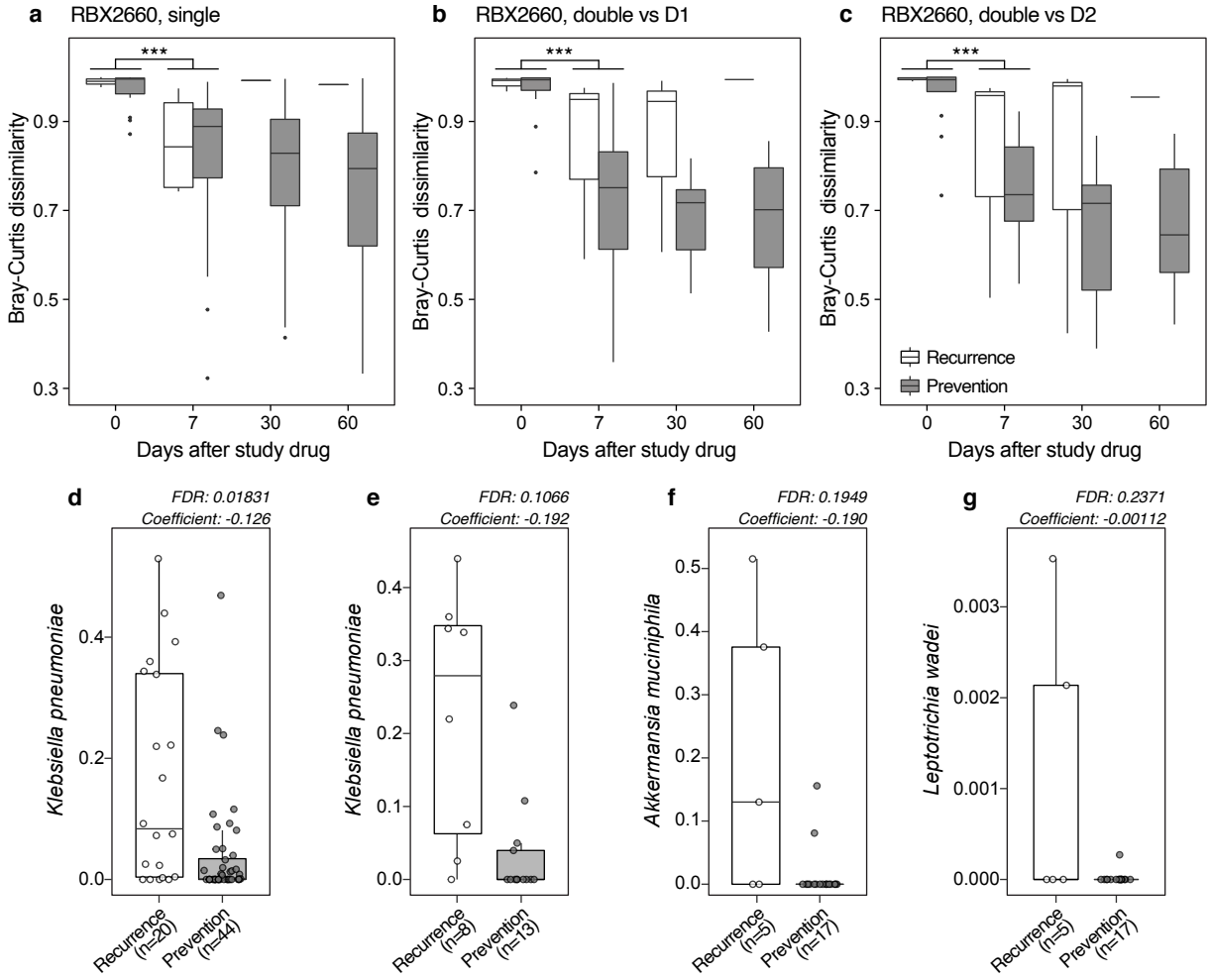

Figure 2.7 Changes in the Bray-Curtis dissimilarities between a patient and corresponding donor after **(a)** single dose RBX2660 and **(b-c)** double dose RBX2660. Changes in taxonomic structures of gut microbiota were significant for the first week after treatment (Kruskal-Wallis test, ***$P < 0.001$). There were no statistically significant differences between patients who experienced recurrent *Clostridioides difficile* infection (rCDI, white) and other successful patients (gray) at all time points (Wilcoxon signed-rank test with Benjamini-Hochberg FDR correction, FDR < 0.05). D1, the first dose; D2, the second dose. **(d)** Relative abundance of *Klebsiella pneumoniae* in all patients. Patients who experienced rCDI (white) exhibited significantly higher *K. pneumoniae* abundance than treatment-success patients (gray). **(d)** Relative abundance comparison of *K. pneumoniae* between treatment-failture and -success patients in placebo recipients. Relative abundance comparison of **(c)** *Akkermansia muciniphila* and **(e)** *Leptotrichia wadei* that were identified by MaAsLin2 as features associated with treatment-failures of single RBX2660 dose recipients. MaAsLin2 could not identify any taxonomic feature associated with treatment outcome from double RBX2660 recipients.

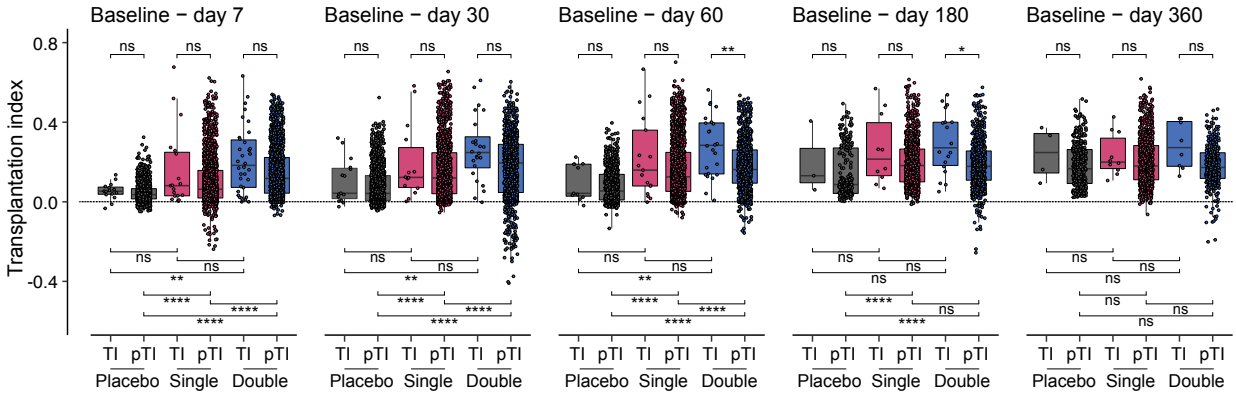

Figure 2.8 Transplantation indices (TIs) and pseudo transplantation indices (pTIs). Pairwise comparisons of all TIs and pTIs of placebo (gray), single dose (red), and double dose recipients (blue) in each time frame were simultaneously performed (Wilcoxon signed-rank test with Benjamini-Hochberg FDR correction, FDR < 0.05). Transplantation indices of double dose recipients include both indices for the first and second RBX2660 doses. **P* ≤ 0.05, ***P* ≤ 0.01, ****P* ≤ 0.001, *****P* ≤ 0.0001.

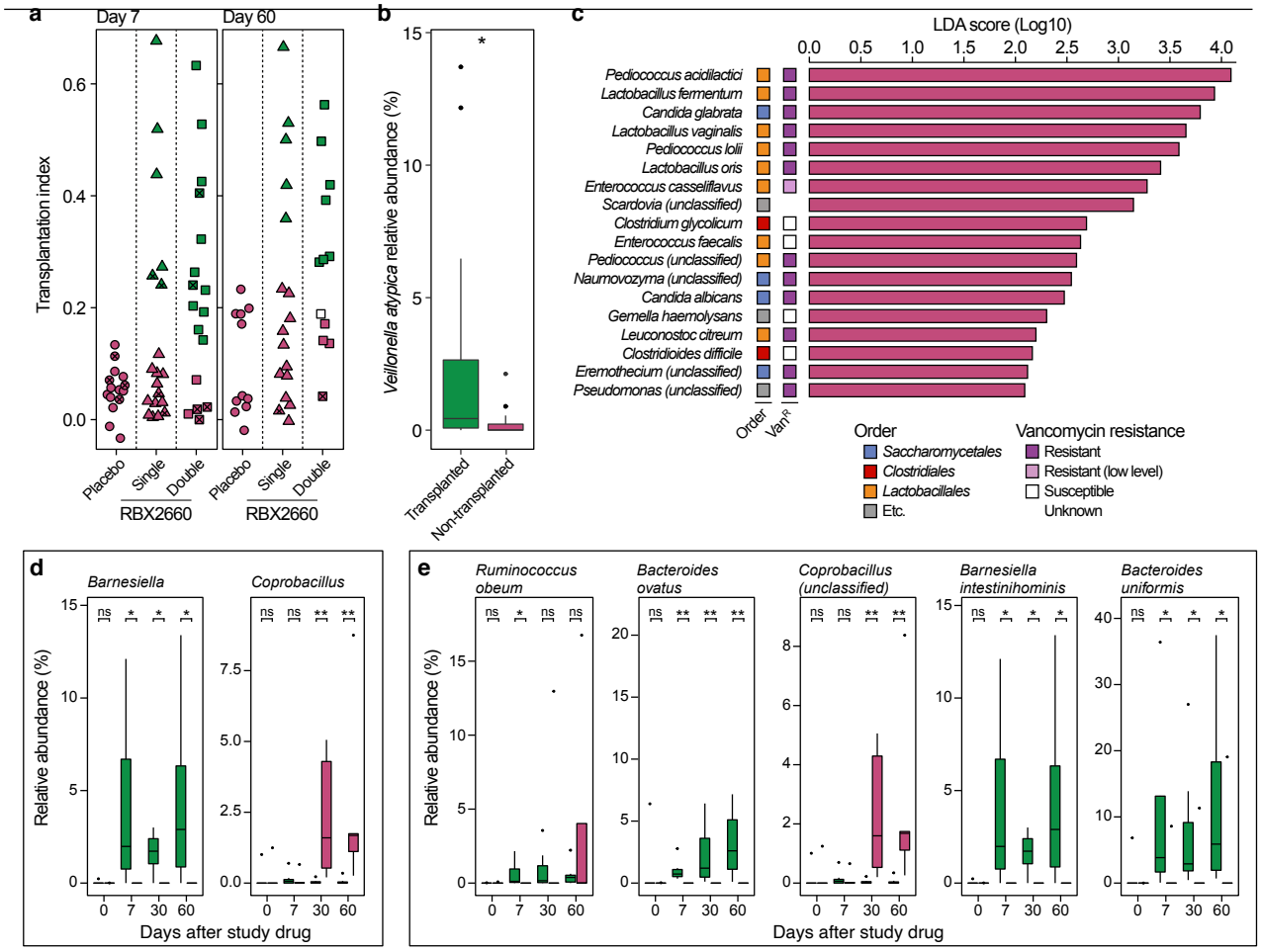

Figure 2.9 Discriminative taxonomic features of RBX2660 transplantation. (a) Transplantation index of patients on day 7 and 60. We defined taxonomic transplantation as a state showing higher transplantation index than that of all placebo recipients (green). The patients who were declared rCDI within 60 days were marked (x). The white square represents one patient who exhibited a lower transplantation index for the first dose but a higher transplantation index for the second dose than placebo patients (R2-21, Figure 2.10a). **(b)** Higher baseline relative abundances of *Veillonella atypica* in patients who showed durable taxonomic transplantation by day 60 in both single and double RBX2660 treatment groups (Wilcoxon signed-rank test, P=0.027). **(c)** Linear discriminant analysis Effect Size (LEfSe) determined baseline taxonomic features of the non-transplanted patients who exhibited lower transplantation indices than placebo recipients at day 60 after double RBX2660 treatment. 13 species among 18 taxonomic features were intrinsically vancomycin resistant (violet square, including *E. casseliflavus* of low resistance). There was no taxonomic feature specific to transplanted patients determined by LEfSe. Genus **(d)** and species enrichment **(e)** associated with taxonomic transplantation (transplanted, green; non-transplanted, purple) were identified through a two-part zero-inflated Beta regression model with random effects (ZIBR) test. *$P \leq 0.05$, **$P \leq 0.01$.

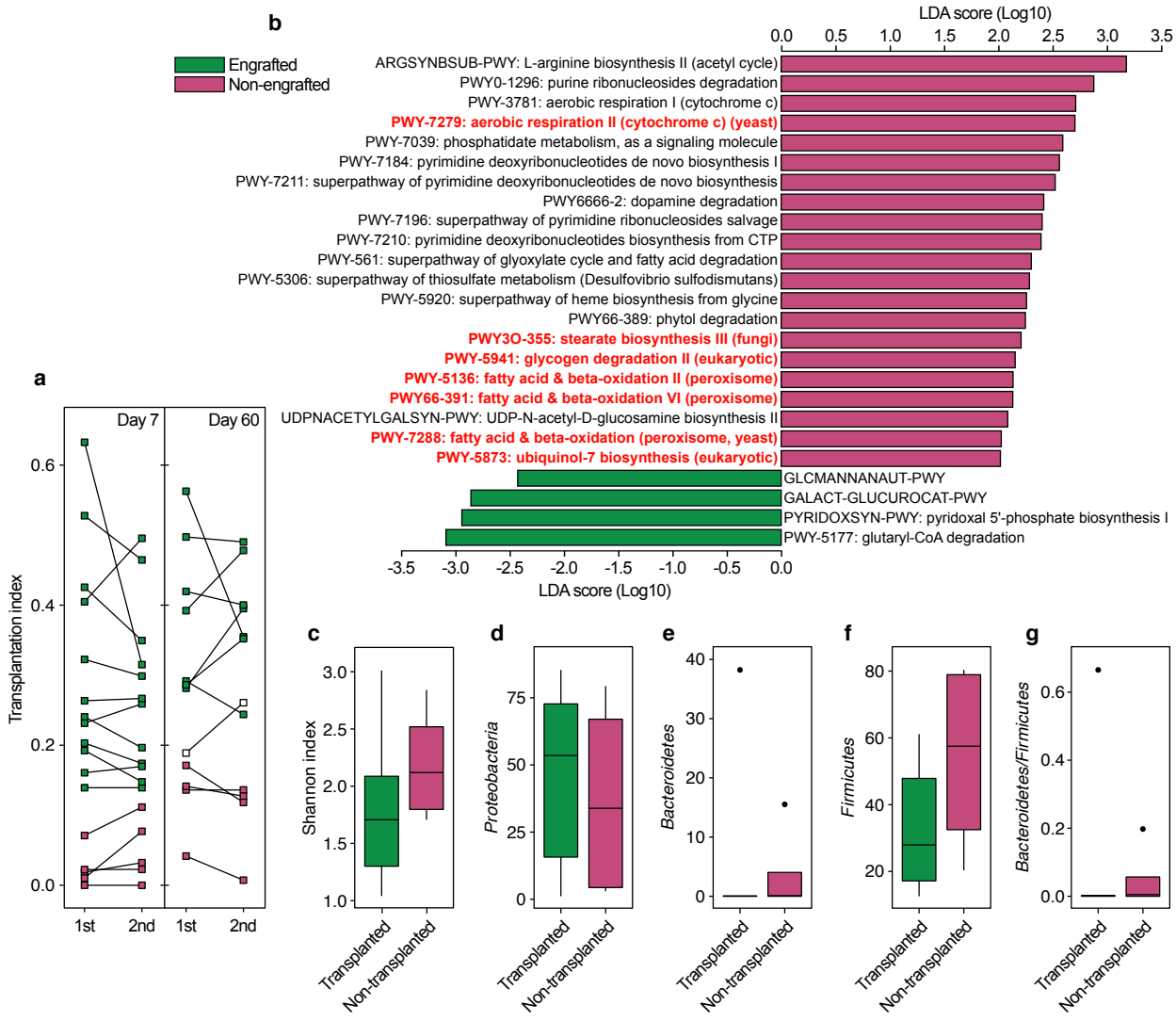

Figure 2.10 Additional discriminative features of the non-transplated patients. (a) Comparison of transplantation indices of double RBX2660 recipients for the first and the second RBX2660 dose at day 7 and day 60. R2-21 exhibited lower engraftment index for the first dose but higher engraftment index for the second dose of RBX2660 than placebo patients at day 60 (white square). **(b)** Metabolic pathway features of the double RBX2660 recipients whose taxonomic structures were engrafted and maintained until day 60 (green) and other non-engrafted patients (purple). Yeast-specific metabolic pathways are marked in red. GLCMANNANAUT-PWY, superpathway of N-acetylglucosamine, N-acetylmannosamine and N-acetylneuraminate degradation; GALACT-GLUCUROCAT-PWY, superpathway of hexuronide and hexuronate degradation. At baseline, **(c)** Shannon index (Wilcoxon signed-rank test, P=0.41), **(d)** *Proteobacteria* (P=0.79), **(e)** *Bacteroidetes* (P=0.92), **(f)** *Firmicutes* (P=0.32), and **(g)** the ratio between *Bacteroidetes* and *Firmicutes* (P=0.92) were not significantly different between the engrafted and non-engrafted double RBX2660 recipients.

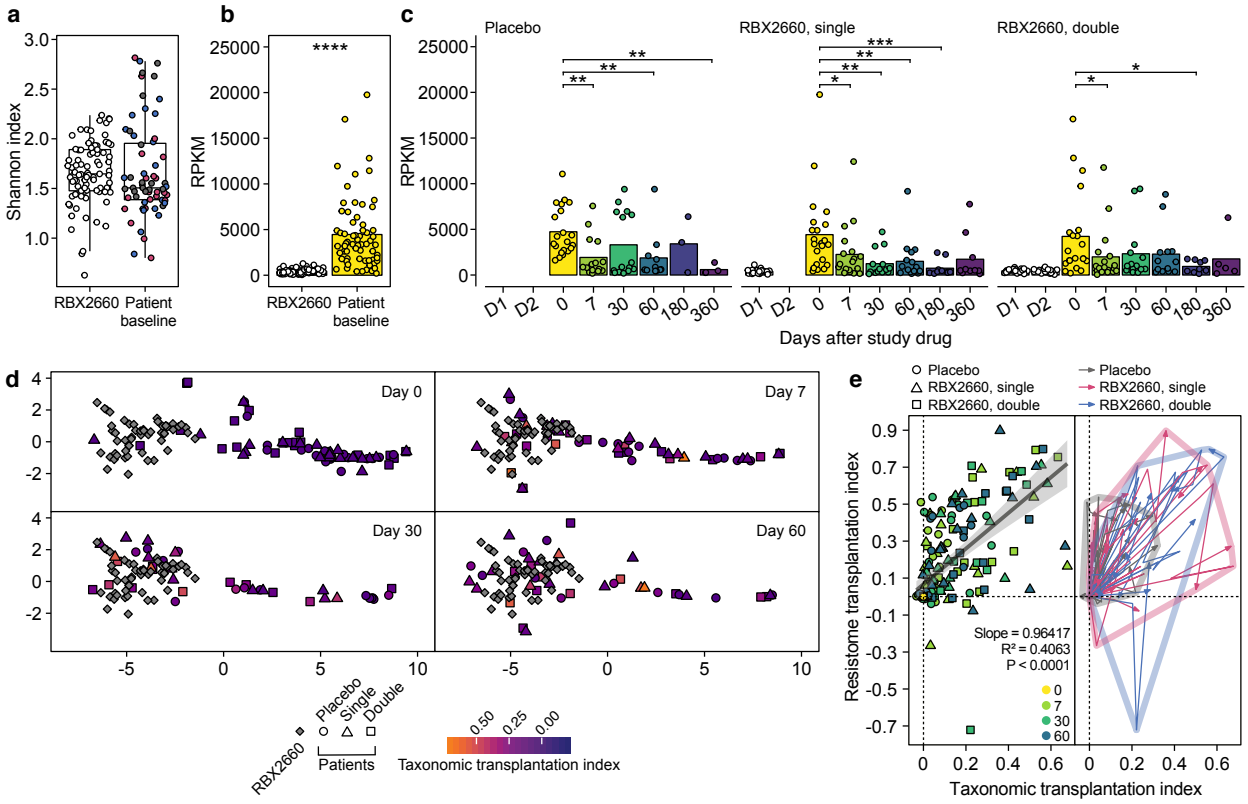

Figure 2.11 RBX2660 fluctuated resistome structures of patients via taxonomic transplantation. (a) Alpha diversity of baseline patient resistomes was comparable to that of RBX2660 (P=0.18). **(b)** However, baseline patient resistomes had a greater antibiotic resistant gene (ARG) reads per kilobase per million sample reads (RPKM, Wilcoxon signed-rank test). **(c)** Significant decrease in ARG RPKM was observed over time in all treatment groups (Wilcoxon signed-rank test with Benjamini-Hochberg FDR correction, FDR < 0.05). Bars indicate mean of individual ARG relative abundances. D1, the first dose; D2, the second dose. **(d)** Patients and RBX2660 products were clustered separately in t-Distributed Stochastic Neighbor Embedding (t-SNE) analysis of resistome structures at day 0. Patient resistomes became similar to RBX2660 over time, but the speed of change varied for each patient regardless of RBX2660 dose and taxonomic transplantation index. **(e)** RBX2660 simultaneously fluctuated both taxonomic and resistome structures more dynamically as compared to placebo. *$P \leq$ 0.05, **$P \leq 0.01$, ***$P \leq 0.001$, ****$P \leq 0.0001$.

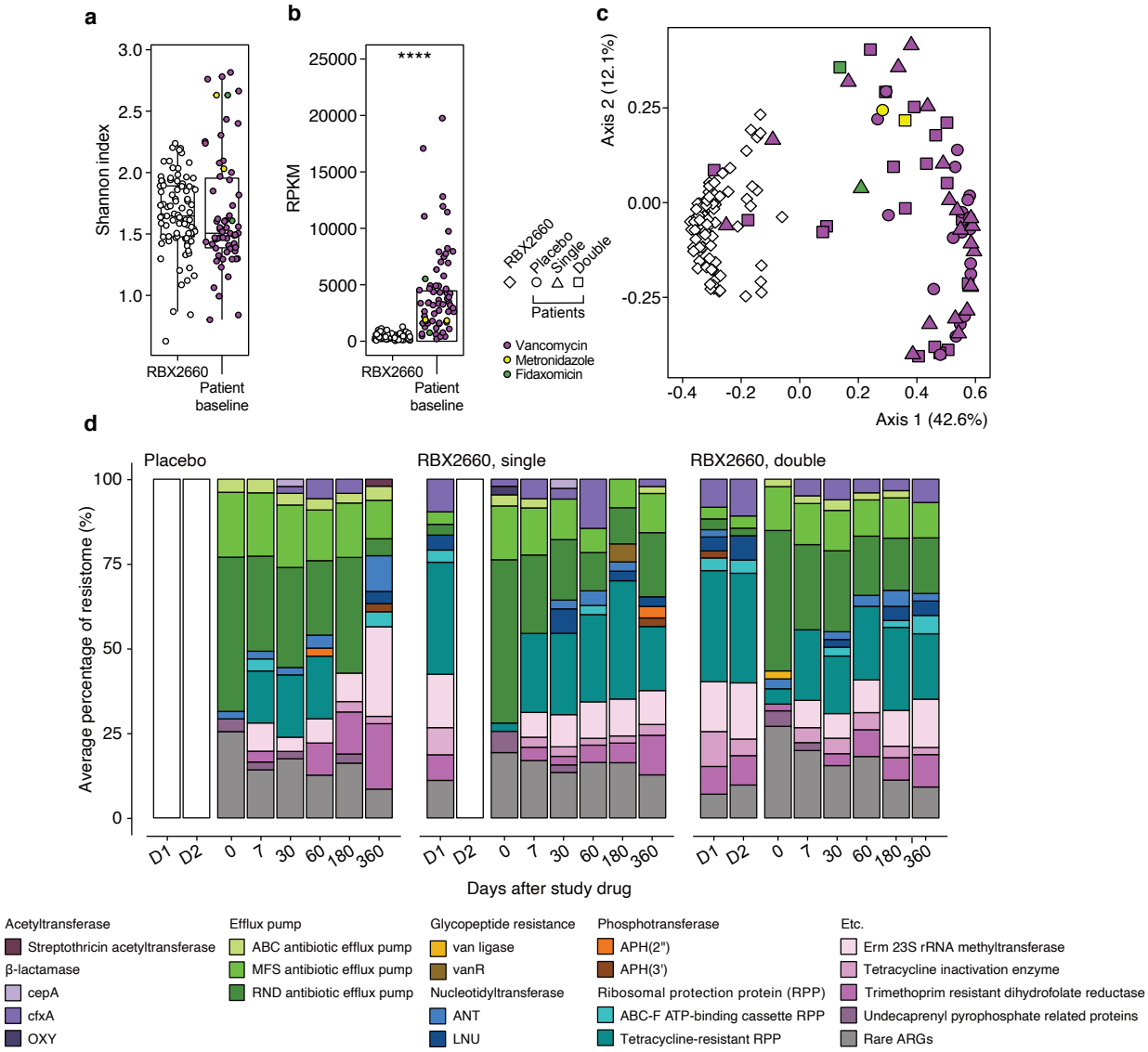

Figure 2.12 Comparison of resistome compositions. (a) Alpha diversity of baseline patient resistomes was comparable to that of RBX2660 (only patients who received vancomycin, *P*=0.066; all patients, *P*=0.180). **(b)** Baseline patient resistomes had a greater antibiotic resistant gene (ARG) reads per kilobase per million sample reads (RPKM, Wilcoxon signed-rank test). ****$P \leq 0.0001$. **(c)** Principal coordinates analysis (PCoA) of resistome composition showed a clustering of RBX2660 (white). Baseline resistomes of metronidazole and fidaxomicin recipients were more closely clustered with other baseline resistomes of vancomycin recipients (*P*=0.0120, PERMANOVA and pairwise comparison with Pillai-Bartlett non-parametric trace and Benjamini-Hochberg FDR correction) than RBX2660 (*P*=0.0015). **(d)** Individual loads of an antibiotic resistant gene (ARG) in a treatment arm were averaged. ARGs whose average portion in the treatment arm was smaller than 2% were combined as "Rare ARGs."

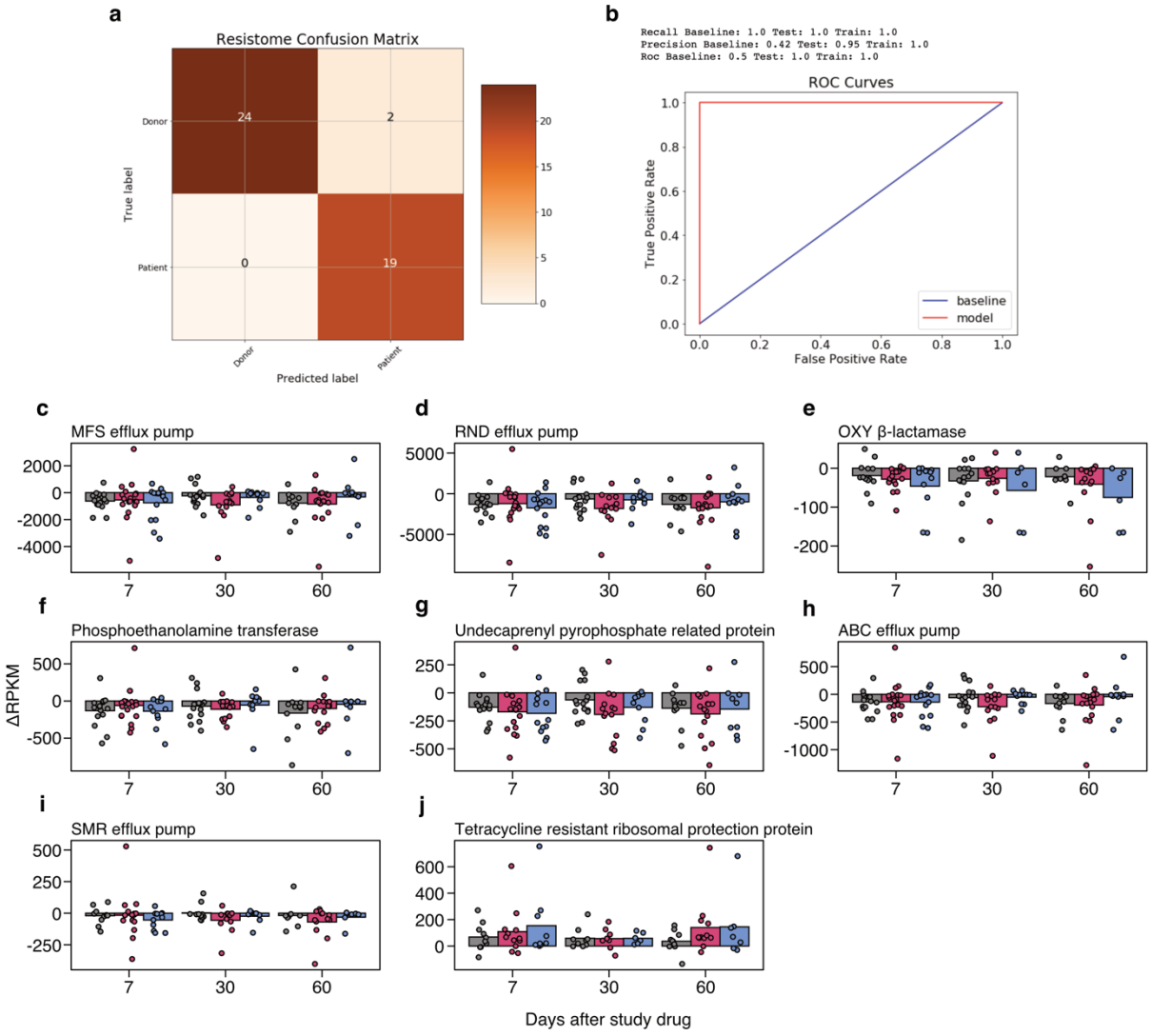

Figure 2.13 Random forest classifier successfully distinguished between donor and patient baseline resistomes. (a) Confusion matrix depicting predicted and true labels for the test set (n = 45). All but 2 samples were correctly categorized. **(b)** A receiver operating characteristic (ROC) curve showed high recall, precision, and area under curve (AUC) for the model. **(c–j)** Individual changes in abundance (reads per kilobase per million sample reads, RPKM) of selected antibiotic resistant genes from baseline were similar among patients in the three treatment groups.

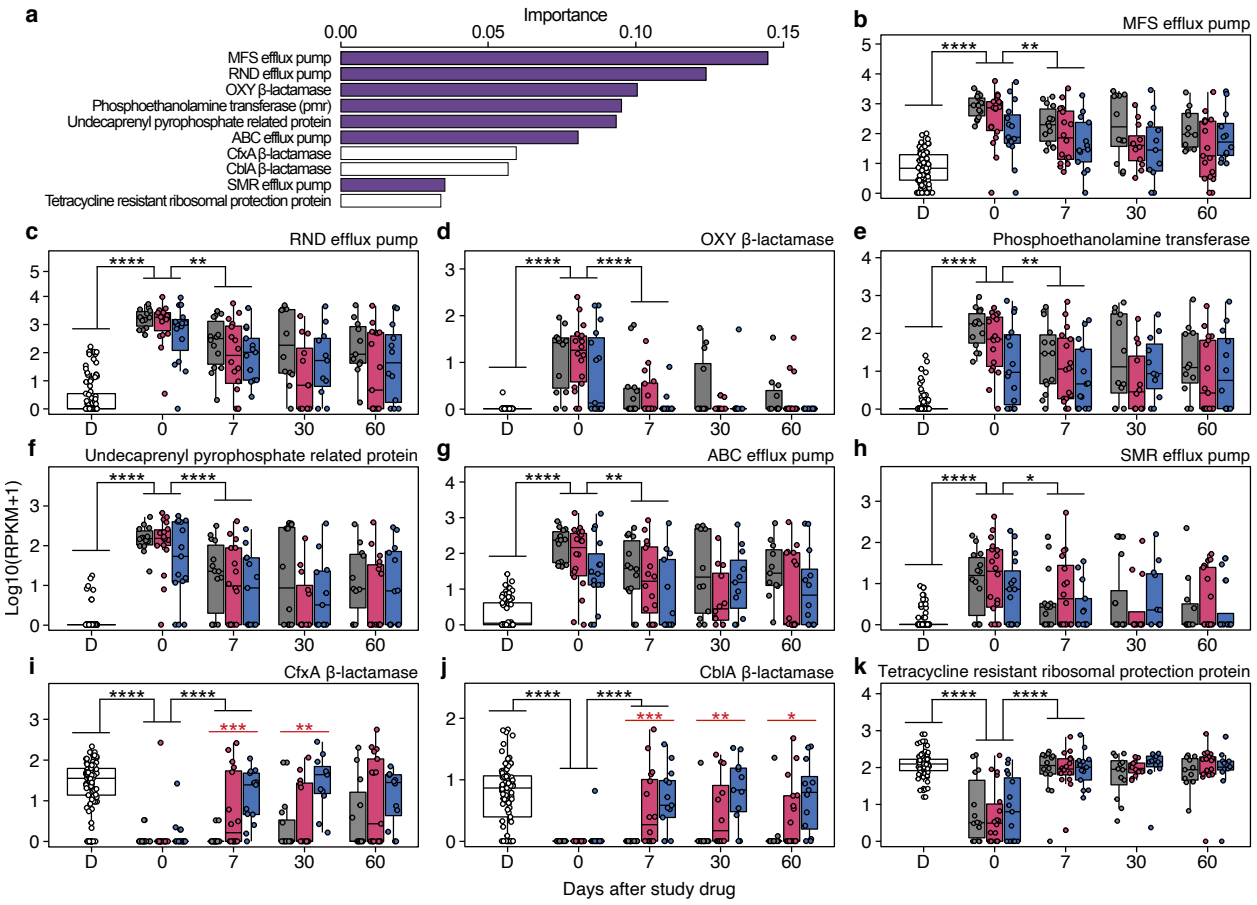

Figure 2.14 Recipients adopted a resistome profile similar to that of donors. (a) Ten most important patient-specific (violet) and RBX2660-specific (white) antibiotic resistant gene (ARG) families were identified through the Random Forest classifier. **(b−k)** Relative abundance of the selected 10 ARGs in RBX2660 ("D") and patients who received placebo (gray), single RBX2660 (red), and double RBX2660 (blue). Relative abundance of patient-specific ARGs decreased over time in all patients without statistically significant difference among treatment arms **(b−h)**. Relative abundance of the two RBX2660-specific beta-lactamases in patients increased by RBX2660 administration in a dose-dependent manner **(i−j)** (red, Kruskal-Wallis test). Tetracycline resistant ribosomal protection protein was a RBX2660-specific ARG, but its relative abundance in placebo recipients also increased after the treatment **(k)**. These changes were significant during the first week after the treatment (black, Wilcoxon signed-rank test). *$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$, ****$P \leq 0.0001$.

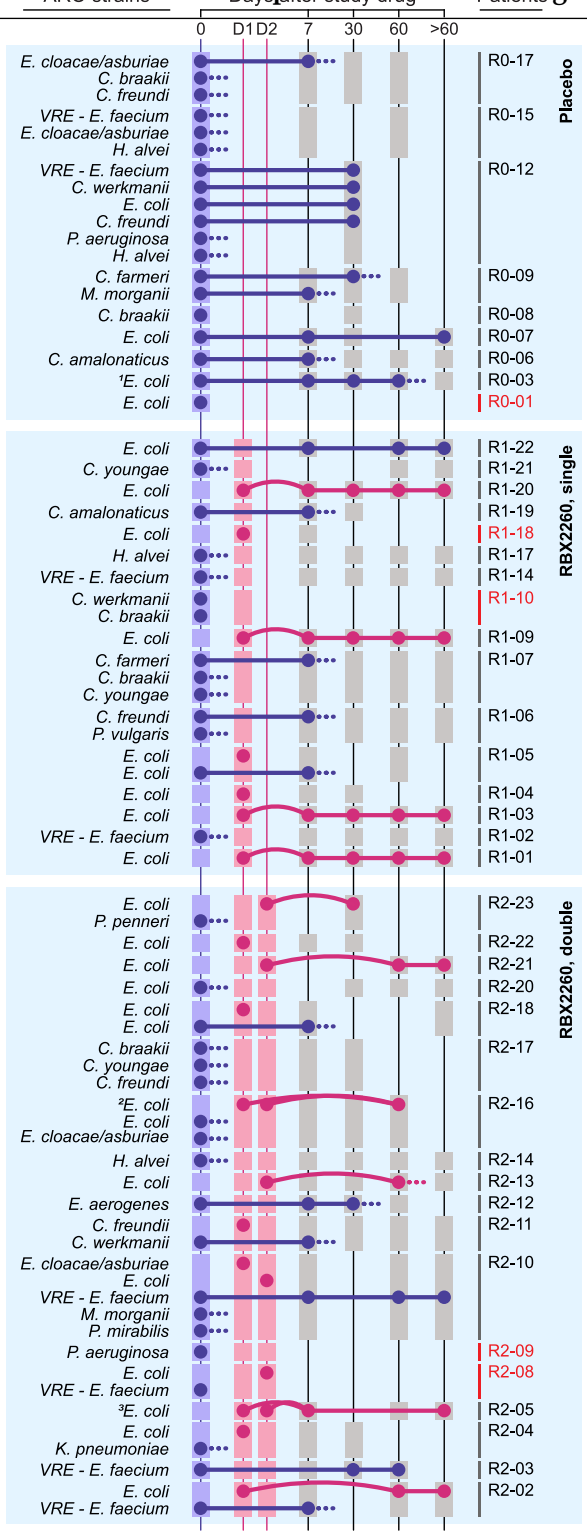

Figure 2.15 RBX2660 effectively cleared antibiotic resistant organisms (AROs) compared to placebo and simultaneously introduced new AROs. We specifically tracked patient-derived (blue dot) and RBX2660-derived AROs (red dot). Patients with no ARO detected from both the baseline sample and corresponding RBX2660 were excluded. Persistency (solid line), disappearance (dash line), and introduction (curved

line) of the AROs were determined by genomic comparison of AROs. Squares indicate sample availability (blue, patient baseline samples; red, RBX2660; gray, patient samples after RBX2660 administration). Patients with no samples after day 7 were marked with red.

[1]R0-03 showed 2−3 separate lineages of *E. coli* prior to day 30, which were reduced to 1 lineage by day 60.
[2]Patient R2-16 received the same RBX2660 product twice.
[3]Although the two RBX2660 products for patient R2-05 were prepared from different donor samples, ARO *E. coli* strains screened from those appeared to be clonal (distance = 8 SNPs).

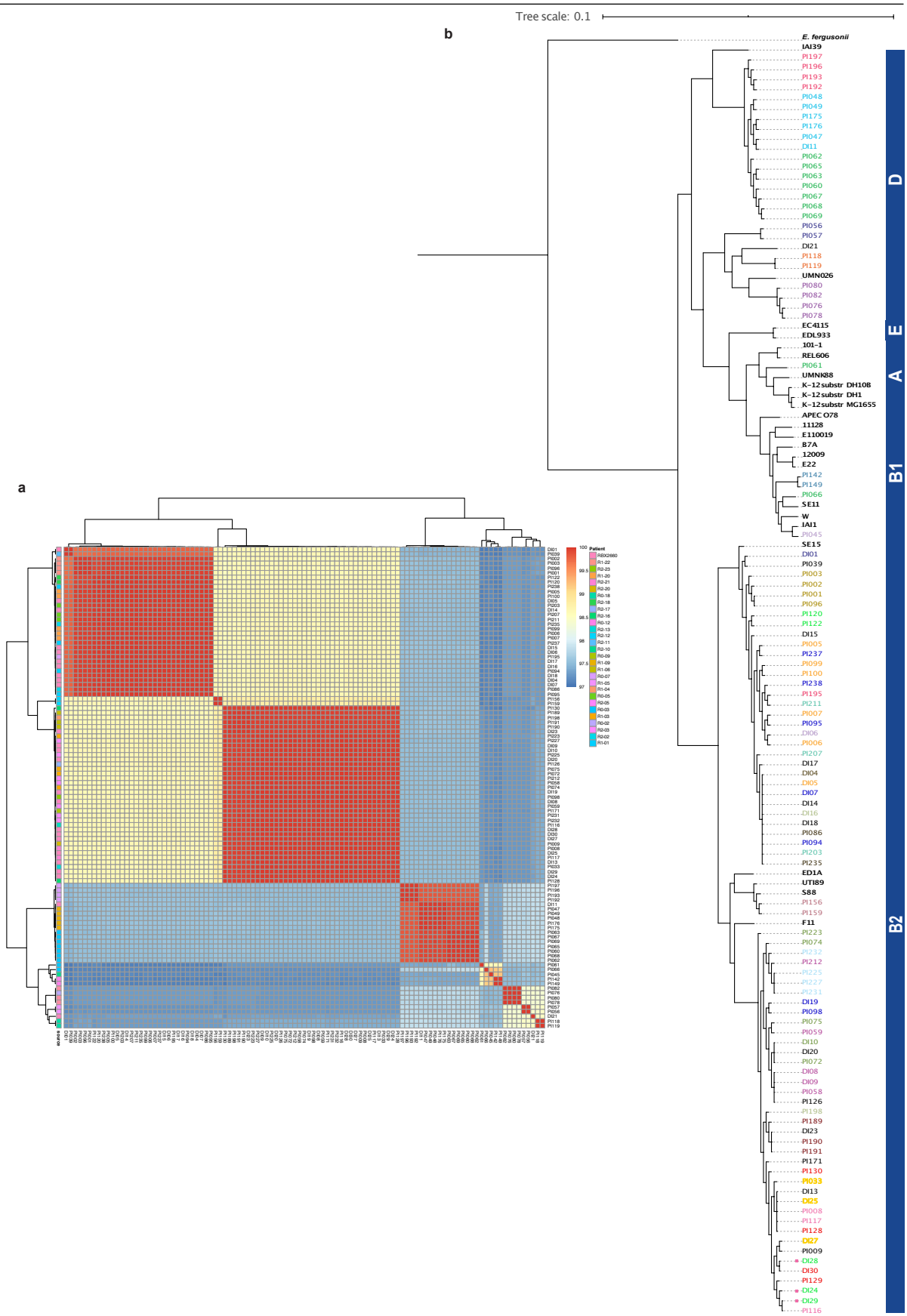
Tree scale: 0.1
b
a
E. fergusonii
IAI39
PI197
PI196
PI193
PI192
PI048
PI049
PI175
PI176
PI047
DI11
PI062
PI065
PI063
PI060
PI067
PI068
PI069
PI056
PI057
DI21
PI118
PI119
UMN026
PI080
PI082
PI076
PI078
EC4115
EDL933
101-1
REL606
PI061
UMNK88
K-12 substr. DH10B
K-12 substr. DH1
K-12 substr. MG1655
APEC O78
11128
E110019
B7A
12009
E22
PI142
PI149
PI066
SE11
W
IAI1
PI045
SE15
DI01
PI039
PI003
PI002
PI001
PI096
PI120
PI122
DI15
PI005
PI237
PI099
PI100
PI238
PI195
PI211
PI007
PI095
DI06
PI006
PI207
DI17
DI04
DI05
DI07
DI14
DI16
DI18
PI086
PI094
PI203
PI235
ED1A
UTI89
S88
PI156
PI159
F11
PI223
PI074
PI232
PI212
PI225
PI227
PI231
DI19
PI098
PI075
PI059
DI10
DI20
PI072
DI08
DI09
PI058
PI126
PI198
PI189
DI23
PI190
PI191
PI171
PI130
PI033
DI13
DI25
PI008
PI117
PI128
DI27
PI009
DI28
DI30
PI129
DI24
DI29
PI116
A
B1
B2
D
E
Patient
RBX2660
100
99.5
99
98.5
98
97.5
97

Figure 2.16 Average nucleotide identity (ANI) and core genome phylogeny of *E. coli* isolates. (a) ANI for all *E. coli* isolates pairwise comparisons. All isolates show at least 97% pairwise identity. **(b)** Core genome phylogeny of *E. coli* isolates with 24 NCBI reference strains and *E. fergusonii* as outgroup. Right panel indicates *E. coli* phylogroup. Isolates originated from the same patient or donor were labelled in the same color. Reference strains were marked in black and bold. Small colored squares indicate isolates from donor product that was administered to multiple patients, where the color of the text and squares correspond to the different patients.

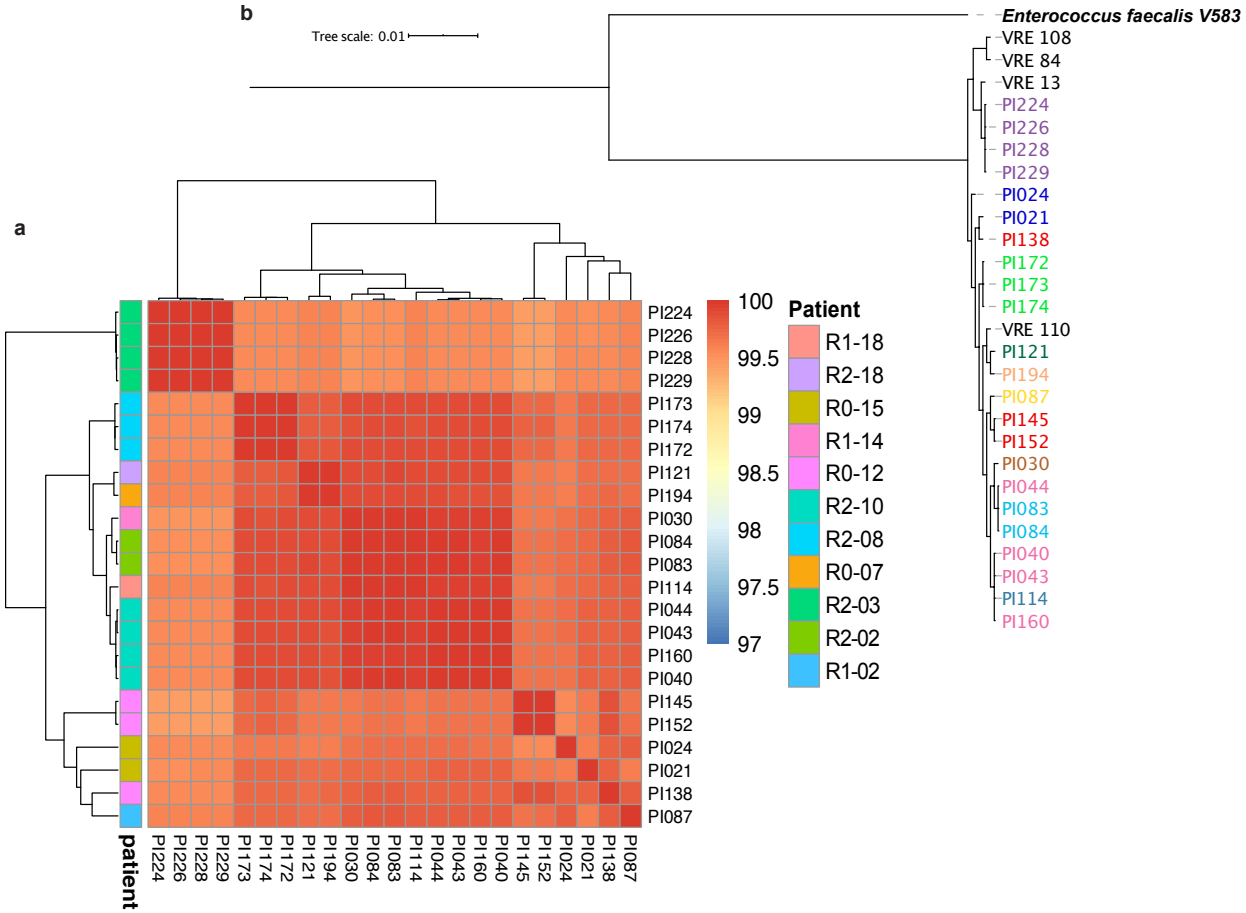

Figure 2.17 Average nucleotide identity (ANI) and core genome phylogeny of VRE isolates. (a) ANI for all VRE isolates pairwise comparisons. All isolates show at least 99.43% pairwise identity. **(b)** Core genome phylogeny of VRE isolates, with 4 NCBI reference VRE strains and V583 *Enterococcus faecalis* as outgroup. Isolates originated from the same patient are labelled in the same color. Reference strain names are marked in black.

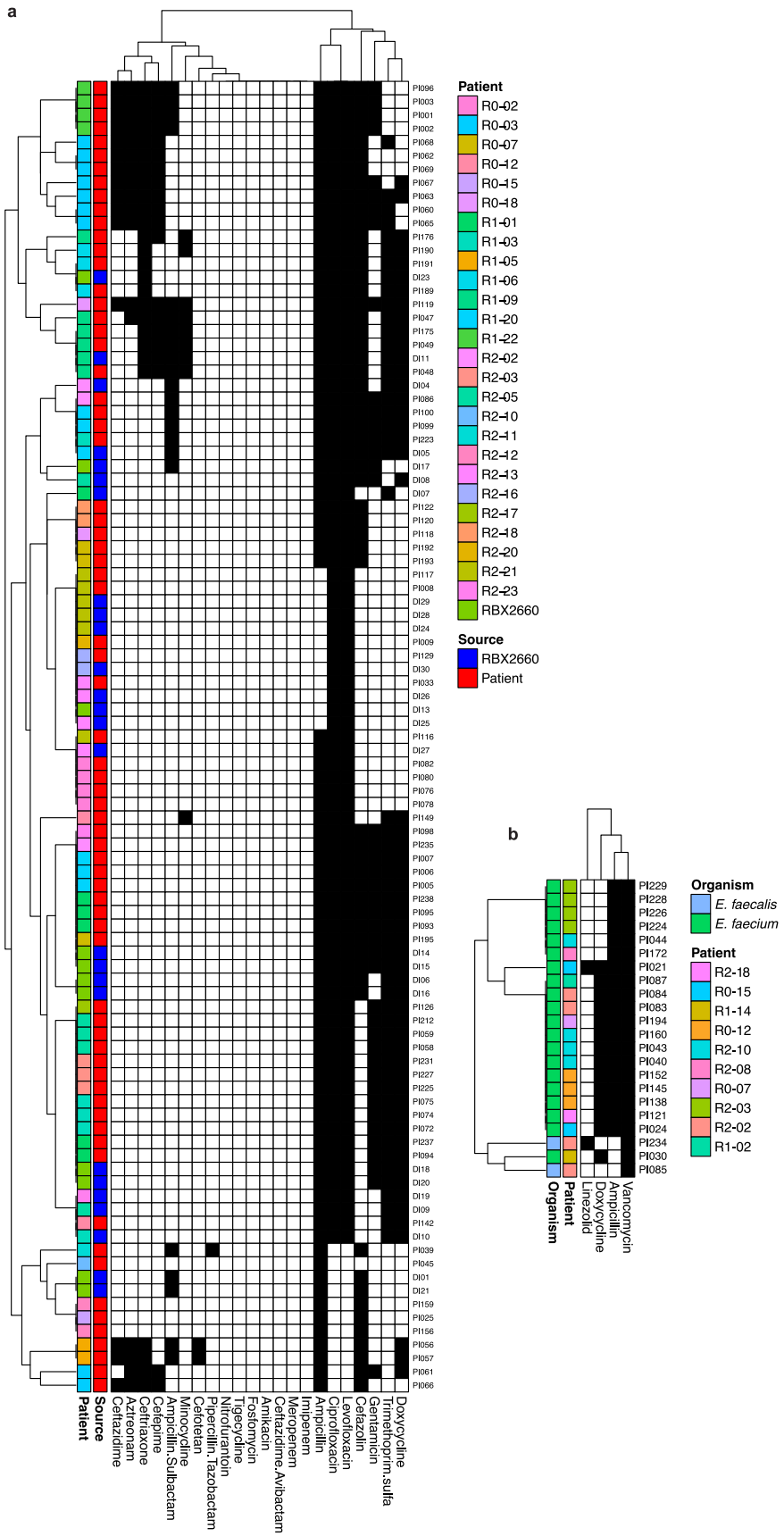
a
Patient
R0-02
R0-03
R0-07
R0-12
R0-15
R0-18
R1-01
R1-03
R1-05
R1-06
R1-09
R1-20
R1-22
R2-02
R2-03
R2-05
R2-10
R2-11
R2-12
R2-13
R2-16
R2-17
R2-18
R2-20
R2-21
R2-23
RBX2660
Source
RBX2660
Patient
Patient
Source
Ceftazidime
Aztreonam
Ceftriaxone
Cefepime
Ampicillin.Sulbactam
Minocycline
Cefotetan
Piperacillin.Tazobactam
Nitrofurantoin
Tigecycline
Fosfomycin
Amikacin
Ceftazidime.Avibactam
Meropenem
Imipenem
Ampicillin
Ciprofloxacin
Levofloxacin
Cefazolin
Gentamicin
Trimethoprim.sulfa
Doxycycline
b
Organism
E. faecalis
E. faecium
Patient
R2-18
R0-15
R1-14
R0-12
R2-10
R2-08
R0-07
R2-03
R2-02
R1-02
Organism
Patient
Linezolid
Doxycycline
Ampicillin
Vancomycin

Figure 2.18 Antibiotic susceptibility testing (AST) results. (a) *E. coli* and **(b)** VRE isolates. AST showed whether isolates are susceptible (white) or intermediate/resistant (black) to a variety of antibiotics. Source of isolate and specific patient of origin were depicted in the sidebars. RBX2660-derived antibiotic resistant organisms engrafted in patients were colored as their corresponding patient. RBX2660 samples without a corresponding patient are otherwise denoted as "RBX2660".

Table 2.1 Patient drug identifiers

Patients	1st dose	2nd dose	New rCDI within 60 days
R0-01	.	.	Yes
R0-02	.	.	.
R0-03	.	.	.
R0-04	.	.	.
R0-05	.	.	Yes
R0-06	.	.	.
R0-07	.	.	Yes
R0-08	.	.	Yes
R0-09	.	.	Yes
R0-10	.	.	.
R0-11	.	.	.
R0-12	.	.	Yes
R0-13	.	.	.
R0-14	.	.	.
R0-15	.	.	.
R0-16	.	.	Yes
R0-17	.	.	.
R0-18	.	.	Yes
R0-19	.	.	.
R0-20	.	.	Yes
R0-21	.	.	.
R1-01	D0-34	.	.
R1-02	D0-63	.	.
R1-03	D0-28	.	.
R1-04	D0-35	.	Yes
R1-05	D0-74	.	.
R1-06	D0-15	.	Yes
R1-07	D0-61	.	.
R1-08	D0-78	.	Yes
R1-09	D0-20	.	.
R1-10	D0-57	.	Yes
R1-11	D0-43	.	.
R1-12	D0-73	.	.
R1-13	D0-14	.	.
R1-14	D0-26	.	.

R1-15	D0-50	.	.
R1-16	D0-78	.	.
R1-17	D0-67	.	.
R1-18	D0-29	.	Yes
R1-19	D0-64	.	.
R1-20	D0-39	.	.
R1-21	D0-26	.	.
R1-22	D0-60	.	.
R2-01	D0-11	D0-62	.
R2-02	D0-40	D0-49	.
R2-03	D0-31	D0-45	Yes
R2-04	D0-38	D0-64	Yes
R2-05	D0-32	D0-30	.
R2-06	D0-18	D0-17	.
R2-07	D0-62	D0-08	Yes
R2-08	D0-80	D0-36	Yes
R2-09	D0-07	D0-07	Yes
R2-10	D0-52	D0-36	.
R2-11	D0-16	D0-73	.
R2-12	D0-75	D0-58	.
R2-13	D0-82	D0-83	.
R2-14	D0-21	D0-21	.
R2-15	D0-56	D0-66	Yes
R2-16	D0-76	D0-76	Yes
R2-17	D0-46	N/A	.
R2-18	D0-77	D0-53	.
R2-19	D0-10	D0-18	Yes
R2-20	D0-71	D0-58	.
R2-21	D0-72	D0-77	.
R2-22	D0-81	D0-59	.
R2-23	D0-51	D0-29	.

Table 2.2 Pairwise SNP distances

Patient	Isolate#1	Isolate#2	Isolate#1 Time (days)	Isolate#2 Time (days)	Species	SNP distance
R1-22	PI003	PI096	60	90	*E.coli*	4
R1-22	PI003	PI002	60	7	*E.coli*	2
R1-22	PI096	PI002	90	7	*E.coli*	4
R1-22	PI001	PI002	0	7	*E.coli*	8
R1-22	PI001	PI003	0	60	*E.coli*	6
R1-22	PI001	PI096	0	90	*E.coli*	10
R2-23	DI19	PI098	RBX2660	30	*E.coli*	6
R1-19	PI103	PI104	0	7	*C. amalonaticus*	2
R1-20	DI05	PI005	RBX2660	7	*E.coli*	2
R1-20	DI05	PI006	RBX2660	30	*E.coli*	2
R1-20	DI05	PI007	RBX2660	60	*E.coli*	1
R1-20	DI05	PI099	RBX2660	180	*E.coli*	3
R1-20	DI05	PI100	RBX2660	360	*E.coli*	7
R1-20	PI006	PI100	30	360	*E.coli*	5
R1-20	PI006	PI099	30	180	*E.coli*	3
R1-20	PI006	PI005	30	7	*E.coli*	0
R1-20	PI006	PI007	30	60	*E.coli*	1
R1-20	PI100	PI099	360	180	*E.coli*	4
R1-20	PI100	PI005	360	7	*E.coli*	5
R1-20	PI100	PI007	360	60	*E.coli*	6
R1-20	PI099	PI005	180	7	*E.coli*	3
R1-20	PI099	PI007	180	60	*E.coli*	4
R1-20	PI005	PI007	7	60	*E.coli*	1
R2-21	DI29	DI28	RBX2660	RBX2660	*E.coli*	0
R2-21	DI29	PI117	RBX2660	180	*E.coli*	5
R2-21	DI29	PI116	RBX2660	90	*E.coli*	4
R2-21	DI29	PI008	RBX2660	60	*E.coli*	6

R2-21	DI28	PI117	RBX2660	180	*E.coli*	5
R2-21	DI28	PI116	RBX2660	90	*E.coli*	4
R2-21	DI28	PI008	RBX2660	60	*E.coli*	6
R2-21	PI117	PI116	180	90	*E.coli*	5
R2-21	PI117	PI008	180	60	*E.coli*	5
R2-21	PI116	PI008	90	60	*E.coli*	4
R2-21	DI24	DI28	RBX2660	RBX2660	*E.coli*	4
R2-21	DI24	DI29	RBX2660	RBX2660	*E.coli*	4
R2-21	DI24	PI008	RBX2660	60	*E.coli*	4
R2-21	DI24	PI116	RBX2660	90	*E.coli*	6
R2-21	DI24	PI117	RBX2660	180	*E.coli*	5
R0-17	PI012	PI018	0	7	*C.braakii*	10615
R0-17	PI011	PI016	0	7	*E. cloacae/asburiae*	9
R2-18	DI28	PI122	RBX2660	7	*E.coli*	38522
R2-18	DI28	DI24	RBX2660	RBX2660	*E.coli*	3
R2-18	DI28	PI120	RBX2660	0	*E.coli*	38504
R2-18	DI28	DI29	RBX2660	RBX2660	*E.coli*	0
R2-18	PI122	DI24	7	RBX2660	*E.coli*	38519
R2-18	PI122	PI120	7	0	*E.coli*	1832
R2-18	PI122	DI29	7	RBX2660	*E.coli*	38522
R2-18	DI24	PI120	RBX2660	0	*E.coli*	38503
R2-18	DI24	DI29	RBX2660	RBX2660	*E.coli*	3
R2-18	PI120	DI29	0	RBX2660	*E.coli*	38504
R0-15	PI021	PI024	0	7	*VRE*	4687
R2-16	DI30	PI127	RBX2660	0	*E.coli*	798
R2-16	PI127	PI129	0	60	*E.coli*	798
R2-16	DI30	PI129	RBX2660	60	*E.coli*	7
R0-12	PI143	PI151	0	30	*C. freundii*	21
R0-12	PI143	PI146	0	30	*C. freundii*	21
R0-12	PI146	PI151	30	30	*C. freundii*	13679

R0-12	PI140	PI148	0	30	*C. werkmanii*	9
R0-12	PI152	PI145	30	30	*VRE*	89
R0-12	PI138	PI145	0	30	*VRE*	15
R0-12	PI138	PI152	0	30	*VRE*	16
R0-12	PI142	PI149	0	30	*E.coli*	17
R2-12	PI155	PI157	3	15	*E. aerogenes*	47
R2-12	PI155	PI158	3	30	*E. aerogenes*	95
R2-12	PI157	PI158	15	30	*E. aerogenes*	68
R2-11	PI034	PI037	0	7	*C. werkmanii*	2
R2-10	PI040	PI043	0	7	*VRE*	2
R2-11	PI040	PI044	0	60	*VRE*	6
R2-12	PI040	PI160	0	180	*VRE*	5
R2-13	PI043	PI160	7	180	*VRE*	3
R2-14	PI043	PI044	7	60	*VRE*	4
R2-15	PI160	PI044	180	60	*VRE*	7
R2-13	DI27	DI26	RBX2660	RBX2660	*E.coli*	1
R2-13	DI27	PI033	RBX2660	60	*E.coli*	1
R2-13	DI26	PI033	RBX2660	60	*E.coli*	0
R2-13	DI25	DI26	RBX2660	RBX2660	*E.coli*	1
R2-13	DI25	DI27	RBX2660	RBX2660	*E.coli*	2
R2-13	DI25	PI033	RBX2660	60	*E.coli*	1
R2-10	DI06	PI045	RBX2660	60	*E.coli*	64151
R0-09	PI166	PI170	0	30	*C. farmerii*	5
R0-09	PI167	PI169	0	7	*M. morganii*	6
R1-07	PI052	PI053	0	15	*C. farmerii*	6
R1-06	PI184	PI186	0	7	*C. freundii*	0
R1-09	DI11	PI047	RBX2660	7	*E.coli*	14
R1-09	DI11	PI048	RBX2660	30	*E.coli*	11
R1-09	DI11	PI049	RBX2660	60	*E.coli*	15
R1-09	DI11	PI175	RBX2660	180	*E.coli*	15
R1-09	DI11	PI176	RBX2660	360	*E.coli*	18

R1-09	PI175	PI047	180	7	*E.coli*	5
R1-09	PI175	PI049	180	60	*E.coli*	4
R1-09	PI175	PI048	180	30	*E.coli*	4
R1-09	PI175	PI176	180	360	*E.coli*	3
R1-09	PI047	PI049	7	60	*E.coli*	5
R1-09	PI047	PI048	7	30	*E.coli*	5
R1-09	PI047	PI176	7	360	*E.coli*	8
R1-09	PI049	PI048	60	30	*E.coli*	4
R1-09	PI049	PI176	60	360	*E.coli*	7
R1-09	PI048	PI176	30	360	*E.coli*	7
R0-07	PI193	PI195	7	30	*E.coli*	59329
R0-07	PI192	PI193	0	7	*E.coli*	4
R0-07	PI192	PI195	0	30	*E.coli*	59325
R0-06	PI054	PI055	0	7	*C. amalonaticus*	0
R1-05	PI056	DI01	0	RBX2660	*E.coli*	59228
R1-05	PI057	DI01	7	RBX2660	*E.coli*	56521
R1-05	PI057	PI056	7	0	*E.coli*	6483
R2-05	PI058	DI09	7	RBX2660	*E.coli*	8
R2-05	PI058	PI212	7	180	*E.coli*	1
R2-05	PI058	PI059	7	60	*E.coli*	1
R2-05	PI058	DI08	7	RBX2660	*E.coli*	2
R2-05	DI09	PI212	RBX2660	180	*E.coli*	7
R2-05	DI09	PI059	RBX2660	60	*E.coli*	7
R2-05	DI09	DI08	RBX2660	RBX2660	*E.coli*	8
R2-05	PI212	PI059	180	60	*E.coli*	0
R2-05	PI212	DI08	180	RBX2660	*E.coli*	1
R2-05	PI059	DI08	60	RBX2660	*E.coli*	1
R0-03	PI060	PI062	0	7	*E.coli*	17
R0-03	PI060	PI063	0	7	*E.coli*	18
R0-03	PI060	PI065	0	7	*E.coli*	9
R0-03	PI060	PI067	0	30	*E.coli*	10

R0-03	PI060	PI068	0	30	*E.coli*	13
R0-03	PI060	PI069	0	60	*E.coli*	12
R0-03	PI060	PI061	0	0	*E.coli*	71906
R0-03	PI060	PI066	0	30	*E.coli*	17825
R0-03	PI062	PI061	7	0	*E.coli*	71901
R0-03	PI062	PI069	7	60	*E.coli*	10
R0-03	PI062	PI065	7	7	*E.coli*	7
R0-03	PI062	PI063	7	7	*E.coli*	14
R0-03	PI062	PI068	7	30	*E.coli*	11
R0-03	PI062	PI066	7	30	*E.coli*	17677
R0-03	PI062	PI067	7	30	*E.coli*	8
R0-03	PI061	PI069	0	60	*E.coli*	71898
R0-03	PI061	PI065	0	7	*E.coli*	71897
R0-03	PI061	PI063	0	7	*E.coli*	71904
R0-03	PI061	PI068	0	30	*E.coli*	71899
R0-03	PI061	PI066	0	30	*E.coli*	65471
R0-03	PI061	PI067	0	30	*E.coli*	71896
R0-03	PI069	PI065	60	7	*E.coli*	3
R0-03	PI069	PI063	60	7	*E.coli*	8
R0-03	PI069	PI068	60	30	*E.coli*	5
R0-03	PI069	PI066	60	30	*E.coli*	17674
R0-03	PI069	PI067	60	30	*E.coli*	4
R0-03	PI065	PI063	7	7	*E.coli*	9
R0-03	PI065	PI068	7	30	*E.coli*	4
R0-03	PI065	PI066	7	30	*E.coli*	17671
R0-03	PI065	PI067	7	30	*E.coli*	1
R0-03	PI063	PI068	7	30	*E.coli*	5
R0-03	PI063	PI066	7	30	*E.coli*	17676
R0-03	PI063	PI067	7	30	*E.coli*	10
R0-03	PI068	PI066	30	30	*E.coli*	17673
R0-03	PI068	PI067	30	30	*E.coli*	5

R0-03	PI066	PI067	30	30	*E.coli*	17672
R1-03	DI10	PI072	RBX2660	7	*E.coli*	0
R1-03	DI10	PI074	RBX2660	30	*E.coli*	2
R1-03	DI10	PI075	RBX2660	60	*E.coli*	4
R1-03	DI10	PI223	RBX2660	180	*E.coli*	33
R1-03	PI075	PI223	60	180	*E.coli*	33
R1-03	PI075	PI072	60	7	*E.coli*	4
R1-03	PI075	PI074	60	30	*E.coli*	2
R1-03	PI223	PI072	180	7	*E.coli*	33
R1-03	PI223	PI074	180	30	*E.coli*	31
R1-03	PI072	PI074	7	30	*E.coli*	2
R0-02	PI076	PI078	0	7	*E.coli*	9
R0-02	PI076	PI080	0	30	*E.coli*	8
R0-02	PI076	PI082	0	60	*E.coli*	12
R0-02	PI082	PI080	60	30	*E.coli*	4
R0-02	PI082	PI078	60	7	*E.coli*	5
R0-02	PI080	PI078	30	7	*E.coli*	3
R2-03	PI226	PI229	30	50	*VRE*	0
R2-03	PI226	PI228	30	40	*VRE*	0
R2-03	PI229	PI228	50	40	*VRE*	0
R2-03	PI224	PI226	0	30	*VRE*	2
R2-03	PI224	PI228	0	40	*VRE*	2
R2-03	PI224	PI229	0	50	*VRE*	2
R2-02	PI083	PI084	0	7	*VRE*	1
R2-02	DI04	PI086	RBX2660	60	*E.coli*	2
R2-02	DI04	PI235	RBX2660	120	*E.coli*	4
R2-02	PI086	PI235	60	120	*E.coli*	4
R1-01	DI07	PI093	RBX2660	7	*E.coli*	0
R1-01	DI07	PI094	RBX2660	30	*E.coli*	5
R1-01	DI07	PI095	RBX2660	60	*E.coli*	6
R1-01	DI07	PI237	RBX2660	360	*E.coli*	20

R1-01	PI093	PI095	7	60	*E.coli*	6
R1-01	PI093	PI094	7	30	*E.coli*	5
R1-01	PI093	PI237	7	360	*E.coli*	20
R1-01	PI095	PI094	60	30	*E.coli*	3
R1-01	PI095	PI237	60	360	*E.coli*	18
R1-01	PI094	PI237	30	360	*E.coli*	19

Table 2.3 NCBI references

species	strain	GenBank assembly
VRE	108	GCA_000395825.1
VRE	110	GCA_000395845.1
VRE	13	GCA_000395865.1
VRE	84	GCA_000395885.1
Enterococcus faecalis		GCA_000007785.1_ASM778v1
Escherichia fergusonii	ATCC 35470	GCA_008064915.1
Escherichia coli	DH1	GCA_000023365
Escherichia coli	K-12 substr. MG1655	GCA_000005845.2
Escherichia coli	DH10B	GCA_006352235.1
Escherichia coli	UMNK88	GCA_000212715.2
Escherichia coli	REL606	GCA_000017985.1
Escherichia coli	101.1	GCA_000168095.1
Escherichia coli	EC4115	GCA_000021125.1
Escherichia coli	EDL933	GCA_000732965.1
Escherichia coli	IAI39	GCA_000026345.1
Escherichia coli	UMN026	GCA_000026325.2
Escherichia coli	IAI1	GCA_000026265.1
Escherichia coli	11128	GCA_000010765.1
Escherichia coli	APEC O78	GCA_000332755.1
Escherichia coli	E110019	GCA_000167875.1
Escherichia coli	B7A	GCA_000725265.1
Escherichia coli	12009	GCA_000010745.1
Escherichia coli	E22	GCA_000167855.1
Escherichia coli	SE11	GCA_000010385.1
Escherichia coli	W	GCA_000184185.1
Escherichia coli	SE15	GCA_000010485.1
Escherichia coli	ED1A	GCA_000026305.1
Escherichia coli	UTI89	GCA_000013265.1
Escherichia coli	S88	GCA_000026285.2
Escherichia coli	F11	GCA_000167835.1

Table 2.4 Double disk test results

Sample	Time point	Identification	Ceftazidime	Ceftazidime / Clavulanic Acid	Difference in Zone Sizes	Interpretation	Cefotaxime	Cefotaxime / Clavulanic Acid	Zone Size Difference (mm)	Interpretation
DI11	Donor	*E. coli*	29	32	3	Negative	14	32	18	Positive
PI047	7	*E. coli*	29	32	3	Negative	15	33	18	Positive
PI048	30	*E. coli*	29	32	3	Negative	14	31	17	Positive
PI049	60	*E. coli*	29	33	4	Negative	16	32	16	Positive
PI175	180	*E. coli*	30	33	3	Negative	14	32	18	Positive
PI176	360	*E. coli*	31	34	3	Negative	18	33	15	Positive
Quality Control										
Negative		*E. coli* 25922	33	33	0	<2mm increase in zone diameter	37	37	0	<2mm increase in zone diameter
Positive		*K. pneumoniae* 700603	15	28	13	≥ 3mm increase in zone diameter	27	36	9	≥ 5mm increase in zone diameter

Chapter 3

Genomic Analyses of Longitudinal *Mycobacterium abscessus* Isolates in a Multi-Center Cohort Reveal Parallel Signatures of In-Host Adaptation

3.1 Abstract

Nontuberculous mycobacteria (NTM) are ubiquitous in the environment and are increasingly causing opportunistic infections. *Mycobacterium abscessus* complex (MAB) is one of the major NTM lung pathogens which disproportionately colonize and infect the lungs of individuals with cystic fibrosis (CF). MAB can persist in the lungs of these individuals for years, and antimicrobial treatment is frequently ineffective. Understanding the in-host adaptation of MAB in people who are chronically colonized or infected has the potential to inform new and future approaches to development of novel therapies. Here, we leveraged a cohort of 175 longitudinal isolates from 30 patients with MAB lung infection in two hospital centers to identify genomic markers of in-host adaptation. Utilizing isolate whole genome sequencing, we quantified the relatedness of isolates both within our cohort and in the broader global context of MAB genomes and found highly related isolate pairs across different hospital centers, despite low likelihood of transmission. We further investigated genes undergoing parallel adaptation in the host lung environment and demonstrated reduced macrolide susceptibility co-occurring with *whiB1* mutations. Finally, we characterized a 23kb mercury resistance plasmid found in two isolates, whose loss confers phenotypic susceptibility to organic and non-organic mercury compounds, suggesting adaptation to the low-mercury lung environment.

3.2 Introduction

Nontuberculous *Mycobacterium* spp. (NTM) are a diverse group of mycobacteria outside the *Mycobacterium tuberculosis* and *Mycobacterium leprae* complexes[1]. Commonly found in soil and water[2], NTM are mostly considered environmental saprophytes. However, NTM can cause opportunistic infection, particularly in humans who are immunocompromised or have preexisting lung conditions such as chronic obstructive pulmonary disease (COPD), cystic fibrosis (CF), or non-CF bronchiectasis[3]. Within the NTM, *Mycobacterium abscessus* complex (MAB) disproportionately colonize and infect patients with CF[4-5]. Colonization and infection can persist for years, and even with prolonged multidrug therapy[6] only ~30% of patients experience treatment success[7]. With infection rates increasing worldwide[4], MAB pose an imminent public health challenge.

The *M. abscessus* complex comprises three subspecies: *Mycobacterium* subsp. *bolletii*, *Mycobacterium* subsp. *massiliense*, and *Mycobacterium* subsp. *abscessus*[8]. Of these, 70% of global clinical isolates have been shown to belong to three clusters of dominant circulating clones (DCCs)[9]. However, it is unclear how frequently infection is caused by direct transmission between patients relative to independent environmental acquisition. Recent studies demonstrating the intercontinental presence of highly related pairs of clinical isolates (less than 20 single nucleotide polymorphisms; SNPs) suggest transmission between patients may occur more often than previously thought[10]. DCCs have also been found to exhibit elevated drug resistance and intracellular survival[9], suggestive of adaptation to a pathogenic lifestyle. However, direct epidemiological

evidence of transmission is often lacking[11-12]. It remains unclear whether the prevalence of DCCs is a result of recent transmission, widespread environmental presence, sampling bias, a slower mutation rate among DCCs maintaining fitness benefits in the lung milieu– or a combination of these factors. Thus, understanding how this environmental species survives and persists in the host is critical for elucidating MAB transmission and evolution.

Despite their ability to cause chronic infection, little is known of the in-host adaptive behavior of MAB. MAB is intrinsically resistant to many antimicrobials[13,14] but can also acquire antimicrobial resistance through SNPs in ribosomal rRNA genes *rrl* and *rrs*[14]. A smooth-to-rough morphotype switch involving loss of glycopeptidolipid (GPL) on the cell surface has been shown to correspond with heightened virulence[15], but the precise genetic factors involved are unknown. To date, most studies investigating in-host adaptation of MAB have been limited to isolates from a single patient[16,17]. One study querying longitudinal samples from 18 patients found evidence for convergent evolution in 30 genes, including the GPL locus and virulence regulators[10]. Some of these mutations were demonstrated to impair survival on fomites[10], suggesting fitness trade-offs between the environment and host. Thus, utilizing high-resolution sequencing to explicitly investigate MAB in-host adaptation may inform prevention strategies and effective treatment development.

Here, we leveraged a multi-center cohort of 175 isolates, longitudinally collected from 30 patients with MAB infection who were treated at academic medical centers in the United States between 2002 and 2020. Through whole genome sequencing (WGS)

we first contextualized the genomic relatedness of isolates, incorporating an additional 1,455 published MAB genomes for comparative genomic analyses. Next, we investigated within-lineage genomic diversity and characterized parallel in-host adaptation. Finally, we probed differences in antimicrobial resistance and mercury compound resistance correlated with specific genomic variance. This work provides high-resolution insight into MAB adaptive mechanisms and identifies novel candidate genes for parallel adaptation. We highlight how MAB adapts to the host milieu while shedding obsolete functions, furthering our understanding of MAB evolution during chronic infection.

3.3 Results

3.3.1 Cohort comprises two major subspecies of MAB spanning the global phylogeny

To understand the phylogenetic relationships in the cohort, we annotated the genomes of all 175 clinical isolates, as well as 3 NCBI reference genomes for each MAB subspecies (subsp. *abscessus*, subsp. *massiliense*, subsp. *bolletii*). We then aligned 2,693 core genes to generate a maximum likelihood phylogenetic tree (Figure 3.1a, Table 3.1). Most isolates (74.9%, 131/175) belonged to subsp. *abscessus*, and a smaller group of subsp. *massiliense* genomes (25.1%, 44/175) was also identified. No subsp. *bolletii* were present in the cohort. Therefore, the tree was re-rooted with the subsp. *bolletii* reference genome as the outgroup.

We then calculated pairwise average nucleotide identity (ANI) for all isolate pairs (Figure 3.2). All isolate pairs showed at least 97% ANI, fulfilling the genomic gold

standard for microbial species[18]. Again, two major clusters were identified corresponding to each of the major subspecies present in the cohort. Pairwise comparisons within the same subspecies were at least 98.49% ANI (Figure 3.1b).

To contextualize our cohort within the global phylogeny, we conducted another core genome alignment (1,423 core genes), incorporating 1,452 additional published MAB genomes downloaded from NCBI and the European Nucleotide Archive (ENA) for a total of 1,630 genomes included in the analysis (Figure 3.3a, Table 3.1, genomes downloaded Dec. 28, 2021). Most genomes in this larger cohort belonged to subsp. *abscessus* (72,1%, 1175/1630), followed by subsp. *massiliense* (27.2%, 444/1630). There were 8 subsp. *bolletii* genomes (0.49%, 8/1630), all contributed from downloaded genomes. The 175 isolates in our study were broadly distributed throughout the species phylogeny, suggesting a large range of genomic diversity is represented in the study cohort. ANI was also measured against each of the 3 subspecies reference genomes, and again the same subspecies were determined to be at least 98.5% ANI against a reference genome (Figure 3.3b).

3.3.2 Genomic relatedness indicates within-patient diversity of subspecies and lineages

To define lineages, we measured the relatedness of intra-patient isolate pairs in the 175 isolate core genome alignment. When looking at the overall distribution of pairwise core genome SNP distances, isolate pairs > 40,000 SNPs apart belonged to different subspecies (Figure 3.1c, Figure 3.4). While same-patient isolate pairs generally exhibited low core genome SNP distance (median: 134; mean: 1,671; range: 1- 62,011), pairs of

isolates belonging to the same subspecies but corresponding to large SNP distances (10,000-14,000 SNPs) were also identified, pointing to diverse MAB populations within some hosts (Figure 3.1c).

To further quantify within-host diversity, we applied pairwise ANI as an orthogonal approach and found that at a 99.99% ANI cutoff, we could distinguish clusters of highly related isolates (Figure 3.1c, Figure 3.5a). While most of these clusters comprised isolates from a single patient, there were also groups of closely related isolates found across multiple patients (L1-L4, Figure 3.5a). Interestingly, these multi-patient groups were closely related (<200 core genome SNPs) to published DCC genomes. Whole genome alignment of multi-patient clusters further revealed SNP distances ranging from 7 to 490 SNPs (mean: 177, median: 135). Two patient pairs (MAB_04 and MAB_26, MAB_04 and MAB_30) exhibited DCC1 isolates less than 10 SNPs apart but were treated at different sites. Three other patient pairs exhibited DCC1 and DCC3 isolates less than 38 SNPs apart (suggested as an indicator of possible transmission[9]), with one pair (MAB_25 and MAB_05) treated at different sites in different years, one pair (MAB_26 and MAB_30) treated at the same site in different years, and just one pair (MAB_06 and MAB_17) at the same site with overlapping treatment years. However, there were no further metadata or documented hospital outbreaks. Overall, the observation of highly related isolate pairs from distinct locations and years suggests presence of nearly clonal DCCs on a broad geographic scale.

To contextualize the evolutionary history of MAB lineages, we generated maximum parsimony trees using PHYLIP[19]. From these trees' ancestral nodes, we were

able to infer the distance from the Most Recent Common Ancestor (dMRCA) to be an average of 4.82 SNPs (95% CI [2.41, 7.23]). We then applied the estimated molecular clock for each corresponding subspecies[8] (1.8 SNPs/year for subsp. *abscessus*, 0.46 SNPs/year for subsp. *massiliense*) and calculated the estimated average time from Most Recent Common Ancestor (tMRCA) to be 3.72 years (95% CI [2.18,5.34]). This value corresponded closely to the actual average time since initial positive NTM isolate culture of 3.98 years (95% CI [2.75, 5.21]), lending credence to the accuracy of our genomic measurements.

3.3.3 Antimicrobial Resistance Genes are prevalent and conserved within lineages

Next, we sought to characterize the antimicrobial resistance genes (ARGs) that may confer a survival advantage to MAB in the host. We found that 100% (175/175) of isolates carry *bla*, *arr*, and *cmx_cmrA* genes, which confer resistance to beta lactams, rifamycin, and chloramphenicol, respectively. 92.6% (162/175) of isolates carried the *aph(3'')* gene conferring resistance to streptomycin. 74.3% (130/175) of isolates, all subsp. *abscessus*, carried the *erm(41)* gene for macrolide (induced clarithromycin) resistance. We did not observe substantial within-lineage variance for ARGs, and in 94.1% of cases (32/34 lineages) whole lineages were identical in predicted ARG profile.

3.3.4 Lineages undergo parallel within-host adaptation in Mycobacterial virulence genes

We sought to identify adaptive mutations that may have conferred fitness advantages for long-term survival in the host. For each lineage, we identified polymorphisms by aligning isolate reads against the genome from the lineage's earliest collection time and annotating whole genome SNPs and insertions/deletions (indels). At this higher resolution, isolate pairs from the same lineage and patient were on average 10 SNPs apart (95% CI [8.74, 11.3]).

In total we found 29 genes mutated in parallel across multiple patients and lineages (Figure 3.5b, Table 3.2). We then applied a permutation test with 10,000 iterations to assess the significance (non-randomness) of these parallel findings (Figure 3.5b, Table 3.2). 79% (23/29) of the genes were also significant by the permutation test (*P*-value < 0.05, Benjamini Hochberg). Some of these genes were also found to be variable in isolates collected just 9 days apart from multiple body sites (Figure 3.5c). Significant hits included genes implicated in mycobacterial virulence and drug resistance, such as the PE/PPE family immunomodulator PE5[20] and the anti-tuberculosis drug target EmbC[21]. A number of these genes were previously reported in the literature as evidence of within-host parallel evolution, including *crp*, *embC*, *whiB1*, and *espR*[10], suggesting diverse populations of MAB undergo similar adaptive trajectories during chronic infection.

3.3.5 Within-lineage diversity affects phenotypic antimicrobial resistance

We sought to test the phenotypic effects of our parallel mutated genes by examining a lineage from patient MAB_18, which featured variability in the *whiB1* gene (Figure

3.6a). *whiB1* (MAB_3539) encodes for a nitric oxide-sensitive transcriptional repressor in the WhiB family of proteins, and has been implicated in regulation of the ESX-1 secretion system[22]. Deletion of the *whiB7* gene has been reported to confer sensitivity to the ribosome targeting drugs amikacin, clarithromycin, erythromycin, tetracycline, and spectinomycin[23]. In MAB_18, three isolates exhibited a nonsynonymous *whiB1* mutation resulting in a glycine (Gly) to alanine (Ala) switch at the Gly24 locus conserved between *M. tuberculosis* H37rV and *M. abscsessus* ATCC 19977. We hypothesized that the nonsynonymous *whiB1* mutation at this conserved locus would affect clinical isolates' antimicrobial susceptibility. Thus, we measured their MICs in amikacin, erythromycin, and clarithromycin using a resazurin microplate assay[24,25]. We observed that the three isolates with a *whiB1* mutation were significantly more susceptible to erythromycin and clarithromycin, but not amikacin (Figure 3.6b, clarithromycin P=0.00105, erythromycin P =0.000789, Kruskal-Wallace rank sum test), demonstrating within-lineage variability in phenotypic macrolide resistance corresponding to *whiB1* mutation.

3.3.6 Loss of 23kb mercury resistance plasmid is linked to mercury susceptibility

We observed the loss of a 23kb mercury resistance plasmid in isolates from patient MAB_14. This patient had seven isolates belonging to two distinct lineages (Figure 3.7a). The initial isolate MAB_14_01 genome contained a 23kb mercury resistance plasmid identical to one in ATCC19977, reported to have originated in *M. marinum*[26,27] and predicted to encode a 472 amino acid (AA) MerA protein and 218 AA MerB protein. This plasmid was not present in subsequent isolate genomes of the same

lineage (MAB_14_02 through MAB_14_05). The MAB_14_06 and MAB_14_07 genomes of a separate lineage contained a 7kb contig carrying both predicted 474 AA MerA and 219 AA MerB (Figure 3.7b). Genomes without MerA contained a predicted 281 AA MerB in the chromosome, in a region encoding for cell wall components MmpL4 and PPE4 (Figure 3.7c). This region was highly conserved across all subsp. *abscessus* isolates.

MerA is a mercuric reductase which reduces inorganic mercury Hg(II) to Hg(0), while MerB is an organic mercury lysase which cleaves the Hg-C bond in organic mercury compounds to generate inorganic mercury Hg(II) (Figure 3.7d). We sought to validate the phenotypic differences between isolates in MAB_14 that had different combinations and alleles of *merA* and *merB*, by conducting disc diffusion assays with an inorganic mercury compound, mercury chloride ($HgCl_2$), and an organic mercury compound, phenylmercury acetate (PMA). We found striking differences in phenotypic resistance to both compounds between the two genotypes (Figure 3.7e). Isolates encoding *merA* had higher resistance to both PMA and $HgCl_2$, but this difference was significant only at the highest concentration of the inorganic compound, $HgCl_2$ (Kruskal-Wallis test, P=0.011). In contrast, resistance to the organic compound PMA was significantly higher in the *merA* isolates across three different concentrations (Kruskal-Wallis test, P=0.011, 0.0064, 0.0016). Despite *merB* being present in all tested isolates, the resistance to the organic mercury compound PMA was lower among isolates with just the chromosomal copy, compared to isolates with an additional copy of *merB* accompanied by *merA*.

3.4 Discussion

The possibility of patient-to-patient or fomite-directed transmission of MAB has been debated. Here, we found instances of highly related isolate pairs (<20 whole genome SNPs) across different hospital centers as well as within the same center, but no additional data to suggest an outbreak. Considering how well our cohort encompasses the global MAB phylogeny, these occurrences likely indicate widely circulating lineages of MAB acquired through separate infection events, as have been reported in similar studies[11-12]. In the absence of recent transmission, the high genomic relatedness of these isolates could be explained by a slow rate of mutation among highly successful host-adapted pathogens. Whether this is a species-wide trend towards obligate pathogenicity (as was the case for *M. tuberculosis*)[10], or driving a chasm between clinical and environmental populations of MAB, warrants exploration. Further research on the species' molecular clock and genomic comparisons with environmental isolates will provide more clarity on the transmission dynamics and evolutionary trajectory of MAB.

We identify candidate genes as hotspots of in-host adaptation, potentially conferring advantages for survival within the lung milieu. It is possible that even greater diversity of MAB was present but not captured due to limitations in study design. Here we only obtained one isolate per timepoint, but there may be additional co-existing lineages, or lineages that emerged prior to the dates to which we attributed them. Furthermore, CF patients exhibit polymicrobial lung infections[28], and cross-species interactions such as horizonal gene transfer or competition may occur. Future

studies may capture more diversity by picking multiple colonies[29], conducting plate sweeps[9,12], or conducting metagenomic sequencing of sputum samples[30,31].

We demonstrated *in vitro* that within-lineage variations are associated with diverse phenotypic susceptibility to drugs. We found that a nonsynonymous mutation at a conserved site in *whiB1* was associated with increased susceptibility to clarithromycin and erythromycin, but not amikacin. The patient was on azithromycin therapy for CF at earlier timepoints (MAB_18_01 through MAB_18_05), while latter isolates (MAB_18_07 through MAB_18_09) were exposed to azithromycin and cefoxitin, and briefly amikacin (Figure 3.8). Exposure to amikacin may have contributed to sustained resistance, while azithromycin treatment did not induce widespread macrolide resistance. WhiB1 is a repressor regulating the mycobacterial ESX-1 system, which disrupts the innate immune response by targeting host membranes[35]. It is possible that *whiB1*, similar to *whiB7*[28], regulates antimicrobial resistance as well as virulence. Thus, the observed *whiB1* mutation may pose a trade-off between increased macrolide susceptibility and greater immunomodulation. This trade-off may present an opportune window for greater macrolide susceptibility, informing effective treatment options. A greater understanding of the regulatory pathways of *whiB1* in MAB is required to test this hypothesis.

Finally, we demonstrate how the loss of a 23kb mercury resistance plasmid is correlated with increased susceptibility to both organic and inorganic mercury compounds. Mercury exists naturally in the soil, water, and atmosphere, and cycles on a global scale through both anthropogenic and natural processes such as industrial

wastewater, landfills, burning fossil fuels, and processing through microorganisms[33]. Bacteria use MerA and MerB to break down organomercury into Hg(II), and subsequently to the much less toxic elemental mercury (Hg(0)), which is highly volatile and rapidly diffused out of the bacterial cell[34,35]. A study of *Arabidopsis thaliana* found that insertion of bacterial *merA* and *merB* genes together conferred tenfold higher resistance to organic methylmercury than insertion of the *merB* gene alone[34]. These findings reflect our own observations, and potentially signify that both genes are imperative to successfully break down and remove mercury from the bacterial cell. Loss of the mercury resistance plasmid, then, is likely due to the fitness trade-off of maintaining the plasmid in the low-mercury pulmonary environment[36,37].

In this study we highlight genomic processes through which MAB adapts to promote its own survival within the host. Many of these events occur in parallel across patients and hospital sites and include DCCs circulating on a global scale. In the absence of evidence of recent transmission, we suggest highly infectious strains of MAB exhibit low rates of mutation to maintain a pathogen lifestyle. Further, the within-lineage polymorphisms we observed have phenotypic effects, potentially benefiting fitness in the host, at the putative detriment of environmental survival. This work thus contributes to our understanding of in-host survival of MAB and may inform development of treatment strategies against these chronic infections.

3.5 Methods

3.5.1 Isolate collection

122 isolates from 22 patients were recovered from clinical specimens collected as a part of routine clinical care at the Barnes-Jewish Hospital (BJH) microbiology laboratory. Another 53 isolates from 8 patients were obtained from clinical samples at Michigan Medicine (University of Michigan (UM)) (Table 3.1). Specimens were primarily respiratory and included MAB isolated from sputum, tracheal aspirates, and bronchial alveolar lavage fluids. Isolates were cultured onto Middlebrook 7H11 agar (Hardy Diagnostics, Santa Maria, CA) and incubated at 35°C (BJH) or 37°C in air (Michigan). Isolates recovered from these specimens were stored at -80°C. For isolates at BJH, the identity of each isolate was confirmed using Vitek matrix-assisted laser desorption ionization–time of flight mass-spectrometry (MALDI-TOF MS) with Knowledge Base 3.0 (bioMérieux, Durham, NC, USA) as previously described[38].

3.5.2 DNA extraction and sequencing

Isolate DNA was extracted using the QIAamp BiOstic bacteremia DNA kit (Qiagen, Germantown, MD, USA) using manufacturer instructions, adjusting for a 2-minute mechanical lysis step (Mini-Beadbeater-24; BioSpec, Bartlesville, OK, USA) at the start of the protocol. Sequencing libraries were prepared using 0.5 ng genomic DNA using the Nextera XT kit (Illumina, San Diego, CA, USA) and methods from Baym et al.[39] Pooled libraries were sequenced on a NextSeq500 System (Illumina, San Diego, CA, USA) to 2.5 million paired-end reads (2 x 150 bp).

3.5.2 Genome assembly and annotation

Published genomes and sequence reads were downloaded from NCBI (PRJNA398137 and PRJNA523365) and ENA (ERP001039) to capture multiple studies from different geographic sites. Demultiplexed reads were trimmed using Trimmomatic v0.38[40] (leading, 10; trailing, 10; sliding window, 4:15; minimum length, 60) and assembled using Unicycler v0.4.7[41] with default parameters. Genes were annotated using Prokka v1.12[42] (default parameters, contigs > 500 bp). All genomes were queried using CheckM 1.0.7[43], and only assemblies with >95% completeness, >5% strain heterogeneity, and < 2% contamination were included. Antimicrobial resistance genes were queried using AMRfinder v3.10.16[44] and Resfinder v4.0[45]. MLST was queried using mlst v2.19.0[4647] under the 'mabscessus' scheme.

3.5.3 Core genome alignment and Average Nucleotide Identity

Core genome alignments were conducted using Roary v3.12.0[48] (-cd 100; -n; -e; -i 85) and .gff files from Prokka as input. Resulting alignments were converted to Newick trees using FastTree v2.1.10[49] and visualized on iTOL v5[50]. Core genome SNP distances were calculated using SNP-sites v2.4.0.[51] and visualized using ggplot2[52] on R v3.6.3[53]. Pairwise average nucleotide identity (ANI) was calculated using dnadiff on MUMmer v4.0.0[54]. Heatmaps were visualized using pheatmap package[55] on R v.3.6.3[53]. Networks were visualized by filtering for at least 99.99% ANI on Cytoscape v3.8.0[56].

3.5.4 Characterization of within-lineage diversity

Lineage-specific alignments were generated by first creating custom indices for each lineage's temporally initial isolate in Bowtie 2 v2.3.5[57] using the bowtie2-build command. Subsequent isolate reads were then aligned against the corresponding index (-X 2000; -no-mixed; -very-sensitive -n-ceil 0,0.01). Resulting alignments were annotated for SNPs and insertions/deletions using SAMtools v1.12[58] and BCFtools v1.9[58] (bcftools call -c DP>10 QS>0.95; bcftools view -i FQ<-85). Unrooted SNP trees were visualized in R v3.6.3[53] using the ape package[59].

For the permutation analysis, mutations were randomly generated across each initial isolate's genome length, and then annotated for which genes they landed on. For each round of permutations, the total number of lineages randomly mutated in parallel for a given gene was noted. This process was repeated for a total of 10,000 permutations to generate a hypothetical distribution. P-values were calculated by calculating the percentile of the actual number of lineages within the hypothetical distribution. Bubble plots were visualized in R v3.6.3[53] using ggplot2[52].

3.5.5 dMRCA

Maximum parsimony trees were generated in PHYLIP v3.697[19] (PHYLogeny Inference Package) using the closest ANI-matching isolate as an outgroup. Whole genome alignments were generated for each lineage with at least two samples and inputted to PHYLIP for a total of 30 trees. The average branch length from the most recent common ancestor (MRCA; node 1 in the maximum parsimony tree) was derived as average distance to MRCA (dMRCA). These values were then divided by each subspecies'

estimated molecular clock (subsp. *abscessus*: 1.8 SNPs/year; subsp. *massiliense*: 0.46 SNPs/year[8]) to calculate time to MRCA (tMRCA). Isolates from MAB_24 were excluded from this analysis, as the patient had started care outside of the University of Michigan hospital system and initial NTM infection date was unknown. However, average estimates did not change with the inclusion of isolates from MAB_24 (average 4.82 SNPs 95% CI [2.49, 7.15], estimated tMRCA 3.72 years, 95% CI [2.19, 5.25]).

3.5.6 Antimicrobial resistance assays

Isolates were inoculated into 7H9 broth supplemented with OADC. After growth to mid-log phase in 37 °C in air, suspensions were diluted to 0.05 OD_{600} in 7H9, corresponding to approximately 10^8 CFU. In a 96-well plate, serial dilutions of each drug were prepared by adding 100 µL antimicrobial solution to 100 µL 7H9 broth. 100 µL of the diluted sample was then added to each well and mixed by pipetting up and down. The plate was placed in a sealed container to grow shaking for 4 days in 37 °C in air, after which 10uL 10% resazurin was added to each well. After 24 hours (shaking in 37 °C in air), the minimum inhibitory concentration was recorded as the minimum concentration observed to inhibit cell growth (growth determined from a color change from blue to pink). ATCC19977 was used as a control strain, along with a positive control row (containing just sample and 7H9 broth), and negative control row (containing just 7H9 broth) for each sample.

3.5.7 Mercury resistance assays

Isolates were grown out and diluted to 0.05 OD_{600} as described above. Mercury resistance was tested in the methodology described by Steingrube et al[60]: 600uL of each isolate suspension was spread on an 7H10 plate supplemented with ADC. 6mm discs were loaded with 20 µL $HgCl_2$ or PMA at tenfold diluted concentrations: 10^{-2}M, 10^{-3}M, 10^{-4}M, 10^{-4}M. A blank disc was added as a negative control. The plates were set to grow for 72 hours at 37°C in air, after which the zone of inhibition was measured.

3.6 Data availability

The sequencing data supporting the conclusions of this article is available in the NCBI repository under Bioproject PRJNA882917.

3.7 References

1. Faria S, Joao I, Jordao L. General Overview on Nontuberculous Mycobacteria, Biofilms, and Human Infection. *Journal of Pathogens*. 2015;2015:e809014. doi:10.1155/2015/809014

2. Falkinham JO. Nontuberculous mycobacteria in the environment. *Clin Chest Med*. 2002;23(3):529-551. doi:10.1016/s0272-5231(02)00014-x

3. Griffith DE, Aksamit T, Brown-Elliott BA, et al. An Official ATS/IDSA Statement: Diagnosis, Treatment, and Prevention of Nontuberculous Mycobacterial Diseases. *Am J Respir Crit Care Med*. 2007;175(4):367-416. doi:10.1164/rccm.200604-571ST

4. Lopeman RC, Harrison J, Desai M, Cox JAG. Mycobacterium abscessus: Environmental Bacterium Turned Clinical Nightmare. *Microorganisms*. 2019;7(3):90. doi:10.3390/microorganisms7030090

5. Levy I, Grisaru-Soen G, Lerner-Geva L, et al. Multicenter Cross-Sectional Study of Nontuberculous Mycobacterial Infections among Cystic Fibrosis Patients, Israel. *Emerg Infect Dis*. 2008;14(3):378-384. doi:10.3201/eid1403.061405

6. Martiniano SL, Nick JA, Daley CL. Nontuberculous Mycobacterial Infections in Cystic Fibrosis. *Thoracic Surgery Clinics*. 2019;29(1):95-108. doi:10.1016/j.thorsurg.2018.09.008

7. Dorn A van. Multidrug-resistant Mycobacterium abscessus threatens patients with cystic fibrosis. *The Lancet Respiratory Medicine*. 2017;5(1):15. doi:10.1016/S2213-2600(16)30444-1

8. Bryant JM, Grogono DM, Greaves D, et al. Whole-genome sequencing to identify transmission of Mycobacterium abscessus between patients with cystic fibrosis: a

retrospective cohort study. *Lancet*. 2013;381(9877):1551-1560. doi:10.1016/S0140-6736(13)60632-7

9. Bryant JM, Grogono DM, Rodriguez-Rincon D, et al. Population-level genomics identifies the emergence and global spread of a human transmissible multidrug-resistant nontuberculous mycobacterium. *Science*. 2016;354(6313):751-757. doi:10.1126/science.aaf8156

10. Bryant JM, Brown KP, Burbaud S, et al. Stepwise pathogenic evolution of Mycobacterium abscessus. *Science*. 2021;372(6541):eabb8699. doi:10.1126/science.abb8699

11. Lipworth S, Hough N, Weston N, et al. Epidemiology of Mycobacterium abscessus in England: an observational study. *Lancet Microbe*. 2021;2(10):e498-e507. doi:10.1016/S2666-5247(21)00128-2

12. Waglechner N, Tullis E, Stephenson AL, et al. Genomic epidemiology of Mycobacterium abscessus in a Canadian cystic fibrosis centre. *Sci Rep*. 2022;12(1):16116. doi:10.1038/s41598-022-19666-8

13. Sanguinetti M, Ardito F, Fiscarelli E, et al. Fatal pulmonary infection due to multidrug-resistant Mycobacterium abscessus in a patient with cystic fibrosis. *J Clin Microbiol*. 2001;39(2):816-819. doi:10.1128/JCM.39.2.816-819.2001

14. Nessar R, Cambau E, Reyrat JM, Murray A, Gicquel B. Mycobacterium abscessus: a new antibiotic nightmare. *J Antimicrob Chemother*. 2012;67(4):810-818. doi:10.1093/jac/dkr578

15. Howard ST, Rhoades E, Recht J, et al. Spontaneous reversion of Mycobacterium abscessus from a smooth to a rough morphotype is associated with reduced expression of

glycopeptidolipid and reacquisition of an invasive phenotype. *Microbiology*. 152(6):1581-1590. doi:10.1099/mic.0.28625-0

16. Lewin A, Kamal E, Semmler T, et al. Genetic diversification of persistent Mycobacterium abscessus within cystic fibrosis patients. *Virulence*. 2021;12(1):2415-2429. doi:10.1080/21505594.2021.1959808

17. Kreutzfeldt KM, McAdam PR, Claxton P, et al. Molecular longitudinal tracking of Mycobacterium abscessus spp. during chronic infection of the human lung. *PLoS One*. 2013;8(5):e63237. doi:10.1371/journal.pone.0063237

18. Richter M, Roselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition | PNAS. *2009*. Accessed June 27, 2022. https://www.pnas.org/doi/full/10.1073/pnas.0906412106

19. Felsenstein J. PHYLIP (Phylogeny Inference Package) version 3.697. Published online 2009.

20. Choo SW, Wee WY, Ngeow YF, et al. Genomic reconnaissance of clinical isolates of emerging human pathogen Mycobacterium abscessus reveals high evolutionary potential. *Sci Rep*. 2014;4(1):4061. doi:10.1038/srep04061

21. Korkegian A, Roberts DM, Blair R, Parish T. Mutations in the Essential Arabinosyltransferase EmbC Lead to Alterations in Mycobacterium tuberculosis Lipoarabinomannan. *J Biol Chem*. 2014;289(51):35172-35181. doi:10.1074/jbc.M114.583112

22. Kudhair BK, Hounslow AM, Rolfe MD, et al. Structure of a Wbl protein and implications for NO sensing by M. tuberculosis. *Nat Commun*. 2017;8(1):2280. doi:10.1038/s41467-017-02418-y

23. Hurst-Hess K, Rudra P, Ghosh P. Mycobacterium abscessus WhiB7 Regulates a Species-Specific Repertoire of Genes To Confer Extreme Antibiotic Resistance. *Antimicrob Agents Chemother*. 2017;61(11). doi:10.1128/AAC.01347-17

24. Garcia de Carvalho NF, Sato DN, Pavan FR, Ferrazoli L, Chimara E. Resazurin Microtiter Assay for Clarithromycin Susceptibility Testing of Clinical Isolates of Mycobacterium abscessus Group. *J Clin Lab Anal*. 2016;30(5):751-755. doi:10.1002/jcla.21933

25. Bich Hanh BT, Quang NT, Park Y, et al. Omadacycline Potentiates Clarithromycin Activity Against Mycobacterium abscessus. *Front Pharmacol*. 2021;12:790767. doi:10.3389/fphar.2021.790767

26. Schué M, Dover LG, Besra GurdyalS, Parkhill J, Brown NL. Sequence and Analysis of a Plasmid-Encoded Mercury Resistance Operon from Mycobacterium marinum Identifies MerH, a New Mercuric Ion Transporter. *Journal of Bacteriology*. 2009;191(1):439-444. doi:10.1128/JB.01063-08

27. Medjahed H, Gaillard JL, Reyrat JM. Mycobacterium abscessus: a new player in the mycobacterial field. *Trends in Microbiology*. 2010;18(3):117-123. doi:10.1016/j.tim.2009.12.007

28. Filkins LM, O'Toole GA. Cystic Fibrosis Lung Infections: Polymicrobial, Complex, and Hard to Treat. *PLOS Pathogens*. 2015;11(12):e1005258. doi:10.1371/journal.ppat.1005258

29. Lieberman TD, Flett KB, Yelin I, et al. Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. *Nat Genet*. 2014;46(1):82-87. doi:10.1038/ng.2848

30. Feigelman R, Kahlert CR, Baty F, et al. Sputum DNA sequencing in cystic fibrosis: non-invasive access to the lung microbiome and to pathogen details. *Microbiome*. 2017;5(1):20. doi:10.1186/s40168-017-0234-1

31. Raghuvanshi R, Vasco K, Vázquez-Baeza Y, et al. High-Resolution Longitudinal Dynamics of the Cystic Fibrosis Sputum Microbiome and Metabolome through Antibiotic Therapy. *mSystems*. 2020;5(3):e00292-20. doi:10.1128/mSystems.00292-20

32. Wong KW. The Role of ESX-1 in Mycobacterium tuberculosis Pathogenesis. *Microbiol Spectr*. 2017;5(3). doi:10.1128/microbiolspec.TBTB2-0001-2015

33. Barkay T, Miller SM, Summers AO. Bacterial mercury resistance from atoms to ecosystems. *FEMS Microbiology Reviews*. 2003;27(2-3):355-384. doi:10.1016/S0168-6445(03)00046-9

34. Bizily SP, Rugh CL, Meagher RB. Phytodetoxification of hazardous organomercurials by genetically engineered plants. *Nat Biotechnol*. 2000;18(2):213-217. doi:10.1038/72678

35. Summers AO. Organization, expression, and evolution of genes for mercury resistance. *Annu Rev Microbiol*. 1986;40:607-634. doi:10.1146/annurev.mi.40.100186.003135

36. Reichl FX, Walther UI, Durner J, et al. Cytotoxicity of dental composite components and mercury compounds in lung cells. *Dental Materials*. 2001;17(2):95-101. doi:10.1016/S0109-5641(00)00029-4

37. Lim HE, Shim JJ, Lee SY, et al. Mercury inhalation poisoning and acute lung injury. *Korean J Intern Med*. 1998;13(2):127-130. doi:10.3904/kjim.1998.13.2.127

38. Comparison of Sample Preparation Methods, Instrumentation Platforms, and Contemporary Commercial Databases for Identification of Clinically Relevant Mycobacteria

by Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry - PubMed. Accessed September 17, 2022. https://pubmed.ncbi.nlm.nih.gov/25972426/

39. Baym M, Kryazhimskiy S, Lieberman TD, Chung H, Desai MM, Kishony R. Inexpensive Multiplexed Library Preparation for Megabase-Sized Genomes. *PLOS ONE*. 2015;10(5):e0128036. doi:10.1371/journal.pone.0128036

40. Trimmomatic: a flexible trimmer for Illumina sequence data | Bioinformatics | Oxford Academic. Accessed July 18, 2022. https://academic.oup.com/bioinformatics/article/30/15/2114/2390096

41. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads | PLOS Computational Biology. Accessed July 18, 2022. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005595

42. Prokka: rapid prokaryotic genome annotation | Bioinformatics | Oxford Academic. Accessed July 18, 2022. https://academic.oup.com/bioinformatics/article/30/14/2068/2390517

43. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Accessed July 18, 2022. https://genome.cshlp.org/content/25/7/1043

44. Feldgarden M, Brover V, Haft DH, et al. Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. *Antimicrobial Agents and Chemotherapy*. 2019;63(11):e00483-19. doi:10.1128/AAC.00483-19

45. Florensa AF, Kaas RS, Clausen PTLC, Aytan-Aktug D, Aarestrup FM. ResFinder – an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. *Microb Genom*. 2022;8(1):000748. doi:10.1099/mgen.0.000748

46. GitHub - tseemann/mlst: Scan contig files against PubMLST typing schemes. Accessed July 18, 2022. https://github.com/tseemann/mlst

47. Jolley KA, Maiden MC. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*. 2010;11(1):595. doi:10.1186/1471-2105-11-595

48. Roary: rapid large-scale prokaryote pan genome analysis | Bioinformatics | Oxford Academic. Accessed July 18, 2022. https://academic.oup.com/bioinformatics/article/31/22/3691/240757

49. Price MN, Dehal PS, Arkin AP. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution*. 2009;26(7):1641-1650. doi:10.1093/molbev/msp077

50. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*. 2021;49(W1):W293-W296. doi:10.1093/nar/gkab301

51. Page AJ, Taylor B, Delaney AJ, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom*. 2016;2(4):e000056. doi:10.1099/mgen.0.000056

52. Hadley W. *Ggplot2: Elegant Graphics for Data Analysis*. Springer International Publishing; 2016. doi:10.1007/978-3-319-24277-4

53. R Core Team. R: A Language and Environment for Statistical Computing. Published online 2020. www.R-project.org

54. Versatile and open software for comparing large genomes | Genome Biology | Full Text. Accessed July 18, 2022. https://genomebiology.biomedcentral.com/articles/10.1186/gb-2004-5-2-r12

55. Kolde R. pheatmap: Pretty Heatmaps. Published online January 4, 2019. Accessed July 18, 2022. https://CRAN.R-project.org/package=pheatmap

56. RCy3: Network biology using Cytoscape from within R | F1000Research. Accessed July 18, 2022. https://f1000research.com/articles/8-1774/v3

57. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357-359. doi:10.1038/nmeth.1923

58. SAMtools/BCFtools/HTSlib - Downloads. Accessed July 18, 2022. https://www.htslib.org/download/

59. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*. 2019;35(3):526-528. doi:10.1093/bioinformatics/bty633

60. Steingrube VA, Wallace RJ, Steele LC, Pang YJ. Mercuric reductase activity and evidence of broad-spectrum mercury resistance among clinical isolates of rapidly growing mycobacteria. *Antimicrob Agents Chemother*. 1991;35(5):819-823.

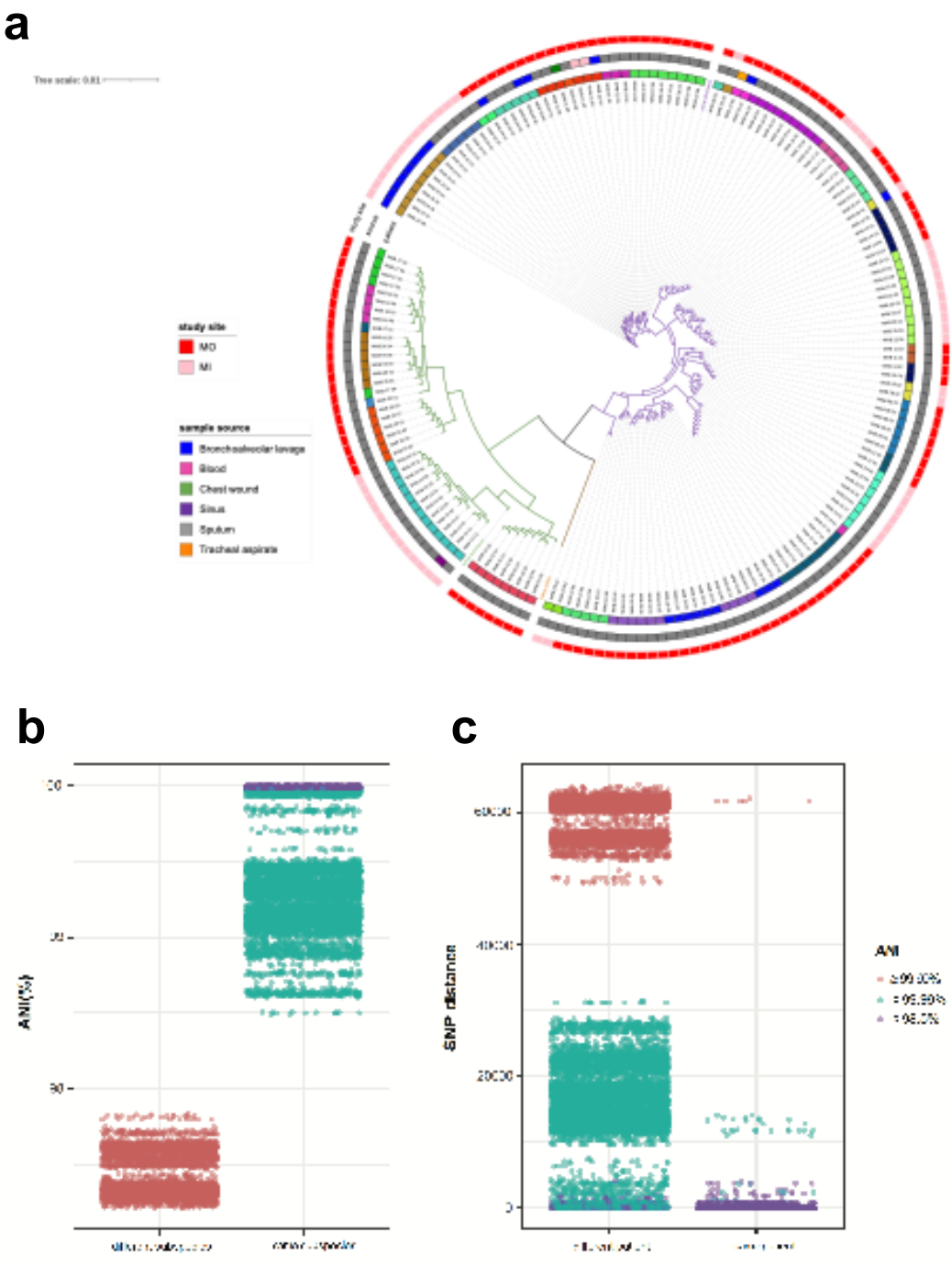
a
study site
MO
MI
sample source
Bronchoalveolar lavage
Blood
Chest wound
Sinus
Sputum
Tracheal aspirate
b
ANI(%)
c
SNP distance
ANI

Figure 3.1 Genomic comparisons of entire cohort (a) Core genome alignment of 175 isolate genomes with 3 reference genomes for each *M. abscessus* subspecies. Aligned with roary, converted with FastTree, visualized with iTOL. 2,693 core genes. Tree is rerooted at the *M. bolletii* reference genome. Branch colors indicate subspecies determined by clade: green for subspecies *massiliense* (*n*=44), purple for subspecies *abscessus* (*n*=131). Outer rings (from inner to outer) denote patient ID, sample source, and study site (MO: Missouri, MI: Michigan, BAL: bronchoalveolar lavage). **(b)** ANI across isolate pairs. X-axis shows comparisons between different subspecies and same subspecies. Y-axis denotes ANI value. Isolate pairs belonging to the same subspecies have pairwise ANI values above 98.5%. **(c)** Core genome SNP distance across isolate pairs. Each point represents a pairwise comparison. Comparisons are grouped as distances between isolates from different patients or the same patient. Points are colored by subspecies comparison as well as corresponding ANI: different subspecies pairs are purple, highly related pairs of at least 99.99% ANI are turquoise, and less related pairs less than 99.99% ANI are salmon.

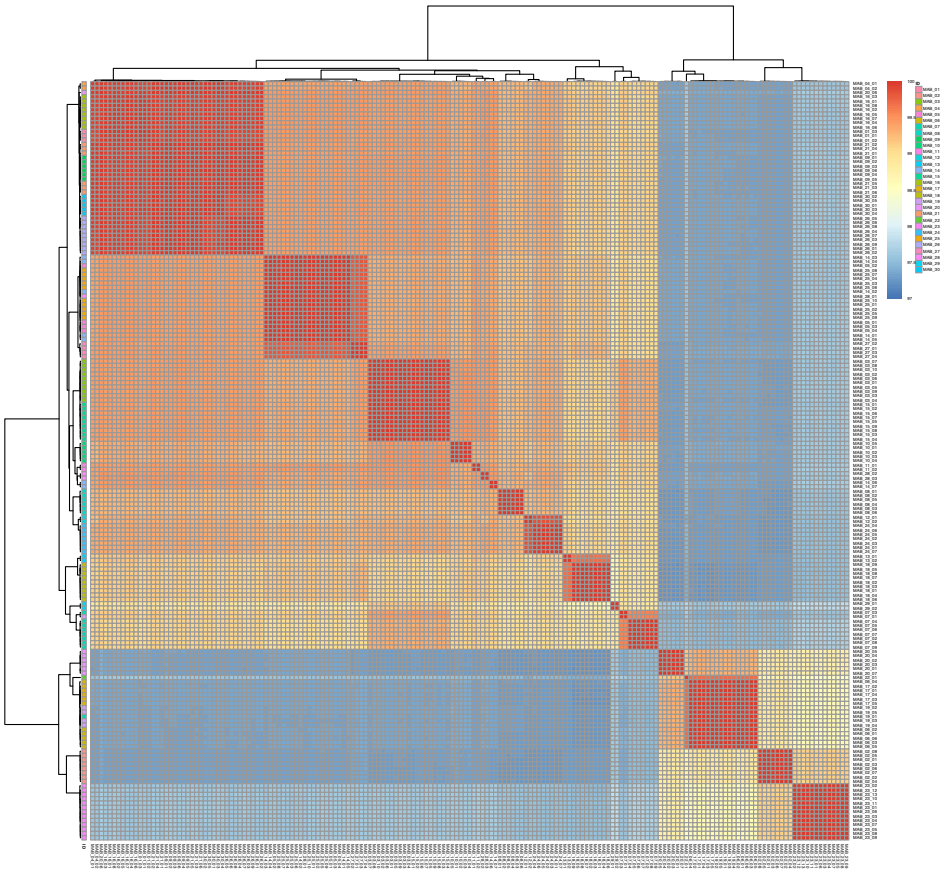

Figure 3.2 Heatmap of pairwise ANI values for all 175 isolate genomes. Each row and column signify an isolate, and color represents ANI values. General clustering of two groups corresponding to the two subspecies (larger group at upper left: subsp. *abscessus*, smaller group at lower right: subp. *massiliense*) is observed.

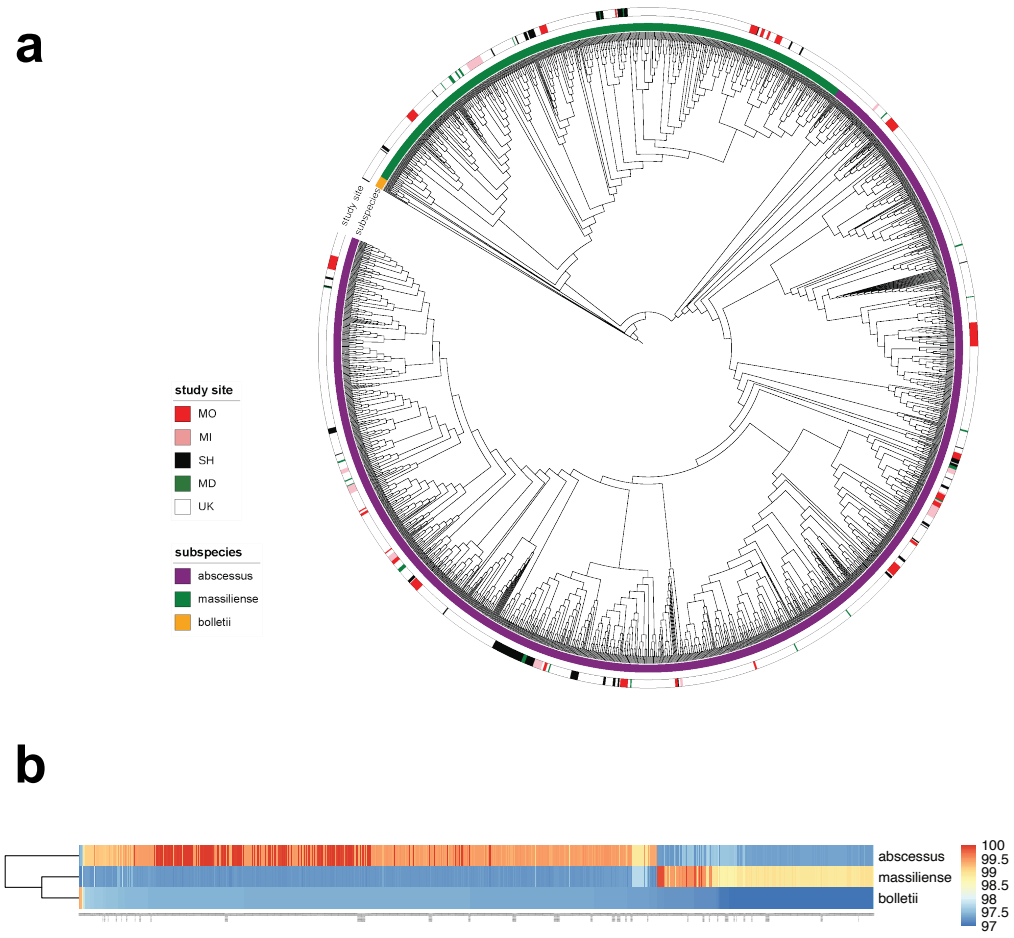

Figure 3.3 Genomic comparisons with 1455 global MABSC genomes. (a) Core genome alignment of our cohort (*n*=175) with 1455 published MABSC genomes, 1,423 core genes. Alignment conduced with roary, converted with FastTree, visualized with iTOL as cladogram, rerooted at the *M. abscessus* subsp. *bolletii* reference genome. Branch lengths do not represent distance. Outer rings denote subspecies (yellow: *bolletii*, green: *massiliense*, purple: *abscessus*) and study site. Study site does not always represent where sample was originally obtained, but rather where the genome was sequenced and reported. MO: Missouri, MI: Michigan, SH: Shanghai, MD: Maryland, UK: United Kingdom **(b)** Average Nucleotide Identity (ANI) of 1630 MABSC isolates against 3 subspecies reference genomes. The rows are each isolate genome, and the columns are each subspecies genome. Isolates at least 98.5% ANI with a given reference genome were classified as belonging to that subspecies.

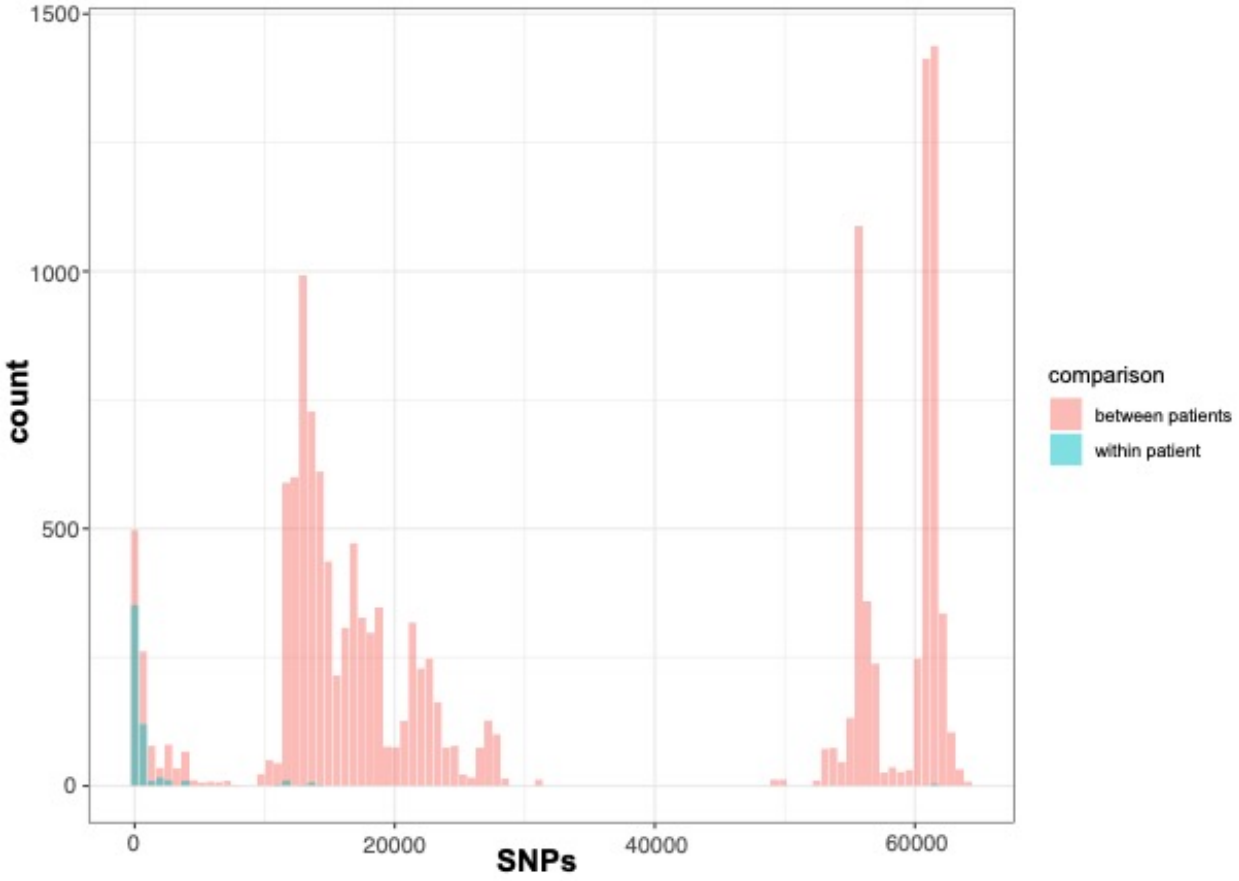

Figure 3.4 Histogram of pairwise core genome Single Nucleotide Polymorphism (SNP) distance across 175 isolate cohort. Measured from alignment of 2,693 core genes using roary and snp-sites. X-axis indicates SNP distance, while Y-axis indicates frequency. Comparisons of same-patient isolates are colored turquoise ("within patients"), and comparisons between different-patient samples are colored pink ("between patients").

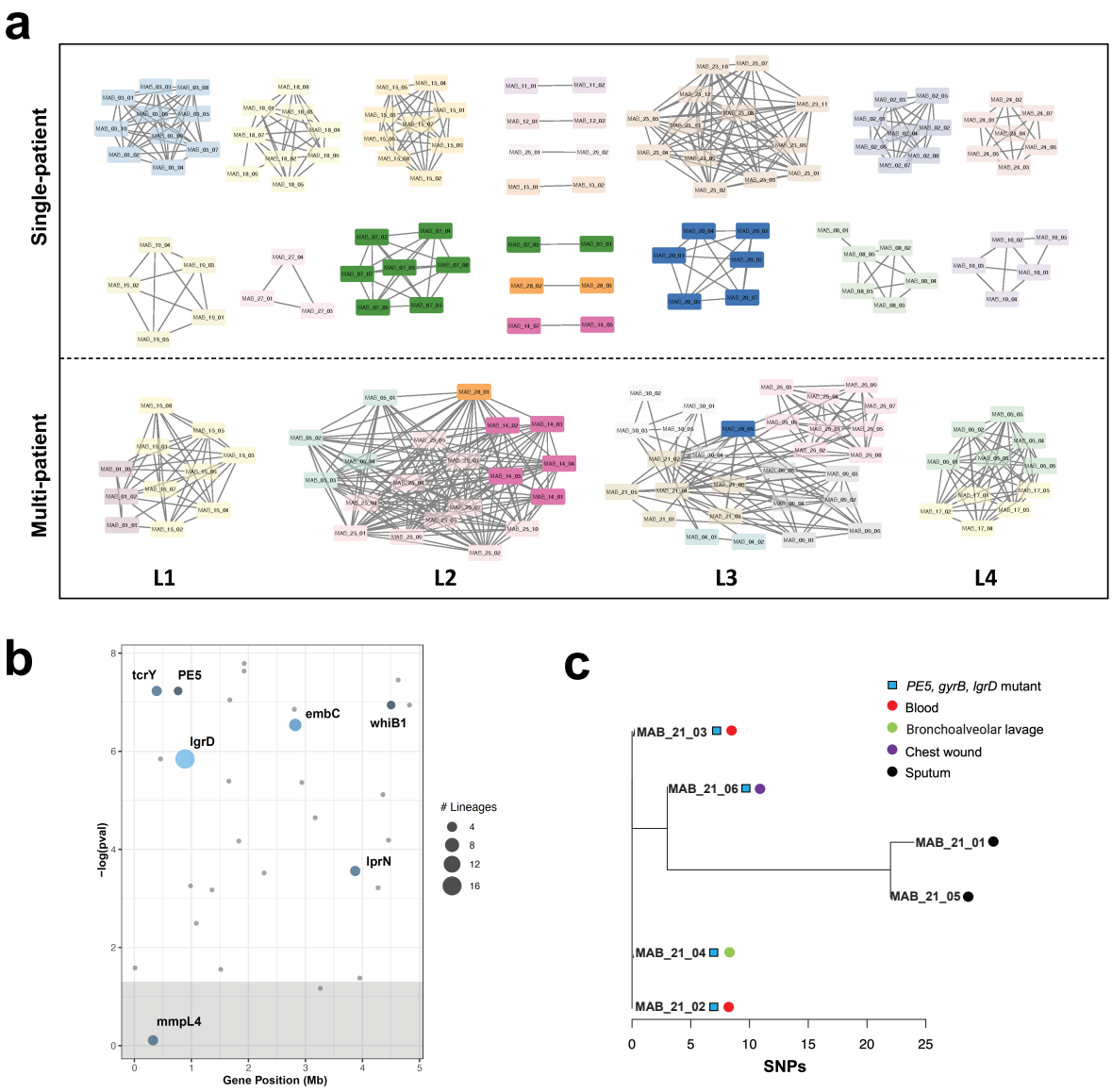

Figure 3.5 Multiple subspecies or lineages coexist within four patients. (a) Network visualization of isolate genomes at least 99.99% ANI. Nodes indicate genomes and edges indicate a pairwise ANI value of at least 99.99%. Colors indicate patient ID; nodes of the same color are isolates from the same patient. Clusters have been divided into two panels to distinguish clusters of isolates coming from a single patient from multi-patient clusters. Multi-patient clusters are labeled L1-L4. Nodes highlighted with higher opacity are isolates from patients with multiple subspecies present (MAB_20, blue) or multiple lineages (MAB_07, green; MAB_14, pink; MAB_28, orange). Two isolates without any 99.99% ANI matches are not pictured: MAB_22_01 and MAB_27_02. ANI measured with dnadiff, clusters visualized on Cytoscape. **(b)** Bubble plot displaying

results of permutation analysis. Mutations were randomly distributed across representative isolate genomes to generate a neutral (expected) distribution for parallel mutations across lineages. This distribution was then compared with the observed number of lineages with mutations in each gene. X-axis denotes position in the ATCC19977 reference genome, Y-axis denotes negative log p-value. Bubble size corresponds to number of lineages the gene was found to be mutated in. Genes mutated in at least 3 lineages are colored and named. Area with grey background indicates *P*-value < 0.05. **(c)** Unrooted tree showing whole genome SNP distances between isolates from MAB_21, which were collected within 9 days of each other. X-axis is SNP distance, tree nodes contain isolate IDs. Nodes are also annotated for site of collection and key mutations observed. SNPs annotated by aligning reads against initial isolate genome MAB_21_01.

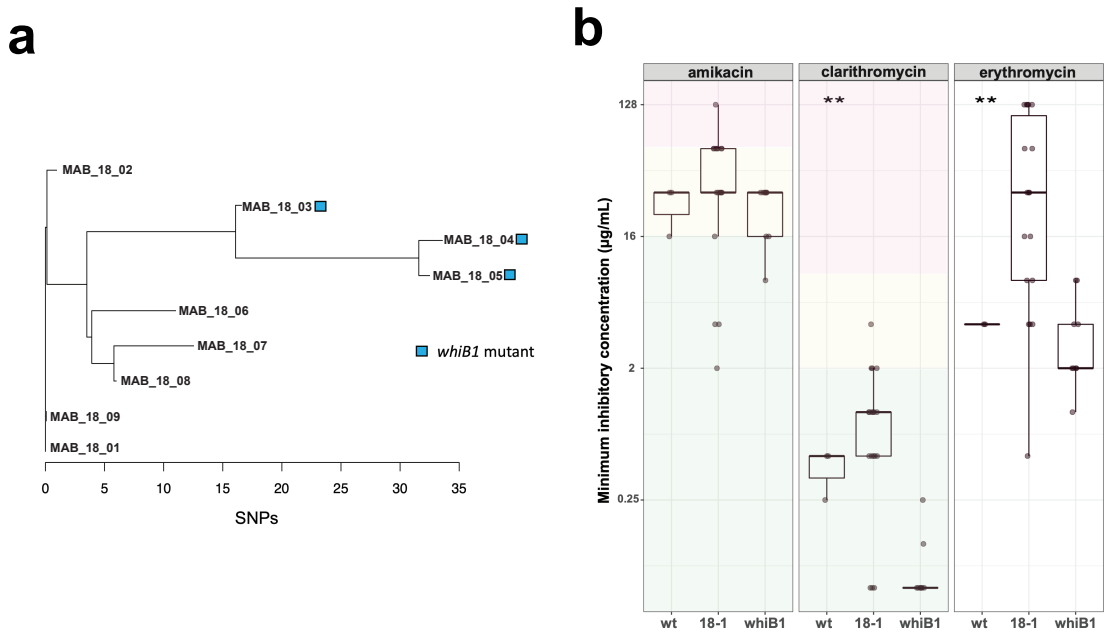

Figure 3.6 MAB_18 isolates show varied macrolide susceptibility. (a) Unrooted tree showing whole genome SNP distances between isolates from MAB_18. X-axis is SNP distance, tree nodes contain isolate IDs. Nodes are also annotated for key observed mutations. **(b)** Results of antimicrobial resistance assay. Each panel indicates a tested drug. Isolates (n=9) are grouped by observed mutation: MAB_18_03, MAB_18_04, MAB_18_05 are categorized as "whiB1" and the remaining MAB_18 isolates are "18-1". ATCC19977 was included as a control ("wt"). Significant differences among the groups were observed for clarithromycin (P=0.00105) and erythromycin (P=0.000789, Kruskal-Wallace rank sum test). Background colors in the panel represent clinical interpretation according to CLSI M24-A2 guidelines: resistant (red), intermediate (yellow) or susceptible (green). No interpretation for erythromycin is available and thus left blank.

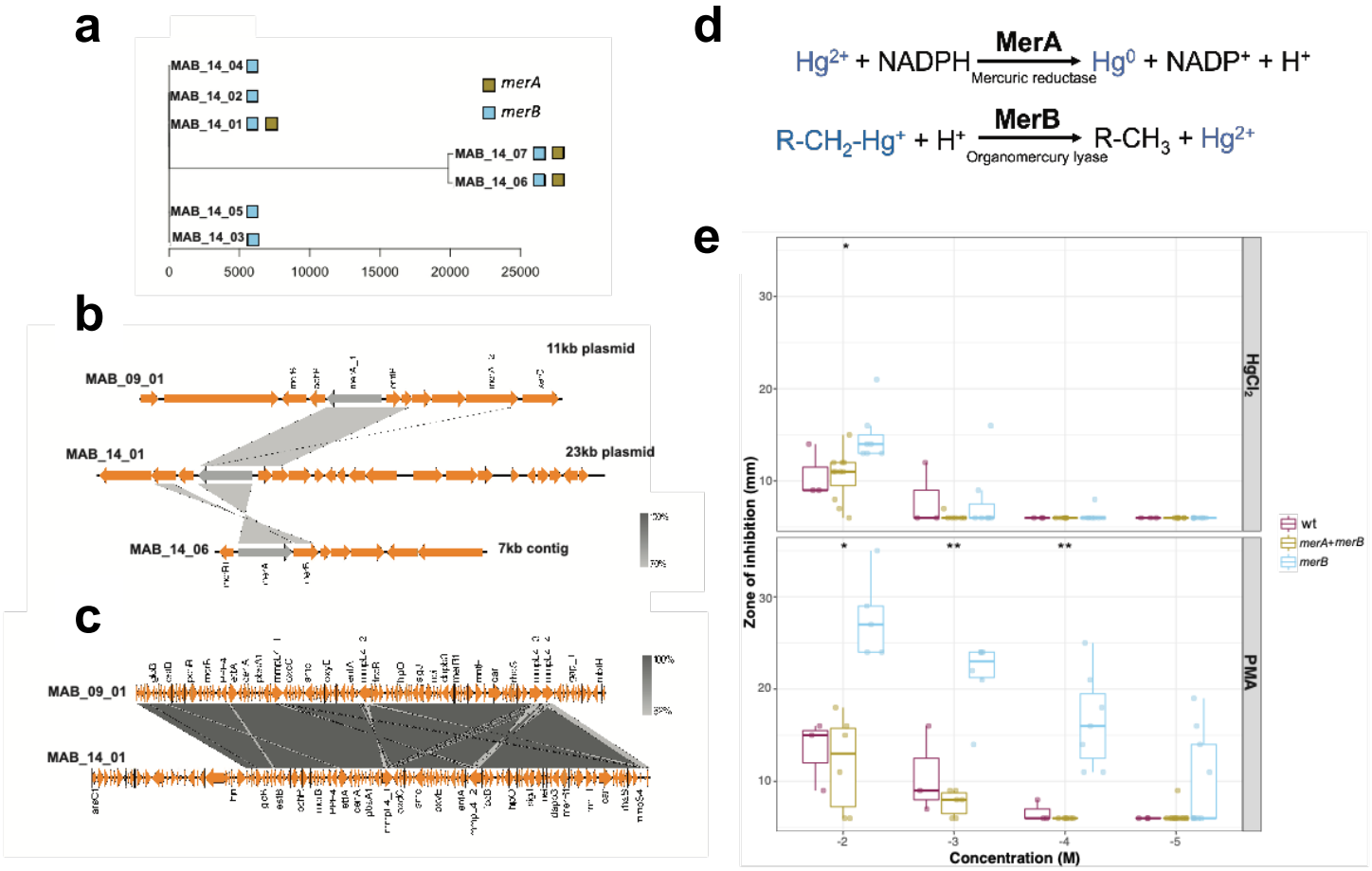

Figure 3.7 Loss of mercury resistance genes affects mercury susceptibility (a) Unrooted tree showing whole genome SNP distances between isolates from MAB_14. X-axis is SNP distance and tree nodes contain isolate IDs. Nodes are also annotated for presence of *merA* or *merB*. Isolates MAB_14_01 through MAB_14_05 belong to one lineage, while MAB_14_06 and MAB_14_07 belong to a second lineage. **(b-c)** Genomic context of *merA* and *merB* genes across isolate genomes. *merA* is colored grey, while other coding sequences are orange. Each row is a visualization of an assembled contig containing *merA* or *merB*. Grey regions between contig rows indicate regions of high percent identity according to BLASTn. Visualized using easyfig. **(d)** Diagram illustrating activity of mercuric reductase MerA and organomercury lyase MerB. MerA reduces inorganic mercury to the inert form, while MerB lyses mercury from methyl compounds. (E) Results of mercury resistance assays. Each clinical isolate was exposed to inorganic ($HgCl_2$) or organic (PMA) mercury compounds via disc diffusion assay. ATCC19977 was included as a control (purple). Isolates are grouped by genotype: *merA* and *merB* (MAB_14_01, MAB_14_07 and MAB_14_08, turquoise) or *merB* only (MAB_14_02 through MAB_14_05, orange). ATCC19977 is "wt" (purple) and contains a

23kb plasmid identical to MAB_14_01. Significant differences were observed between groups in $HgCl_2$ 10^{-2} M (P =0.011, Kruskal-Wallace rank sum test), PMA 10^{-2} M (P =0.011), PMA 10^{-3} M (P=0.0064), and PMA 10^{-4} (P=0.0016). 6mm indicates disc size and no zone of inhibition.

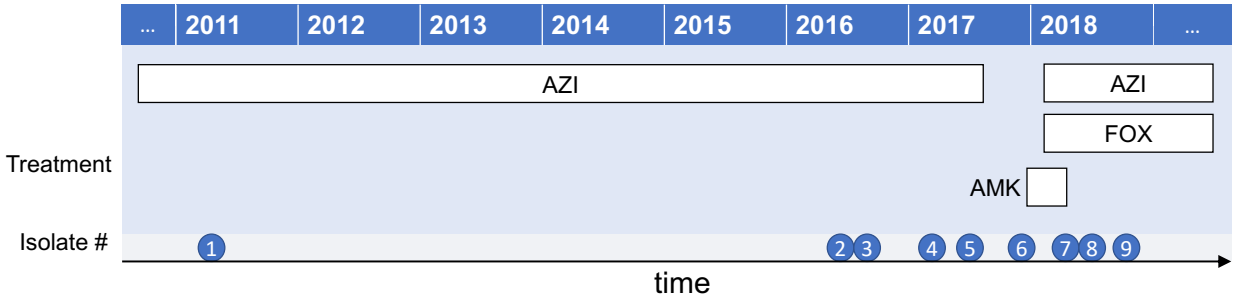

Figure 3.8 Treatment timeline for patient MAB_18. X-axis indicates time in years, and antimicrobials administered to patient are represented as white blocks. AZI: azithromycin, FOX: cefoxitin, AMK: amikacin. Isolates are represented as circles along the time axis with numbers corresponding to isolate number (ex. 1: MAB_18_01).

Table 3.1 Isolate source, year, and lineage

[1]abs=abscessus, mas=masiliense
[2]BAL= bronchoalveolar lavage, RMXSI= right maxillary sinus

Sample	Subspecies[1]	Lineage	Sample source[2]	Year isolated	MLST
MAB_01_01	abs	MAB_01	sputum	2011	5
MAB_01_02	abs	MAB_01	sputum	2011	5
MAB_01_03	abs	MAB_01	sputum	2011	5
MAB_02_01	mas	MAB_02	sputum	2006	-
MAB_02_02	mas	MAB_02	sputum	2006	-
MAB_02_03	mas	MAB_02	sputum	2006	-
MAB_02_04	mas	MAB_02	sputum	2007	-
MAB_02_05	mas	MAB_02	sputum	2007	-
MAB_02_06	mas	MAB_02	sputum	2008	-
MAB_02_07	mas	MAB_02	sputum	2011	-
MAB_02_08	mas	MAB_02	sputum	2011	-
MAB_03_01	abs	MAB_03	sputum	2007	107
MAB_03_02	abs	MAB_03	sputum	2007	107
MAB_03_03	abs	MAB_03	sputum	2007	107
MAB_03_04	abs	MAB_03	sputum	2008	107
MAB_03_05	abs	MAB_03	sputum	2008	107
MAB_03_06	abs	MAB_03	sputum	2008	107
MAB_03_07	abs	MAB_03	sputum	2011	107
MAB_03_08	abs	MAB_03	sputum	2012	107
MAB_03_09	abs	MAB_03	sputum	2012	107
MAB_03_10	abs	MAB_03	sputum	2013	107
MAB_04_01	abs	MAB_04	BAL	2017	5
MAB_04_02	abs	MAB_04	sputum	2018	5
MAB_05_01	abs	MAB_05	sputum	2011	28
MAB_05_02	abs	MAB_05	sputum	2011	28
MAB_05_03	abs	MAB_05	sputum	2011	28
MAB_05_04	abs	MAB_05	sputum	2013	28
MAB_06_01	mas	MAB_06	sputum	2015	-
MAB_06_02	mas	MAB_06	sputum	2015	-
MAB_06_03	mas	MAB_06	sputum	2016	-
MAB_06_04	mas	MAB_06	sputum	2017	-
MAB_06_05	mas	MAB_06	sputum	2017	-

MAB_06_06	mas	MAB_06	sputum	2018	-
MAB_07_01	abs	MAB_07_a	sputum	2006	-
MAB_07_02	abs	MAB_07_b	sputum	2006	275
MAB_07_03	abs	MAB_07_a	sputum	2007	-
MAB_07_04	abs	MAB_07_b	sputum	2008	275
MAB_07_05	abs	MAB_07_b	sputum	2009	275
MAB_07_06	abs	MAB_07_b	sputum	2010	275
MAB_07_07	abs	MAB_07_b	sputum	2010	275
MAB_07_08	abs	MAB_07_b	sputum	2010	275
MAB_07_09	abs	MAB_07_b	sputum	2010	275
MAB_08_01	abs	MAB_08	sputum	2014	97
MAB_08_02	abs	MAB_08	sputum	2014	97
MAB_08_03	abs	MAB_08	sputum	2015	97
MAB_08_04	abs	MAB_08	sputum	2015	97
MAB_08_05	abs	MAB_08	sputum	2016	97
MAB_08_06	abs	MAB_08	sputum	2016	97
MAB_09_01	abs	MAB_09	sputum	2007	5
MAB_09_02	abs	MAB_09	sputum	2008	5
MAB_09_03	abs	MAB_09	sputum	2008	5
MAB_09_04	abs	MAB_09	sputum	2010	5
MAB_09_05	abs	MAB_09	BAL	2013	5
MAB_09_06	abs	MAB_09	BAL	2013	5
MAB_10_01	abs	MAB_10	sputum	2006	-
MAB_10_02	abs	MAB_10	sputum	2007	-
MAB_10_03	abs	MAB_10	sputum	2007	-
MAB_10_04	abs	MAB_10	sputum	2008	-
MAB_10_05	abs	MAB_10	sputum	2009	-
MAB_11_01	abs	MAB_11	sputum	2017	9
MAB_11_02	abs	MAB_11	sputum	2018	9
MAB_12_01	abs	MAB_12	sputum	2017	94
MAB_12_02	abs	MAB_12	sputum	2017	94
MAB_13_01	abs	MAB_13	BAL	2012	34
MAB_13_02	abs	MAB_13	tracheal aspirate	2014	34
MAB_14_01	abs	MAB_14_a	sputum	2013	28
MAB_14_02	abs	MAB_14_a	sputum	2014	28
MAB_14_03	abs	MAB_14_a	sputum	2016	28
MAB_14_04	abs	MAB_14_a	sputum	2016	28
MAB_14_05	abs	MAB_14_a	sputum	2017	28

MAB_14_06	abs	MAB_14_b	sputum	2017	191
MAB_14_07	abs	MAB_14_b	sputum	2018	191
MAB_15_01	abs	MAB_15	sputum	2002	107
MAB_15_02	abs	MAB_15	sputum	2004	107
MAB_15_03	abs	MAB_15	sputum	2006	107
MAB_15_04	abs	MAB_15	sputum	2007	107
MAB_15_05	abs	MAB_15	sputum	2007	107
MAB_15_06	abs	MAB_15	sputum	2007	107
MAB_15_07	abs	MAB_15	sputum	2008	107
MAB_15_08	abs	MAB_15	sputum	2009	107
MAB_15_09	abs	MAB_15	sputum	2010	107
MAB_16_01	abs	MAB_16	sputum	2013	5
MAB_16_02	abs	MAB_16	sputum	2013	5
MAB_16_03	abs	MAB_16	sputum	2013	5
MAB_16_04	abs	MAB_16	sputum	2014	5
MAB_16_05	abs	MAB_16	sputum	2014	5
MAB_16_06	abs	MAB_16	sputum	2014	5
MAB_16_07	abs	MAB_16	sputum	2015	5
MAB_16_08	abs	MAB_16	sputum	2015	5
MAB_17_01	mas	MAB_17	sputum	2015	-
MAB_17_02	mas	MAB_17	sputum	2015	-
MAB_17_03	mas	MAB_17	sputum	2016	-
MAB_17_04	mas	MAB_17	sputum	2016	-
MAB_17_05	mas	MAB_17	sputum	2016	-
MAB_18_01	abs	MAB_18	sputum	2011	34
MAB_18_02	abs	MAB_18	sputum	2016	34
MAB_18_03	abs	MAB_18	sputum	2016	34
MAB_18_04	abs	MAB_18	sputum	2017	34
MAB_18_05	abs	MAB_18	sputum	2017	-
MAB_18_06	abs	MAB_18	sputum	2017	34
MAB_18_07	abs	MAB_18	sputum	2018	34
MAB_18_08	abs	MAB_18	sputum	2018	34
MAB_18_09	abs	MAB_18	sputum	2018	34
MAB_19_01	mas	MAB_19	sputum	2007	-
MAB_19_02	mas	MAB_19	sputum	2007	-
MAB_19_03	mas	MAB_19	sputum	2008	-
MAB_19_04	mas	MAB_19	sputum	2008	-
MAB_19_05	mas	MAB_19	sputum	2009	-

MAB_20_01	mas	MAB_20_a	sputum	2007	-
MAB_20_02	mas	MAB_20_a	sputum	2007	-
MAB_20_03	mas	MAB_20_a	sputum	2007	-
MAB_20_04	mas	MAB_20_a	sputum	2008	-
MAB_20_05	mas	MAB_20_a	sputum	2008	-
MAB_20_06	abs	MAB_20_b	sputum	2009	5
MAB_20_07	mas	MAB_20_a	sputum	2010	-
MAB_21_01	abs	MAB_21	sputum	2019	5
MAB_21_02	abs	MAB_21	blood	2019	5
MAB_21_03	abs	MAB_21	blood	2019	5
MAB_21_04	abs	MAB_21	BAL	2019	5
MAB_21_05	abs	MAB_21	sputum	2019	-
MAB_21_06	abs	MAB_21	chest wound	2019	5
MAB_22_01	mas	MAB_22	sputum	2019	-
MAB_23_01	mas	MAB_23	sputum	2014	-
MAB_23_02	mas	MAB_23	sputum	2014	-
MAB_23_03	mas	MAB_23	sputum	2015	-
MAB_23_04	mas	MAB_23	sputum	2015	-
MAB_23_05	mas	MAB_23	sputum	2016	-
MAB_23_06	mas	MAB_23	sputum	2016	-
MAB_23_07	mas	MAB_23	sputum	2017	-
MAB_23_08	mas	MAB_23	sputum	2017	-
MAB_23_09	mas	MAB_23	sputum	2018	-
MAB_23_10	mas	MAB_23	sputum	2018	-
MAB_23_11	mas	MAB_23	sputum	2018	-
MAB_23_12	mas	MAB_23	sputum	2019	-
MAB_23_13	mas	MAB_23	RMXSI	2019	-
MAB_24_01	abs	MAB_24	sputum	2014	94
MAB_24_02	abs	MAB_24	sputum	2014	94
MAB_24_03	abs	MAB_24	sputum	2015	94
MAB_24_04	abs	MAB_24	sputum	2018	94
MAB_24_05	abs	MAB_24	sputum	2019	94
MAB_24_06	abs	MAB_24	sputum	2019	94
MAB_24_07	abs	MAB_24	sputum	2020	94
MAB_25_01	abs	MAB_25	sputum	2017	28
MAB_25_02	abs	MAB_25	sputum	2017	28
MAB_25_03	abs	MAB_25	sputum	2017	28
MAB_25_04	abs	MAB_25	sputum	2017	28

MAB_25_05	abs	MAB_25	sputum	2017	28
MAB_25_06	abs	MAB_25	sputum	2018	28
MAB_25_07	abs	MAB_25	sputum	2019	28
MAB_25_08	abs	MAB_25	sputum	2019	28
MAB_25_09	abs	MAB_25	sputum	2020	28
MAB_25_10	abs	MAB_25	sputum	2020	28
MAB_26_01	abs	MAB_26	BAL	2014	5
MAB_26_02	abs	MAB_26	BAL	2014	5
MAB_26_03	abs	MAB_26	BAL	2014	5
MAB_26_04	abs	MAB_26	BAL	2014	5
MAB_26_05	abs	MAB_26	BAL	2015	5
MAB_26_06	abs	MAB_26	BAL	2015	5
MAB_26_07	abs	MAB_26	BAL	2015	5
MAB_26_08	abs	MAB_26	BAL	2015	5
MAB_26_09	abs	MAB_26	sputum	2017	5
MAB_27_01	abs	MAB_27	sputum	2014	-
MAB_27_02	abs	MAB_27	sputum	2014	-
MAB_27_03	abs	MAB_27	sputum	2015	-
MAB_27_04	abs	MAB_27	sputum	2016	-
MAB_28_01	abs	MAB_28_a	BAL	2014	28
MAB_28_02	abs	MAB_28_b	sputum	2018	2
MAB_28_03	abs	MAB_28_b	sputum	2019	2
MAB_29_01	abs	MAB_29	sputum	2017	-
MAB_29_02	abs	MAB_29	sputum	2019	-
MAB_30_01	abs	MAB_30	sputum	2017	5
MAB_30_02	abs	MAB_30	sputum	2018	5
MAB_30_03	abs	MAB_30	sputum	2018	5
MAB_30_04	abs	MAB_30	sputum	2019	5
MAB_30_05	abs	MAB_30	sputum	2019	5

Table 3.2 Permutation analysis results

gene	numstrain	SNP	indel	*P*-value	BH-adjusted *P*-value	gene product
lgrD	17	7	11	1.00E-04	0.0029	linear gramicidin synthase subunit D
embC	6	6	0	1.00E-04	0.00145	arabinosyltransferase C
whiB1	3	2	1	1.00E-04	0.000966667	Transcriptional regulator WhiB1
PE5	3	3	0	1.00E-04	0.000725	PE family immunomodulator PE5
nrdI	2	2	0	1.00E-04	0.00058	Protein NrdI
mspB	2	2	0	1.00E-04	0.000483333	Porin MspB
mspA	2	2	0	1.00E-04	0.000414286	Porin MspA
tcrY	4	4	0	0.0002	0.000725	putative sensor histidine kinase TcrY
lppW	2	0	2	0.0003	0.000966667	Putative lipoprotein LppW
folP2	2	2	0	0.0003	0.00087	Inactive dihydropteroate synthase 2
ctaB	2	2	0	0.0004	0.001054545	Protoheme IX farnesyltransferase
crp	2	2	0	0.0012	0.0029	CRP-like cAMP-activated global transcriptional regulator
phoA	2	2	0	0.0021	0.004684615	Alkaline phosphatase
comEC	2	2	0	0.0022	0.004557143	ComE operon protein 3
sdhA	2	2	0	0.0031	0.005993333	Succinate dehydrogenase flavoprotein subunit
infB	2	2	0	0.0053	0.00960625	Translation initiation factor IF-2
secA1	2	2	0	0.0089	0.015182353	Protein translocase subunit SecA 1
espR	2	2	0	0.0096	0.015466667	Nucleoid-associated protein EspR
mshD	2	2	0	0.0194	0.029610526	Mycothiol acetyltransferase
lprN	4	2	2	0.0196	0.02842	Lipoprotein LprN
eccC	2	2	0	0.029	0.040047619	ESX secretion system protein EccC
papA5	2	2	0	0.0292	0.038490909	Phthiocerol/phthiodiolone dimycocerosyl transferase
htrA	2	2	0	0.0332	0.04186087	Putative serine protease HtrA

stp	2	2	0	0.0684	0.08265	Multidrug resistance protein Stp
betI	2	2	0	0.1766	0.204856	HTH-type transcriptional regulator BetI
smc	2	2	0	0.1896	0.211476923	Chromosome partition protein Smc
yhdG	2	2	0	0.2348	0.252192593	putative amino acid permease YhdG
car	2	1	1	0.3009	0.311646429	Carboxylic acid reductase
mmpL4	4	2	2	0.8966	0.8966	Siderophore exporter MmpL4

Chapter 4

Persisting uropathogenic *Escherichia coli* lineages show signatures of niche-specific within-host adaptation mediated by mobile genetic elements

The contents of this chapter are adapted from a manuscript published in *Cell Host & Microbe*:

Thänert R*, **Choi J***, Reske KA, et al. Persisting uropathogenic Escherichia coli lineages show signatures of niche-specific within-host adaptation mediated by mobile genetic elements. *Cell Host & Microbe*. Published online May 10, 2022. doi:10.1016/j.chom.2022.04.008

* = equal contribution

4.1 Abstract

Large-scale genomic studies have identified within-host adaptation as a hallmark of bacterial infections. However, the impact of physiological, metabolic, and immunological differences between distinct niches on the pathoadaptation of opportunistic pathogens remains elusive. Here, we profile the within-host adaptation and evolutionary trajectories of 976 isolates representing 119 lineages of uropathogenic *Escherichia coli* (UPEC) sampled longitudinally from both the gastrointestinal and urinary tracts of 123 patients with urinary tract infections. We show that lineages persisting in both niches within a patient exhibit increased allelic diversity. Habitat-specific selection results in niche-specific adaptive mutations and genes putatively mediating fitness in either environment. Within-lineage inter-habitat genomic plasticity mediated by mobile genetic elements (MGEs) provides the opportunistic pathogens with a mechanism to adapt to the physiological conditions of either habitat, and lower MGE richness is associated with recurrence in gut-adapted UPEC lineages. Collectively, our results establish niche-specific adaptation as a driver of UPEC within-host evolution.

4.2 Introduction

During infection or colonization, bacterial pathogens adapt to their host by optimizing their ability to replicate, disseminate, and evade host immunity[1,2]. Under strong selection, mutations arise continuously within persisting strains but rarely sweep to fixation, resulting in lasting intraspecies allelic diversity that provides a record of the pressures encountered[3,4]. Parallel signatures in unrelated hosts can identify pathoadaptive

mutations in persisting pathogens, revealing common drivers of within-host adaptation[5]. While a wealth of microbial whole genome sequencing (WGS) data has identified common patterns of pathogen adaptation (pathoadaptation)[6–8], studies of within-host evolution have, with few exceptions[9,10], been limited to specific niches in the human body, potentially overlooking population dynamics of opportunistic pathogens occupying multiple body habitats. Accordingly, there is a limited understanding of how physiological barriers between habitats may impact pathoadaptation.

One in four women affected by a UTI will experience a recurrence (rUTI) within 6 months of initial infection[11]. Uropathogenic *Escherichia coli* (UPEC) are the most common cause of UTIs, accounting for approximately 75% of uncomplicated cases[12]. The recovery of UPEC from the gastrointestinal tract at asymptomatic time points before rUTI supports a model in which UPEC lineages can persist intestinally and re-seed the urinary tract[13–15]. Emergence of uro-adaptive mutations of the type 1 fimbrial adhesin FimH in urinary isolates that are rarely present in intestinal isolates suggests rapid adaptation to habitat-specific conditions[16–19]. In some patients, however, the absence of UPEC in the intestine and the recovery of UPEC from urine at asymptomatic timepoints (asymptomatic bacteriuria) highlight that patient-specific patterns of persistence may differentially shape UPEC pathoadaptation[13]. It is unclear how the distinct physiological, metabolic, immunologic, and microbial conditions of the gastrointestinal and urinary tract impact UPEC within-host adaptation. Evolutionary trade-offs between habitats pose the question as to which molecular mechanisms enable UPEC lineages to persist, adapt, and cause repeated episodes of UTI[20].

Here, we investigate the hypothesis that habitat-specific selection in the gastrointestinal and urinary tracts differentially shapes UPEC within-host evolution. To assess this hypothesis, we characterize colonization patterns of persisting UPEC lineages in a longitudinal, prospective cohort of UTI patients. We contrast the adaptation of lineages colonizing the gastrointestinal tract with those also recovered from the urinary tracts to identify habitat-specific adaptations of UPEC. By characterizing within-lineage mutational diversity, we identify distinct patterns of within-host adaptation between UPEC colonization types indicating that niche-adaptation shapes UPEC within-host adaptation. Finally, we identify mobile genetic elements (MGEs) as a major facilitator of within-lineage genomic plasticity associated with a pool of habitat-specific genes, putatively mediating UPEC fitness in either habitat and impacting recurrence in gut-adapted UPEC lineages.

4.3 Results

4.3.1 UPEC lineages persist in the gastrointestinal and urinary tracts

We collected 976 drug-resistant *Escherichia coli* isolates from a prospective, longitudinal cohort study of 123 patients presenting with symptomatic UTI caused by antibiotic resistant (AR) uropathogens. *E. coli* were cultured from 1,752 stool and urine specimens collected at study enrollment and subsequently at 10 asymptomatic time points over a 6-month follow-up period using a home shipment protocol. Patients that experienced a rUTI within the follow-up period were able to restart sample collection (42 patients, 34.15%).

To identify UPEC lineages persisting within patients, we characterized genomic relatedness of same-patient isolates using whole-genome sequencing (WGS) of all 976 *E. coli* isolates (average of 8.2 isolates/patient). Following methodologies implemented in similar studies[21,22], we profiled single nucleotide polymorphism (SNP) distances based on patient-specific core-genomes to differentiate isolates belonging to the same *E. coli* lineage as the causative agent of the index UTI from isolates representing distinct subspecies clusters. We observed that within-patient SNP distances followed a multimodal distribution (Figure 4.1a), with a notable paucity of within-patient pairwise isolate SNP distances between 500 and 10,000 SNPs. To assess plausibility of 500 SNPs as the upper limit of a UPEC lineage definition for this study, we estimated the average duration since last common ancestor (LCA) for each lineage. For each persistent lineage, we generated whole genome SNP trees based on lineage-specific reference assemblies and calculated the median branch length. We then divided this value by a previously reported estimated rate of *E. coli* base substitution (8.9×10^{-11} bp/generation)[23]. Importantly because our estimate is based on within-gut *E. coli* generation times, values for urinary persisters are likely less accurate. We estimated an average of ~0.33 (0-5.39, Figure 4.1b) years since the LCA, consistent with the reported history of recurrent UTIs in our patient cohort. Whole genome pairwise ANI values calculated between same-patient isolates further showed that isolates typed to the same lineage based on the 500 core genome SNPs cutoff exhibited high pairwise ANI values (99.991% (0.0127) - median (IQR)), while isolates from the same patient typed into distinct lineages and from distinct

patients displayed lower, variable ANI values (97.288% (1.531), 97.268% (1.588), Figure 4.1c-d).

We applied the 500 core genome SNPs cutoff to all isolates cultured from the same patient and identified a total of 187 distinct subspecies clusters of *E. coli* (hereafter referred to as 'lineages' - Figure 4.1). 702 isolates recovered at asymptomatic time points belonged to 119 lineages that were isolated as the causative agent of a UTI (diagnostic urinary isolate: DxU) and were defined as UPEC for the purpose of this study. The majority of these lineages belonged to the pandemic ExPEC sequence type complexes (STc) 131 (36.97%, Serotypes O25:H4 and O16:H5), predominately ST131-*fimH*30, and STc14 (21.85%, Serotype O75:H5), predominately ST1193 (Table 4.1, Figure 4.2).

We characterized asymptomatic persistence of UPEC lineages based on longitudinal recovery of same-lineage *E.coli* from patient-matched urine and stool specimens, using standard-of-care clinical microbiology culturing methods (Fig 4.3a, Methods). We classified three distinct patterns of UPEC lineage persistence (see Methods): (1) gastrointestinal persistence ('Gut colonizer', 51 lineages, 46.4%), (2) persistence in both habitats ('Dual colonizer', 32 lineages, 29.1%), or (3) persistence in the urinary tract ('Urinary colonizer', 4 lineages, 3.6%, Fig 4.3a). Isolates belonging to these categories were used in downstream analysis to investigate UPEC within-host evolution. In 23 patients (20.9%) we did not find evidence for UPEC persistence in either the urinary or the gastrointestinal tract. While sequence type distribution did not differ between persistence types (Fig 4.3b), STs of non-persisting lineages differed significantly from that of persisters (Fig 4.3c, Fisher's exact test $P<0.001$), with ST131 and ST1193

underrepresented among non-persisting lineages (Fisher's exact test $P<0.001$). Interestingly, dual colonizers were associated with the majority of rUTI events attributable to a specific lineage during the 6-month follow-up period (57.9% (11/19 lineages), 36.8% (7/19) gut colonizer, 5.3% (1/19) urinary colonizer). Collectively, these observations suggest that colonization of the gut (Gut colonizer) or both environments (Dual colonizer) describe the majority of persistent UPEC.

4.3.2 Urinary persistence is associated with increased allelic diversity of UPEC lineages

To assess the impact of environmental selection on UPEC within-host evolution, we profiled the within-host adaptation of UPEC lineages in their persistence habitats (*i.e.*, gut colonizers in the gut, dual colonizers in gut and urinary tract, and urinary colonizers in the urinary tract). We identified all within-lineage SNPs by aligning sequenced reads against lineage-specific pseudo-assemblies, as previously described[13,24].

By inferring the ancestral sequence through maximum parsimony, we found that urinary persistence is associated with significantly increased distance to the most recent common ancestor (dMRCA) compared to gut colonizing lineages (Figure 4.4a, $n=87$ lineages, Kruskal-Wallis $P=1.38e^{-05}$, Dunn post-hoc test gut vs dual colonizer $P=2.39e^{-05}$, gut vs urinary colonizer $P=3.32e^{-02}$). These observations are consistent with two potential explanations; First, urinary persistence may enable UPEC lineages to persist within a host for longer durations. Alternatively, considering that *E. coli* are native to the gut, disparate selective pressure in the urinary tract could result in habitat-specific fitness maxima

distinct from those of the gastrointestinal tract and extend the spectrum of positively selected mutations, diversifying the allelic repertoire of persisting UPEC lineages.

4.3.3 UPEC niche-specific adaptation shapes within-host adaptation

To test the hypothesis that urinary persistence results in trajectories of within-host adaptation distinct from those observed in the gut, we annotated within-lineage allelic diversity (SNPs, insertions, deletions) at the gene level. We implemented permutation tests, randomly distributing the number of observed mutations over each lineage's pseudo-assembly to generate a null distribution. We then compared observed against expected frequencies to identify genes with signatures of non-random evolution across lineages. Permutation tests were conducted independently for colonization types to characterize the effect of distinct persistence patterns.

Our analysis identified 253 genes with mutational signatures indicating non-random selection (n=87 lineages, Permutation test, confidence interval 95%). To validate that positive selection drives mutations in this gene set, we calculated per gene dN/dS ratios, a canonical metric for selection. We found a robust enrichment of elevated dN/dS values for both genes mutated in a single lineage (Figure 4.4b, m=1, median 11.57±11.41 median absolute deviation (MAD)) or in parallel across multiple lineages (m≥2, 11.52±10.78) compared to genes non-significant by permutation test (median 0.97±0.98). Consistent with this observation, the overall dN/dS value for all genes significant by permutation test and mutated in parallel across lineages, 1.34 (0.96-2.02, 95% confidence interval by binomial sampling), indicated that adaptation drives mutation in these genes.

In contrast, genes carrying mutations but non-significant by permutation test were under purifying selection (dN/dS 0.32, 0.30-0.35), consistent with previous literature[24].

Mutations of a single gene (*wbbL*) was observed in all colonization types, while 12 genes were shared between at least two groups (Table 4.2). Virulence- and drug-associated genes were mutated in parallel frequently across colonization types (Figure 4.4c), including capsule-related genes *neuC* (dN/dS 7.3) and *mprA* (dN/dS 17.5), as well as *wbbL* (dN/dS 59.4), coding a rhamnosyl transferase critical for O-antigen synthesis. As both capsule and O-antigen directly affect UPEC fitness *in vivo*[25], these mutations may also affect UPEC persistence. Further, genes implicated in antibiotic resistance, including *ompC* (dN/dS 17.8), *acrR* (dN/dS 5.8), *nfsA* (dN/dS 17.8), and *nfsB* (dN/dS 10.9)[26–28], were found to be under positive selection across lineages. Interestingly, mutations of the biofilm suppressing antiterminator RfaH encoding gene (dN/dS 33.5) were exclusively found in lineages persisting within the urinary tract. Biofilms are critical UPEC colonization factors, enabling adhesion to abiotic (catheter) and biotic (urinary tract) surfaces[29].

To assess functional adaptation of UPEC during persistence comprehensively, we performed Gene Ontology term overrepresentation analysis (GOOA) in the pool of all genes mutated within-lineages that exhibited a signature of non-random selection. Strikingly, functional categories under selection differed between colonization types, with only a small set of core-functions (sialic acid transport, membrane assembly, antibiotic resistance, negative regulation of transcription) found to be under selection in multiple colonization types (Figure 4.4d). Distinct transport capabilities, response to

environmental stressors, metabolic processes, and regulatory functions were selected in gut-restricted and dual colonizers (Figure 4.4d), indicating that distinct persistence patterns differentially shape within-host adaptation of persisting UPEC lineages. Functions found to be under selection in dual colonizers, including iron ion transport, response to pH, response to nitric oxide, ornithine metabolism, or fumarate metabolism (Figure 4.4d), have been linked to urinary fitness of UPEC and likely direct adaptations towards the habitat-specific conditions of the urinary tract[30,31]. Collectively, these results support the idea that niche-specific selection shapes the evolutionary trajectories of persisting UPEC, altering the landscape of positively selected functionalities for multi-habitat lineages.

4.3.4 Within-host adaptation of UPEC impacts resistance phenotypes

We observed that 79.4% of the within-lineage allelic diversity in genes mutated in parallel among dual colonizing lineages was structured by habitat, with mutations only occurring in a single habitat within a lineage (Figure 4.5a). Similarly, when including 71 additional urinary isolates from the 51 gut colonizing lineages and implementing our permutation test to identify genes under positive selection (Table 4.2), we found that an even larger fraction of mutations in genes with parallel signature across lineages was only found in isolates cultured from one sample type (93.5%, Fisher's exact test, P=0.001). As urinary colonizers had no representative gut isolates, they were not included in this analysis. We reasoned that this phenomenon could result from two potential processes: (1) a consequence of genetic bottlenecks upon habitat transition, or (2) habitat-specific

selection resulting in divergent subpopulations within the same lineage in the gastrointestinal and urinary tract.

To test whether niche-specific adaptation may in fact play a role in shaping allelic breakdown along habitat lines in persisting UPEC lineages, we focused on a subset of mutations with a tractable phenotypic impact. We had previously observed strong selection for mutations in antibiotic-resistance associated genes during persistence (Figure 4.4d) and reasoned that niche-specific adaptation would result in niche-dependent resistance phenotypes. Therefore, we identified mutations in antibiotic resistance genes and profiled isolate resistance phenotypes for both dual and gut colonizing lineages. We found that the nonsynonymous *ompC* R191C mutation in dual colonizing lineage WU-041_1 was exclusively found in urinary isolates and coincided with the gain of ampicillin/sulbactam (Figure 4.5b). Importantly, we found that non-synonymous mutations of *ompC*, including another instance of R191C in lineage PN-004_1, were restricted to urinary isolates. Similarly, we found *nfsA* Q191* mutation in gut colonizing lineage WU-046_2 exclusively in isolates cultured from urine specimens during symptomatic disease and immediately preceding recurrence (Figure 4.5c), associated with the gain of phenotypic nitrofurantoin resistance. Moreover, identified resistance-conferring mutations of *nfsA*, including another premature stop codon in lineage PN-004_1 (*nfsA* W237*), were restricted to urinary isolates. Together, these findings indicate niche-dependent fitness benefits of mutations in these two genes and a role of niche-specific adaptation in shaping within-host adaptation of persisting UPEC lineages.

We further reasoned that if these observed mutations provide UPEC with direct fitness benefits, they may also be found in UPEC genomes sequenced in different studies. To test this, we downloaded a set of 703 UPEC genomes previously curated from multiple studies[32] and profiled allelic identify of *ompC* and *nfsA* at all positions observed to be variable in this study. We found that for *ompC* and *nfsA* in 2/4 cases and 1/4 cases, respectively, the exact mutations identified in our study were observed in published UPEC genomes (Figure 4.6). This suggests that similar selective pressures to the ones characterized in this study are shaping adaptation of *ompC* and *nfsA* in the larger UPEC population.

4.3.5 Genomic plasticity facilitates UPEC niche adaptation

Differential abundance of genes within an otherwise clonal population, termed genomic plasticity, can facilitate rapid adaptation of bacterial pathogens to new environments[33–35]. The distinct physiological conditions of the gastrointestinal and urinary tracts are likely to require disparate metabolic and colonization factors. We therefore hypothesized that genomic plasticity may enable persisting UPEC lineages to maintain fitness in both the gastrointestinal and urinary environment.

Persisting gut populations of gut colonizers exhibited more homogenous gene profiles than dual colonizers (Figure 4.7a, n=87 lineages, Kruskal-Wallis test P=0.009, Dunn post-hoc test P=0.012), indicating that habitat diversification is associated with a larger pool of flexible genes. We hypothesized that this difference may be caused by greater inter-habitat heterogeneity in persisting dual colonizers not observed in lineages

persisting in the gut. To test this hypothesis, we analyzed inter-habitat similarity of same-lineage isolate gene profiles, including all 71 urinary isolates from the 51 gut colonizing lineages. We found that isolates collected from the same sample type were significantly more likely to carry similar genes, while colonization types did not differ significantly (Figure 4.7b, n=87 lineages, Two-way ANOVA, habitat P=5.94e^{-4}, colonization type P>0.05), suggesting that genomic plasticity contributes to niche adaptation of all persisting UPEC lineages.

1,553 genes were restricted to either urinary or stool isolates in the 83 UPEC gut and dual colonizing lineages and therefore may play a role in habitat adaptation (Figure 4.7c). Interestingly, three plasmid-associated genes, *psiA*, *yggR*, and *stbB*, were found to be restricted to gut isolates in 5 independent lineages. To comprehensively profile functional selection on the variable genetic portion of each lineage in either habitat we performed GOOA on the pool of habitat-specific genes. We identified nitrogen compound and iron uptake mechanisms as key factors for urinary adaptation in both dual and gut colonizing lineages (Figure 4.7d, Figure 4.8a, Fisher's exact test GO:0071705 P=0.018 - dual - and P=0.002 - gut, GO:0055072 P=1.81e^{-4} and P=2.51e^{-7}, GO:0044718 P=0.024 and P=0.018). Specifically, systems facilitating the uptake of ferric-citrate complexes that are abundant in urine were found to be habitat-associated in gut as well as dual colonizers (Fig 4d)[36].

Few functionalities were overrepresented in stool isolates of dual colonizing lineages (Figure 4.7E). Conversely, the gut-specific gene pool of gut colonizers exhibited enrichment of multiple functionalities implicated in *E. coli* gut colonization and virulence,

including antibiotic resistance, fumarate transport, type IV secretion, and pilus assembly[37–39]. Notably, GO terms associated with plasmid maintenance genes were found to be enriched in intestinal isolates of gut colonizing lineages, commonly coinciding with presence/absence of virulence and resistance genes (Figure 4.8a-d, Fisher's exact test GO:0030541 P=0.044, GO:0006276 P=1.77e^{-3}). We therefore hypothesized that MGEs may facilitate niche adaptation in persisting UPEC lineages.

4.3.6 Heterogenous MGE carriage facilitates habitat-associated genomic plasticity

To evaluate the role of MGEs in the genomic plasticity of persisting UPEC lineages, we comprehensively identified regions of differential coverage in isolates of the same lineage as previously described[24]. These regions are candidate MGEs differentially abundant in isolates of the same lineage. We annotated the list of putative MGEs (Figure 4.9a), combining *in silico* detection of plasmidic contigs and database-driven annotation of *de novo* identified MGEs as previously described (see Methods, Figure 4.10)[13,40]. 57.1% (887/1553 genes) of the habitat-specific gene pool mapped back to putative MGEs. As expected, we found antibiotic resistance genes (ARGs), proteolysis, and conjugation mechanisms associated with plasmidic MGEs (Figure 4.9b). Pathofunctions that were implicated as habitat-specific in our previous analysis, including iron import systems, type II and type IV secretion systems, and cell adhesion genes, were found to be enriched within MGE subcategories.

To profile potential sharing of UPEC MGEs with other species we mapped all MGE contigs to the NCBI nucleotide database. We found that plasmidic MGEs had the

broadest putative host range (Figure 4.11a). However, plasmidic MGEs exclusively identified in urinary isolates exhibited a trend towards a narrower host range compared to those found in the gut (Figure 4.11a, ANOVA P=0.053, Tukey post-hoc test vs gut-exclusive P=0.053, vs dual-habitat P=0.057). Moreover, these MGEs were significantly less likely to be mapped to common gut residents, including *Salmonella enterica*, *Citrobacter freundii*, or *Enterobacter cloacae* (Figure 4.11b, Fisher's exact test, FDR corrected P<0.05), indicating that gut-associated plasmidic MGEs are more likely be shared with other gut residents.

Contrary to the high intra-habitat dissimilarity of lineage MGE profiles in urinary colonizers (Figure 4.9c), we observed homogenous within-habitat MGE carriage in dual and gut colonizing lineages. In gut colonizing lineages, heterogeneity of MGE carriage was significantly elevated across habitats compared to within-habitat, as well as significantly larger compared to dual colonizers (Figure 4.9c, n=87 lineages, Two-way ANOVA $P \leq 1.57e^{-05}$, Tukey post-hoc P<0.001 and P=0.014, respectively). These results suggest that multi-habitat selection in dual colonizers may stabilize the MGE pool across habitat boundaries. Urinary isolates' MGE pools were significantly smaller compared to intestinal isolates (Figure 4.9d, n=87 lineages, Two-way ANOVA P=0.042). Moreover, we found that habitat-specific genes from metabolic, antibiotic resistance, and virulence-associated functional categories were mapped to MGEs exclusively present in urinary or stool isolates (Figure 4.9e-f). These observations suggest that mobilization of key functions associated with adaptation to either habitat, such as iron acquisition or nitrogen

compound uptake in the urinary tract (Figure 4.7d), may play a key role in UPEC niche adaptation.

Interestingly, the association of MGEs with ARGs resulted in a pool of 'hidden' ARGs not observed in the DxU isolate but present in other isolates of the same lineage (Figure 4.12). Isolates harboring 'hidden' ARGs frequently showed concordant variation in their replicon profile compared to the DxU isolate (66/78 cases, 84.6%), corroborating differential resistance plasmid carriage as a potential driver of within-lineage plasticity of ARGs.

4.3.7 Decreased MGE richness is associated with rUTI in gut-colonizing UPEC lineages

Based on our observation of decreased urinary richness of MGEs, we hypothesized that MGE richness may hamper urinary fitness of gut-adapted lineages of UPEC resulting in an inverse relationship between MGE richness and the likelihood of a lineage causing a rUTI during our follow-up period. In fact, we found that gut colonizer lineages causing rUTI exhibited significantly lower average MGE richness per isolate compared to their non-rUTI counterparts (Figure 4.13a, n=43 lineages, Welch's t-test, FDR corrected P=0.001). Notably, no such relationship was observed for dual colonizers (n=26 lineages, Welch's t-test, FDR corrected P=0.884).

Despite considerable variability in the functional composition of their mobilized gene pool, no functional category was significantly enriched after correcting for multiple hypothesis testing in either rUTI or non-rUTI lineages (Figure 4.14a, n=69 lineages, Fisher's exact test, all FDR corrected P>0.05). However, we observed a trend towards

lower mobilized ARG richness in rUTI lineages compared to non-rUTI lineages (Figure 4.14b-c n=69 lineages, Wilcoxon rank-sum test P=0.055). We found no difference between the mobilized ARG richness of UPEC persistence types (Figure 4.14d-e, n=87 lineages, Kruskal-Wallis P=0.231).

To identify mobilized functions negatively impacting urinary fitness of gut-adapted UPEC lineages, we characterized the habitat association of each putative MGE for all gut colonizer lineages. We identified a large gut-specific MGE pool (238/457, 52.08%) absent from any urinary isolate. GOOA of genes present on these gut-specific MGEs identified 9 out of 94 GO categories significantly depleted in urinary isolates (Figure 4.13b, Fisher's exact test, FDR-corrected P-value<0.05), including DNA-related, lipid biosynthetic, and type-IV secretion system processes. Interestingly, while some gut-specific GO categories were absent from the MGE pool of rUTI-causing gut colonizers (*e.g.*, antibiotic biosynthesis, tryptophan biosynthesis), these GO terms were in general not underrepresented in their MGE pool (Figure 4.13b).

4.4 Discussion

Invasion and colonization of the urinary from the gastrointestinal tract is the first step in the infectious cascade of the majority of UTIs caused by UPEC[41]. While the affordable implementation of WGS in longitudinal cohort studies has uncovered adaptive patterns of various species to specific host environments[6,7], the within-host pathoadaptation of multi-habitat pathogens remains understudied. Here, we characterize the pathoadaptation of UPEC, one of the most common bacterial pathogens recovered from

multiple body sites. Viewing UPEC within-host evolution in the context of their respective niche is key to understanding the origins of urovirulence in inherently intestinal *E. coli*, particularly in light of the lack of a defining genomic signature of UPEC[42].

Our results support three distinct models of UPEC persistence: exclusive persistence in the gastrointestinal tract (gut colonizer), persistence in both the gastrointestinal and urinary tracts (dual colonizer), and exclusive persistence in the urinary tract (urine colonizer). We find that these distinct patterns of persistence differentially shape UPEC within-host pathoadapation. While development of antibiotic resistance is strongly selected for in all persisting UPEC lineages, as previously reported for other pathogens[8,43,44], we find that distinct functions are under selection in gut and dual colonizers. Specifically, signatures of positive selection in distinct transport functions indicate that niche specific adaptation directly impacts evolutionary trajectories of pathoadaptive traits[45]. Further adaptation to multiple habitats diversifies allelic profiles of persisting UPEC lineages. Intriguingly, potential inter-habitat transfer resulting in the influx of uroadaptive mutations back into gut populations may consequentially lower the fitness boundaries for urinary re-colonization by intrinsically gut-adapted *E. coli*. Experimental evidence has shown that virulence factors critical for uro-colonization are similarly beneficial in the intestinal reservoir[14,39,46], mitigating theoretical evolutionary trade-offs. These observations suggest that urovirulence may be a direct consequence of the generalist properties of the *E. coli* virulence repertoire[47], which is, as we show, fine-tuned by habitat-specific adaptations in the urinary tract.

Our observations support the hypothesis that persistent pathogen colonization requires within-lineage genotypic heterogeneity originating from both *in situ* adaptation as well as genomic plasticity[33]. The prevalence of habitat-restricted mutations and genomic plasticity between urine and stool isolates provides strong evidence that niche-specific adaptation dictates within-host evolution during UPEC persistence. We find that habitat-specific genes are associated with functions that increase *E. coli* fitness in the intestinal or urinary habitat, such as piliation, iron acquisition, nitrogen import, or anaerobic respiration[36–39]. Persisting pathogen lineages require mechanisms that facilitate rapid rearrangements of large genomic regions to adapt to the distinct selective regimes of each habitat. Requirements for rapid genomic plasticity have been described for other pathogens, specifically during early stages of habitat colonization[35,48]. Our results support the hypothesis that those genomic rearrangements are in part facilitated by MGEs[49]. Intriguingly, we observed that functions related to DNA repair were depleted in the MGE gene pool of urinary isolates from gut-adapted UPEC. This observation is consistent with the concept that stress-induced mutagenesis enables maladapted bacteria to evolve rapidly to their environment and may therefore be beneficial following urinary inoculation with gut-adapted lineage of UPEC[50]. Heterogenous MGE carriage provides opportunistic pathogens with a unique mechanism to maintain fitness in multiple habitats. *In vitro* experiments have shown that complex environments result in discontinuous plasmid distribution in clonal populations, potentially resulting in fitness benefits in changing environments[51–53]. Our results support the hypothesis that MGE-mediated plasticity in bacterial populations is a key mechanism for habitat adaptation

and may directly impact bacterial fitness upon habitat transition. Our data further suggest that a pool of gut-specific MGEs shared with other gut resident species may be lost in the urinary environment. Moreover, we find that gut colonizing lineages causing rUTI during our follow-up period have significantly lower MGE richness compared to their non-rUTI counterparts, suggesting an inverse relationship between MGE richness and likelihood of rUTI in gut-adapted lineages of UPEC. Consistent with predictions from *in vitro* work[54], the absence of a similar trend in dual colonizers suggests that multi-habitat colonization stabilizes plasmid carriage under spatially heterogenous selection, potentially via mechanisms like compensatory mutations[55,56].

However, important questions remain to be investigated. This study could not address the topic of directionality and inter-habitat transfer, the frequency of which may impact adaptative trajectories of persisting UPEC lineages. Moreover, given the apparent importance of genomic plasticity for UPEC fitness, localization of functions on either the chromosome or MGEs may determine the uropathogenic potential of intestinal *E. coli* lineages. The mosaic structure of plasmids poses the question which functions determine plasmid spread, evolution and persistence in UPEC lineages. While our study represents one of the largest genomic databases of UPEC to date, a number of patients were lost due to drop-out limiting the number of available isolates from follow-up episodes, specifically diagnostic isolates from outpatient settings. Similarly, our study lacked a representative number of lineages persisting exclusively in the urinary tract, that are potentially uniquely adapted to the urinary environment. Large multi-episode sampling

efforts from patients at risk for rUTI are required to support rarity of this persistence type and the novel genomic predictions of our study.

This study, harnessing an expansive, longitudinal patient cohort sampled at multiple habitats, provides a framework for future investigations, studying the role of both *in vivo* mutations and genomic plasticity in the within-host adaptation of bacterial pathogens across niches. Similar investigations in other species may reveal further mechanisms of colonization and aid targeted de-colonization of persisting human pathogens.

4.5 Methods

4.5.1 Patient cohort

Subjects for this prospective, multi-center cohort study were recruited from patients with positive clinically indicated urine cultures at Barnes-Jewish Hospital/Washington University in St. Louis (WU), St. Louis, Missouri, Duke University Hospital (DK), Durham, North Carolina, the Hospital of the University of Pennsylvania (PN), Philadelphia, Pennsylvania and Rush University Medical Center (RH), Chicago, Illinois. This study was approved by the Washington University Human Research Protection Office as the single IRB; local IRB approval was obtained as necessary. Patients with a symptomatic UTI diagnosed and treated by a physician and a urine culture that yielded *E. coli* with one of the following resistances were included in the current analysis: (1) resistance to ciprofloxacin or levofloxacin, (2) resistance to any third generation cephalosporin, (3) resistance to ertapenem and susceptible to meropenem, imipenem,

and/or doripenem, (4) resistance to >2 of the following antimicrobial classes: carbapenems, aminoglycosides, fluoroquinolones, fourth generation cephalosporins, piperacillin/tazobactam, or (5) identification of any of the following resistance mechanisms: ESBL, CRE, KPC, NDM-1, OXA-48, IMP, IMP-1, or VIM.

Patients were excluded if they were younger than 18 years, if more than one organism was detected by the clinical laboratory at or above the clinical significance threshold, had any chronic indwelling urinary device, or any medical or surgical condition leading to intestinal or urinary system disease or anatomic alteration. Written, informed consent was obtained from all patients. Patients age averaged 56.26 years (range: 18-94, median: 59). 93.5% of patients were female, and 6.50% of patients male. 58.54% of patients self-reported their race as White, and 37.40% as Black. 4.07% of patients reported their ethnicity as Hispanic. Pearson's chi-square tests indicated no significant association of age, sex, or race with UTI recurrence or UPEC colonization.

123 of 127 enrolled patients had at least one biological specimen yielding *E. coli* and were included in the current study. This total includes data from 12 patients enrolled at WU reported in a pilot study[13]. In total, 41 patients were enrolled at WU, 22 at DK, 12 at RH and 48 at PN.

4.5.2 Sample collection and processing

Enrolled subjects submitted stool and urine specimens to the study team at eleven sampling points over a 6-month follow-up period; enrollment (sampling point 01); the end of UTI antimicrobial treatment (02); days 3 (03), 7 (04), 14 (05), 30 (06), 60 (07), 90 (08),

120 (09), 150 (10), and 180 (11) post-treatment. If patients experienced rUTI during the 6-month follow-up period, they were invited to continue to participate with a new follow-up period. Visual schematic of the study design was created with BioRender.com. Samples were kept on ice immediately after production and during transport by courier. Upon arrival to the lab, samples were immediately cultured or prepared for long-term storage and frozen at -80 ^{o}C.

Stool and urine samples collected at sampling points 01, 02, 04, 06, and 11 were selectively cultured to assess asymptomatic uropathogen persistence. For stool culturing, ~1 g of stool sample was supplemented with an equal amount of PBS (w/v) and vortexed to homogenize the samples. Ten, 10-fold serial dilutions of the homogenate were prepared in PBS and 10µl of the first 10 dilutions were streaked on selective agar using a 10 µL calibrated loop. For urine culture, urines were directly plated onto selective agar using a 10 µL calibrated loop using a cross-streak pattern. After 20-30 hours of incubation, agar plates were examined for growth of the putative pathogen. Selective agars were selected to be specific to each patient's identified UPEC. MacConkey agar (MAC) supplemented with ciprofloxacin was used for ciprofloxacin-resistant *E. coli*, while ESBL *E. coli* was cultured on Hardy Diagnostic's ESBL agar and MAC agar supplemented with cefotaxime. A single, representative colony of each distinct colony morphology present on a given culture plate was selected for further processing and sequenced-based analysis. The identity of the cultured pathogens was confirmed using MALDI-TOF MS (VITEK MS, bioMérieux, Durham, NC, USA). Single colonies were diluted in TSB/glycerol and stored at -80^{o}C for later sequencing-based and phenotypic analysis. If

patients were unable to submit a specimen at a predetermined sampling point samples collected at the next closest available time point were selected for analysis. Additionally, pre-recurrence specimens of rUTI patients and time-matched samples from non-rUTI were further processed. Non-rUTI patients were matched to rUTI patients based on (1) colonization status (defined below) and (2) treatment antibiotic during the first episode.

4.5.3 Antimicrobial susceptibility testing

Antimicrobial susceptibility testing of pathogens was performed on Mueller Hinton agar (Hardy Diagnostics, Santa Maria, CA, USA) using Kirby Bauer disk diffusion with antibiotic disks purchased from Hardy Diagnostics (Santa Maria, CA, USA) and Becton Dickinson (Franklin Lakes, NJ, USA). Results were interpreted according to consensus-based medical laboratory standards as provided in the Clinical and Laboratory Standards Institute (CLSI) guidelines for antimicrobial susceptibility testing[57], which provide species-specific breakpoint definitions for determining susceptibility or resistance.

4.5.4 DNA extraction, short-read sequencing, and quality filtering

Isolates were streaked onto blood agar (Hardy Diagnostics, Santa Maria, CA, USA) and incubated at 35^{o}C overnight. Genomic DNA was extracted using the QIAamp Bacteremia DNA kit (Qiagen, Germantown, MD, USA). Sequencing libraries from both isolate gDNA and fecal metagenomic DNA were prepared using the Nextera kit (Illumina, San Diego, CA, USA)[58]. Libraries were pooled and sequenced (2 x150 bp) to a depth of ~2.5 million reads on the NextSeq 500 HighOutput platform (Illumina, San Diego, CA, USA). The

resulting reads were trimmed of adapters using Trimmomatic v.36 (parameters: LEADING:10 TRAILING:10 SLIDINGWINDOW:4:15 MINLEN:60)[59].

4.5.5 Isolate genome assembly and annotation

Draft genomes were assembled using SPAdes v.3.11.0 (parameters: -k 21,33,55,77 -careful)[60]. The resulting scaffolds.fasta files were used for analysis. The quality of draft genomes was assessed by calculating assembly statistics using QUAST v5.0.2 and checkM v.1.0.13[61,62]. High-quality assemblies (<300 contigs, >90% of genome in contigs >1000bp, completeness >90%, contamination <5%) were annotated for open reading frames with Prokka v.1.12 (default parameters, contigs > 500 bp)[63]. Twenty-four publicly available *E. coli* genomes of known phylogroup were downloaded from NCBI to use as reference and annotated as described above (Table 4.3). These genomes were used to assign phylogroups to the isolates sequenced in this study based on core-genome relatedness to the set of references. ARGs were annotated *in silico* using RGI-CARD v.5.1.0 (95% identity, 100% coverage) and Resfinder v.4.0 (95% identity, 100% coverage)[64,65].

4.5.6 Phylogenetic analysis and lineage definition

MLST were annotated *in silico* using mlst v2.11 (default parameter) and serotypes were assigned using serotypefinder v2.0.1 (parameters: -mp blast -l 0.8 -t 0.90)[66,67]. Core-genome alignments were generated using Roary v3.8.0 (default parameters, -cd 100)[68]. For sequence type-specific phylogenetic analysis core-genomes were constructed using

all isolates typed to ST 131 or 1193, respectively (Figure S2). To define lineages, all *E. coli* isolates from the same patient were used for core-genome construction. Newick trees of the core genome phylogenies were generated using FastTree v.2.1.10 (parameters: -gtr -nt) and visualized using iTOL v.4[69,70].

To define *E. coli* lineages, patient-specific pairwise core-genome SNP distances were determined from the patient-specific Roary core-genome alignments via snp-sites v.2.4.0 (default parameters)[71]. Output files were converted into SNP distance matrices using custom R and python scripts. Based on the distribution of pairwise SNP distances (Figure 4.1), *E. coli* lineages were herein defined to have <500 SNPs. Lineages were defined to be UPEC for the purpose of this study if they were isolated as the causative agent (DxU isolate) of a UTI. Pairwise ANI values between same-patient isolates were calculated using fastANI v1.3 (parameters: --fragLen 3,000, --minFraction 0.5)[72].

4.5.7 Determination of colonization patterns, lineage persistence, and rUTI causing UPEC

To understand colonization dynamics of UPEC and assess the impact of inter-habitat transfer on UPEC within-host adaptation, each UPEC lineage was categorized into one of four distinct persistence patterns: urinary tract colonization, intestinal colonization, dual, and uncolonized. Lineages were characterized as colonizing a given habitat (1) if the UPEC lineage was recovered from a habitat- specific specimen (stool/urine) at >1 collection point, or (2) if all habitat-specific specimens (stool/urine) from a UTI episode were positive for the UPEC lineage. DxU urine specimens were not considered for

classification purposes. Lineages for which either type of specimen from their corresponding patient was unavailable were left unclassified. Lineages were further classified as rUTI if (1) the patient of isolation experienced a recurrence during the follow-up period and either (2) the same lineage was isolated as the DxU isolate of a rUTI or (3) no other lineage of *E. coli* was isolated at any point during follow-up. Lineages without follow-up DxU isolates or when multiple lineages of *E. coli* were isolated from a rUTI patient were left unclassified. Lineages from non-rUTI patients were classified as non-rUTI.

4.5.8 Characterization of within-lineage allelic diversity

To determine the allelic diversity between isolates from the same lineage, "pseudo-assemblies" were constructed for each UPEC lineage, as previously described[13,24]. Equal proportions of reads from each isolate of a given lineage were pooled, assembled into a draft genome using SPAdes v.3.11.0 (parameters: -k 21,33,55,77 -careful), and annotated using Prokka v.1.12 (default parameters, contigs > 500 bp)[60,63]. These pseudo-assemblies were used as high-resolution reference genomes to characterize within-lineage allelic variation. Isolate reads were mapped to their respective pseudo-assemblies using Bowtie2 v.2.3.4 (parameters: -X 2000 --no-mixed --very-sensitive --n-ceil 0,0.01)[73]. SNPs and insertions/deletions were annotated using SAMtools v.1.9 and BCFtools v.1.9 (parameters: bcftools call -c -I 'DP>10 & QS>0.95', bcftools view -i 'FQ<-85')[74,75]. SNPs were further filtered for major allele frequency >90% and gene presence in >60% of isolates from a given lineage, to exclude SNPs in potential MGEs. Mutated loci were

mapped back to the reference GFF file (from Prokka) to identify corresponding coding sequences. Pairwise SNP distance matrices were used to construct unrooted lineage-specific phylogenetic trees, using the ape package in R v.3.6.3[76]. Time to last common ancestor (LCA) was estimated using median branch lengths of the resulting tree (determined via ape function '*edge.length*') and dividing it by the estimated rate of *E. coli* evolution of 8.9×10^{-11} per base-pair per generation[23], given an intestinal generation time of 80 minutes[77,78].

4.5.9 dMRCA estimation

To estimate dMRCA for each lineage, we generated parsimonious SNP trees using PHYLIP v3.697[79] to infer the ancestral sequence. VCF files resulting from within-lineage SNP characterization above were merged (bcftools merge --merge snps) including an isolate from the closest-related lineage according to ANI as an outgroup. The resulting VCF files were converted to '.phy' format using the s_vcf2phylip.py script (https://github.com/edgardomortiz/vcf2phylip/blob/master/vcf2phylip.py) published by Ortiz et al on Github. Files were used as input in the PHYLIP dnapars program (default parameters). Isolate dMRCA values were determined based on variable positions to the ancestral allele and used to calculate lineage averages. Lineage dMRCA values were compared between colonization types using Kruskal-Wallis with Dunn post-hoc test. *P*-values were adjusted for multiple comparisons using the Benjamini-Hochberg method (FDR).

4.5.10 Permutation test for non-random distribution of mutations

To identify non-random parallel evolution in UPEC lineages separate permutation tests were implemented for the two main colonization types: gut colonizers (gut isolates only) and dual colonizers. Mutations were randomly distributed across the lineage-specific pseudo-reference assemblies (*i.e.*, if a lineage exhibited 10 SNPs total, 10 random SNPs were assigned in the genome). This process was repeated 1000 times for all lineages. The overall simulated distribution was used as the expected (neutral) distribution to test significance. The *P*-value was calculated as the top percentile of the neutral distribution at which the observed lineage count was present. To profile UPEC within-host adaptation, gut colonizers' pseudo-reference assemblies were generated using only gut isolate reads. To profile inter-habitat, within-lineage mutations, 71 urinary isolates from the 51 gut colonizing lineages were added and permutations were re-run.

4.5.11 Estimation of dN/dS

To determine signatures of positive selection at specific genes, isolate gene sequences were aligned using Snippy v4.3.8, using as a reference the corresponding pseudo-assembly .ffn file as annotated by Prokka v3.8.0. STOP codons were masked from the Snippy snps.consensus.fa output files using a custom script. dN/dS values for each gene's lineage-specific alignment were determined in Genomegamap v1.0.1 using the Maximum Likelihood estimation[80]. Overall dN/dS values for gene groups were estimated by generating a codon-based library of all possible mutations and calculating expected N/S ratios for each gene in the gene group. Overall dN/dS values were then

calculated by summarizing the observed non-synonymous and synonymous mutations over all genes within the gene group. 95% confidence intervals were calculated by sampling from a binomial distribution as done previously[24]. Insertions/deletions as well as genes of plasmidic origin, due to their increased genetic variability[53], were masked for group-wise dN/dS calculations.

4.5.12 Identification of within-lineage genomic plasticity

The accessory gene content of each UPEC lineage was identified based on a collapsed set of non-redundant genes. Therefore, clusters homologous genes were identified using CD-HIT[81], clustering translated gene sequences clustering at >90% amino acid identity. Within-lineage Jaccard dissimilarities (distances) of accessory gene content were calculated using the VEGAN package in R v.3.6.3[82]. Average values for each lineage were used in comparisons. Dissimilarities of gene content were compared between colonization types, between and within habitat using ANOVA and Kruskal-Wallis with Dunn post-hoc. *P*-values were adjusted for multiple comparisons using the Benjamini-Hochberg method (FDR).

4.5.13 GO overrepresentation analysis (GOOA)

To gain insights into the functions under selection during UPEC persistence, we annotated GO terms of genes with non-random mutational signatures (as per the permutation test above) or habitat-specific within-lineage abundance patterns using blast2go[83]. We compared gene-set associated GO terms frequencies to their expected

value as determined using a fully GO-annotated colonization-type specific background (*i.e.*, pangenome of each colonization type). To reduce redundancy in the GO term list associated with habitat-specific genes, we clustered overlapping GO terms using REVIGO prior to analysis allowing small similarity (<0.5)[84]. Functional categories under selection during UPEC within-host persistence were identified using one-sided Fisher's exact test (hypergeometric distribution) in R v.3.6.3. *P*-values were adjusted for multiple comparisons using the Benjamini-Hochberg method (FDR). Fold-changes (enrichment scores) were calculated comparing observed vs expected values. For GO network analysis significant GOOA results were clustered semantically using REVIGO and visualized using Cytoscape[84,85].

4.5.14 Comparison with published UPEC genomes

We downloaded raw reads for 703 UPEC genomes previously curated from multiple studies from NCBI[32]. We assembled genomes using SPAdes v.3.11.0 and assemblies using Prokka v.1.12 (default parameters). We extracted the amino acid sequences of OmpC and NsfA, found to be under positive selection and associated with the gain of phenotypic antibiotic resistance in this study, from all assemblies containing these genes. We queried the mutations (SNPs and INDELs) identified in this study against the set of reference sequences and extracted sequences from UPEC genomes containing the same mutations. We performed multiple sequence alignment between variable regions from our study and UPEC genomes using Clustal Omega and visualized alignments using MView[86]. OmpC and NfsA sequences from UTI89 were used as a reference.

4.5.15 MGE identification, annotation and characterization

We identified putative MGEs differentially abundant in isolates of the same lineage by aligning short reads to the pseudo-reference assembly. Candidate regions of at least 500bp length and <0.2X relative coverage in at least one isolate were considered for further analysis. Candidate MGEs in closed genomic proximity (<1 read pair - 300bp apart) were clustered to account for sporadic read mapping into conserved genomic regions interrupting continuous MGE identification. If candidate MGEs covered >90% of a contig in the pseudo-assembly, the whole contig was defined as a candidate MGE. Coverage for all putative within-lineage MGEs was determined for all isolates and a MGE presence/absence matrix was generated based on the average relative coverage for putative MGEs in each isolate's short read alignment. <0.2X relative coverage over the complete length of the MGE equaled absence and >0.8X relative coverage equaled presence in an isolate. Intermediate values were defined to be unclear evidence of MGE presence/absence. Within-lineage similarity of isolate MGE profiles was assessed using Jaccard dissimilarities (distances) calculated using the VEGAN package in R v.3.6.3 [82]. Comparison of MGE profiles was performed using ANOVA with Tukey post-hoc test and Welch's t-test. *P*-values were adjusted for multiple comparisons using the Benjamini-Hochberg method (FDR).

MGEs were annotated similarly to a previously published protocol for *de novo* MGE identification[40]. The pool of within-lineage MGEs was queried for prophages using PHASTER[87]. MGE contigs of plasmidic origin were identified combining replicon typing using 'Plasmid MLST' with mapping within-lineage MGE contigs to the complete pool

of plasmidic contigs identified *de novo* in the draft assemblies of all isolate as previously described[13,88]. This "lineage-plasmidome" was identified using plasmidSPAdes v.3.11.0 (parameters: --plasmid -k 21,33,55,77 –careful), Recycler v.0.6.2 (parameters: -k 77 -i True), and PlasmidFinder v.4.0 (parameters: -p enterobacteriaceae -k 95.00)[89–91]. A non-redundant list of putative plasmidic contigs was validated against the NCBI plasmid database using ncbi-blast v.2.6.0+[92]. Contigs with >90% identity and >90% coverage of plasmid in the database were retained. This total "lineage-plasmidome" was annotated using Prokka v.1.12 (default parameters), the eggnog-mapper v.6.8 (parameters: -m diamond --query-cover 0.9), RGI-CARD v.5.1.0 (95% identity, 100% coverage), and Resfinder v.4.0 (95% identity, 100% coverage)[63–65,93]. MGEs were determined to be plasmidic if they (1) had an exact replicon match in the Plasmid MLST database or (2) if they aligned to a contig of *de novo* identified plasmidic origin at >80% coverage and 99% identity using ncbi-blast v.2.6.0+[92]. Insertion sequences (IS) and transposases were identified in MGEs by blasting against the ISfinder database[94]. As the repetitive nature of IS frequently causes short-read assemblies to break, incomplete IS are often found at the edge of contigs. To account for this, IS were determined to be present if either (1) a partial IS match was identified at the edge of contig with >95% identity or (2) an IS was identified at >90% identity and >80% coverage. IS elements were defined as elements that only contained an IS/Transposase and no other genes. Lastly, recombinases were identified in the Prokka annotations of the MGE pool.

Consistent with previous methods[40], the final annotation for each MGE was assigned hierarchically from specific to general as follows; (1) Intact phages, (2) Plasmid,

(3) IS element, (4) CDS+Transposase, (5) Recombinase, (6) Questionable/Incomplete phage, (7) Contains CDS, and (8) No CDS. Habitat-specific genes were identified in the MGE pool using ncbi-blast v.2.6.0+ and determined to be present if (1) coverage >90% at 99% identity or (2) coverage >10% at 100% identify and the gene was determined to be located at the edge of a contig[92].

To reduce the likelihood of false positives, GOOA of mobilized functions between rUTI and non-rUTI lineages (Figure 4.13B) was performed after filtering out GO-terms present in less than 5% of all analyzed lineages. GO term overrepresentation in the mobilized gene pool of either rUTI or non-rUTI lineages was assessed using Fisher's exact test. *P*-values were adjusted for multiple comparisons using the Benjamini-Hochberg method (FDR). Pseudo enrichment scores were calculated comparing observed GO term abundances between compared groups adding the minimal value in the array as a pseudocount.

We further assessed MGE host ranges by aligning putative MGE contigs against the NCBI nucleotide database using ncbi-blast v.2.6.0+[92], filtering for hits with >95% identity and 95% query coverage. Uncultured bacteria, eukaryotes, synthetic constructs/vectors, and mixed communities were filtered from the resulting hits. Taxa IDs were converted to species-level annotations and the number of species-level blast hits was summarized per MGE category. Statistical comparisons were performed using ANOVA and species under-represented in the urinary MGE pool were determined using one-sided Fisher's exact test. The 25 species most abundant in the blast hitlist were

considered for statistical analysis. *P*-values were corrected for multiple-hypothesis testing using the Benjamini-Hochberg method (FDR).

4.6 Data Availability

Raw sequencing data has been deposited at the NCBI SRA database under PRJNA682246.

4.7 References

1. Marvig RL, Sommer LM, Molin S, Johansen HK. Convergent evolution and adaptation of Pseudomonas aeruginosa within patients with cystic fibrosis. *Nature Genetics*. 2015;47(1):57-64. doi:10.1038/ng.3148

2. Sheppard SK, Guttman DS, Fitzgerald JR. Population genomics of bacterial host adaptation. *Nature Reviews Genetics*. 2018;19(9):549-565. doi:10.1038/s41576-018-0032-z

3. Lieberman TD, Flett KB, Yelin I, et al. Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. *Nature genetics*. 2014;46(1):82-87. doi:10.1038/ng.2848

4. Lourenço M, Ramiro RS, Güleresi D, et al. A Mutational Hotspot and Strong Selection Contribute to the Order of Mutations Selected for during Escherichia coli Adaptation to the Gut. Cooper TF, ed. *PLOS Genetics*. 2016;12(11):e1006420. doi:10.1371/journal.pgen.1006420

5. Lieberman TD, Michel JB, Aingaran M, et al. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nature Genetics*. 2011;43(12):1275-1280. doi:10.1038/ng.997

6. Gatt YE, Margalit H. Common Adaptive Strategies Underlie Within-Host Evolution of Bacterial Pathogens. *Molecular Biology and Evolution*. 2021;38(3):1101-1121. doi:10.1093/molbev/msaa278

7. Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. Within-host evolution of bacterial pathogens. *Nature Reviews Microbiology*. 2016;14(3):150-162. doi:10.1038/nrmicro.2015.13

8. Rossi E, La Rosa R, Bartell JA, et al. Pseudomonas aeruginosa adaptation and evolution in patients with cystic fibrosis. *Nature Reviews Microbiology*. 2020;19(5):331-342. doi:10.1038/s41579-020-00477-5

9. Lees JA, Kremer PHC, Manso AS, et al. Large scale genomic analysis shows no evidence for pathogen adaptation between the blood and cerebrospinal fluid niches during bacterial meningitis. *Microbial Genomics*. 2017;3(1). doi:10.1099/mgen.0.000103

10. Young BC, Wu CH, Gordon NC, et al. Severe infections emerge from commensal bacteria by adaptive evolution. *eLife*. 2017;6:e30637. doi:10.7554/eLife.30637

11. Foxman B. Urinary tract infection syndromes. Occurrence, recurrence, bacteriology, risk factors, and disease burden. *Infectious Disease Clinics of North America*. 2014;28(1):1-13. doi:10.1016/j.idc.2013.09.003

12. Flores-Mireles AL, Walker JN, Caparon M, Hultgren SJ. Urinary tract infections: epidemiology, mechanisms of infection and treatment options. *Nature Reviews Microbiology*. 2015;13(5):269-284. doi:10.1038/nrmicro3432

13. Thänert R, Reske KA, Hink T, et al. Comparative Genomics of Antibiotic-Resistant Uropathogens Implicates Three Routes for Recurrence of Urinary Tract Infections. Parkhill J, ed. *mBio*. 2019;10(4):e01977-19. doi:10.1128/mBio.01977-19

14. Chen SL, Wu M, Henderson JP, et al. Genomic diversity and fitness of E. coli strains recovered from the intestinal and urinary tracts of women with recurrent urinary tract infection. *Science translational medicine*. 2013;5(184):184ra60. doi:10.1126/scitranslmed.3005497

15. Nielsen KL, Stegger M, Godfrey PA, Feldgarden M, Andersen PS, Frimodt-Møller N. Adaptation of Escherichia coli traversing from the faecal environment to the urinary tract. *International Journal of Medical Microbiology*. 2016;306(8):595-603. doi:10.1016/j.ijmm.2016.10.005

16. Chattopadhyay S, Feldgarden M, Weissman SJ, Dykhuizen DE, Van Belle G, Sokurenko E V. Haplotype diversity in "source-sink" dynamics of Escherichia coli urovirulence. *Journal of Molecular Evolution*. 2007;64(2):204-214. doi:10.1007/s00239-006-0063-5

17. Sokurenko E V. Selection Footprint in the FimH Adhesin Shows Pathoadaptive Niche Differentiation in Escherichia coli. *Molecular Biology and Evolution*. 2004;21(7):1373-1383. doi:10.1093/molbev/msh136

18. Weissman SJ, Beskhlebnaya V, Chesnokova V, et al. Differential stability and trade-off effects of pathoadaptive mutations in the Escherichia coli FimH adhesin. *Infection and Immunity*. 2007;75(7):3548-3555. doi:10.1128/IAI.01963-06

19. Schwartz DJ, Kalas V, Pinkner JS, et al. Positively selected FimH residues enhance virulence during urinary tract infection by altering FimH conformation. *Proceedings of the National Academy of Sciences*. 2013;110(39):15530-15537. doi:10.1073/pnas.1315203110

20. Bricio-Moreno L, Sheridan VH, Goodhead I, et al. Evolutionary trade-offs associated with loss of PmrB function in host-adapted Pseudomonas aeruginosa. *Nature Communications*. 2018;9(1):1-12. doi:10.1038/s41467-018-04996-x

21. Bronson RA, Gupta C, Manson AL, et al. Global phylogenomic analyses of Mycobacterium abscessus provide context for non cystic fibrosis infections and the evolution of antibiotic resistance. *Nature Communications 2021 12:1*. 2021;12(1):1-10. doi:10.1038/s41467-021-25484-9

22. Coll F, Harrison EM, Toleman MS, et al. Longitudinal genomic surveillance of MRSA in the UK reveals transmission patterns in hospitals and the community. *Science translational medicine*. 2017;9(413):953. doi:10.1126/SCITRANSLMED.AAK9745

23. Wielgoss S, Schneider D, Barrick JE, et al. Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with escherichia coli. *G3: Genes, Genomes, Genetics*. 2011;1(3):183-186. doi:10.1534/g3.111.000406

24. Zhao S, Lieberman TD, Poyet M, et al. Adaptive Evolution within Gut Microbiomes of Healthy People Article Adaptive Evolution within Gut Microbiomes of Healthy People. *Cell Host and Microbe*. 2019;25:656-667.e8. doi:10.1016/j.chom.2019.03.007

25. Sarkar S, Ulett GC, Totsika M, Phan MD, Schembri MA. Role of capsule and O antigen in the virulence of uropathogenic Escherichia coli. *PloS one*. 2014;9(4):e94786. doi:10.1371/JOURNAL.PONE.0094786

26. Su CC, Rutherford DJ, Yu EW. Characterization of the multidrug efflux regulator AcrR from Escherichia coli. *Biochemical and Biophysical Research Communications*. 2007;361(1):85-90. doi:10.1016/j.bbrc.2007.06.175

27. Osei Sekyere J. Genomic insights into nitrofurantoin resistance mechanisms and epidemiology in clinical Enterobacteriaceae. *Future Science OA*. 2018;4(5):FSO293. doi:10.4155/fsoa-2017-0156

28. Choi U, Lee CR. Distinct Roles of Outer Membrane Porins in Antibiotic Resistance and Membrane Integrity in Escherichia coli. *Frontiers in Microbiology*. 2019;10(APR). doi:10.3389/fmicb.2019.00953

29. Beloin C, Michaelis K, Lindner K, et al. The Transcriptional Antiterminator RfaH Represses Biofilm Formation in Escherichia coli. *Journal of Bacteriology*. 2006;188(4):1316-1331. doi:10.1128/JB.188.4.1316-1331.2006

30. Mann R, Mediati DG, Duggin IG, Harry EJ, Bottomley AL. Metabolic adaptations of Uropathogenic E. coli in the urinary tract. *Frontiers in Cellular and Infection Microbiology*. 2017;7(JUN):241. doi:10.3389/fcimb.2017.00241

31. Hibbing ME, Dodson KW, Kalas V, Chen SL, Hultgren SJ. Adaptation of arginine synthesis among uropathogenic branches of the escherichia coli phylogeny reveals adjustment to the urinary tract habitat. *mBio*. 2020;11(5):1-15. doi:10.1128/MBIO.02318-20/SUPPL_FILE/MBIO.02318-20-SF009.PDF

32. Biggel M, Xavier BB, Johnson JR, et al. Horizontally acquired papGII-containing pathogenicity islands underlie the emergence of invasive uropathogenic Escherichia coli lineages. *Nature Communications 2020 11:1*. 2020;11(1):1-15. doi:10.1038/s41467-020-19714-9

33. Hammond JA, Gordon EA, Socarras KM, Mell JC, Ehrlich GD. Beyond the pan-genome: Current perspectives on the functional and practical outcomes of the distributed genome hypothesis. *Biochemical Society Transactions*. 2020;48(6):2437-2455. doi:10.1042/BST20190713

34. Darch SE, McNally A, Harrison F, et al. Recombination is a key driver of genomic and phenotypic diversity in a Pseudomonas aeruginosa population during cystic fibrosis infection. *Scientific Reports*. 2015;5:1-12. doi:10.1038/srep07649

35. Gabrielaite M, Johansen HK, Molin S, Nielsen FC, Marvig RL. Gene loss and acquisition in lineages of pseudomonas aeruginosa evolving in cystic fibrosis patient airways. *mBio*. 2020;11(5):1-16. doi:10.1128/mBio.02359-20

36. Robinson AE, Heffernan JR, Henderson JP. The iron hand of uropathogenic Escherichia coli: The role of transition metal control in virulence. *Future Microbiology*. 2018;13(7):813-829. doi:10.2217/fmb-2017-0295

37. Elhenawy W, Hordienko S, Gould S, et al. High-throughput fitness screening and transcriptomics identify a role for a type IV secretion system in the pathogenesis of Crohn's disease-associated Escherichia coli. *Nature Communications*. 2021;12(1):2032. doi:10.1038/s41467-021-22306-w

38. Jones SA, Gibson T, Maltby RC, et al. Anaerobic Respiration of Escherichia coli in the Mouse Intestine. *Infection and Immunity*. 2011;79(10):4218-4226. doi:10.1128/IAI.05395-11

39. Spaulding CN, Klein RD, Ruer S, et al. Selective depletion of uropathogenic E. coli from the gut by a FimH antagonist. *Nature*. 2017;546(7659):528-532. doi:10.1038/nature22972

40. Durrant MG, Li MM, Siranosian BA, Montgomery SB, Bhatt AS. A Bioinformatic Analysis of Integrative Mobile Genetic Elements Highlights Their Role in Bacterial Adaptation. *Cell Host and Microbe*. 2020;27(1):140-153.e9. doi:10.1016/j.chom.2019.10.022

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

42. Schreiber HL, Spaulding CN, Dodson KW, Livny J, Hultgren SJ. One size doesn't fit all: Unraveling the diversity of factors and interactions that drive E. coli urovirulence. *Annals of Translational Medicine*. 2017;5(2). doi:10.21037/atm.2016.12.73

43. Khademi SMH, Sazinas P, Jelsbak L. Within-Host Adaptation Mediated by Intergenic Evolution in Pseudomonas aeruginosa. Golding B, ed. *Genome Biology and Evolution*. 2019;11(5):1385-1397. doi:10.1093/gbe/evz083

44. Fajardo-Lubián A, Ben Zakour NL, Agyekum A, Qi Q, Iredell JR. Host adaptation and convergent evolution increases antibiotic resistance without loss of virulence in a major human pathogen. *PLoS Pathogens*. 2019;15(3):e1007218. doi:10.1371/journal.ppat.1007218

45. Tang F, Saier MH. Transport proteins promoting Escherichia coli pathogenesis. *Microbial Pathogenesis*. 2014;71-72(1):41-55. doi:10.1016/j.micpath.2014.03.008

46. Russell CW, Fleming BA, Jost CA, et al. Context-dependent requirements for FimH and other canonical virulence factors in gut colonization by extraintestinal pathogenic Escherichia coli. *Infection and Immunity*. 2018;86(3):e00746-17. doi:10.1128/IAI.00746-17

47. Brown SP, Cornforth DM, Mideo N. Evolution of virulence in opportunistic pathogens: Generalism, plasticity, and control. *Trends in Microbiology*. 2012;20(7):336-342. doi:10.1016/j.tim.2012.04.005

48. Rau MH, Marvig RL, Ehrlich GD, Molin S, Jelsbak L. Deletion and acquisition of genomic content during early stage adaptation of Pseudomonas aeruginosa to a human host environment. *Environmental Microbiology*. 2012;14(8):2200-2211. doi:10.1111/j.1462-2920.2012.02795.x

49. Sokurenko E V., Gomulkiewicz R, Dykhuizen DE. Source-sink dynamics of virulence evolution. *Nature Reviews Microbiology*. 2006;4(7):548-555. doi:10.1038/nrmicro1446

50. Shee C, Gibson JL, Darrow MC, Gonzalez C, Rosenberg SM. Impact of a stress-inducible switch to mutagenic repair of DNA breaks on mutation in Escherichia coli. *Proceedings of the National Academy of Sciences*. 2011;108(33):13659-13664. doi:10.1073/PNAS.1104681108

51. Slater FR, Bruce KD, Ellis RJ, Lilley AK, Turner SL. Heterogeneous selection in a spatially structured environment affects fitness tradeoffs of plasmid carriage in pseudomonads. *Applied and Environmental Microbiology*. 2008;74(10):3189-3197. doi:10.1128/AEM.02383-07

52. Slater FR, Bruce KD, Ellis RJ, et al. Determining the Effects of a Spatially Heterogeneous Selection Pressure on Bacterial Population Structure at the Sub-millimetre Scale. *Microbial Ecology*. 2010;60:873-884. doi:10.1007/s00248-010-9687-5

53. Rodríguez-Beltrán J, DelaFuente J, León-Sampedro R, MacLean RC, San Millán Á. Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nature Reviews Microbiology*. Published online January 19, 2021:1-13. doi:10.1038/s41579-020-00497-1

54. Harrison E, Hall JPJ, Brockhurst MA. Migration promotes plasmid stability under spatially heterogeneous positive selection. *Proceedings of the Royal Society B: Biological Sciences*. 2018;285(1879):20180324. doi:10.1098/rspb.2018.0324

55. Harrison E, Guymer D, Spiers AJ, Paterson S, Brockhurst MA. Parallel Compensatory Evolution Stabilizes Plasmids across the Parasitism-Mutualism Continuum. *Current Biology*. 2015;25(15):2034-2039. doi:10.1016/J.CUB.2015.06.024

56. Hall JPJ, Wright RCT, Harrison E, et al. Plasmid fitness costs are caused by specific genetic conflicts enabling resolution by compensatory mutation. *PLOS Biology*. 2021;19(10):e3001225. doi:10.1371/JOURNAL.PBIO.3001225

57. Melvin P. Weinstein M. *M100Ed29 | Performance Standards for Antimicrobial Susceptibility Testing, 29th Edition*. 29th ed.; 2018.

58. Baym M, Kryazhimskiy S, Lieberman TD, Chung H, Desai MM, Kishony R. Inexpensive Multiplexed Library Preparation for Megabase-Sized Genomes. Green SJ, ed. *PLOS ONE*. 2015;10(5):e0128036. doi:10.1371/journal.pone.0128036

59. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-2120. doi:10.1093/bioinformatics/btu170

60. Bankevich A, Nurk S, Antipov D, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*. 2012;19(5):455-477. doi:10.1089/cmb.2012.0021

61. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*. 2015;25(7):1043-1055. doi:10.1101/gr.186072.114

62. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29(8):1072-1075. doi:10.1093/bioinformatics/btt086

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

64. Jia B, Raphenya AR, Alcock B, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic acids research*. 2017;45(D1):D566-D573. doi:10.1093/nar/gkw1004

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

66. Larsen M V., Cosentino S, Rasmussen S, et al. Multilocus Sequence Typing of Total-Genome-Sequenced Bacteria. *Journal of Clinical Microbiology*. 2012;50(4):1355-1361. doi:10.1128/JCM.06094-11

67. Joensen KG, Tetzschner AMM, Iguchi A, Aarestrup FM, Scheutz F. Rapid and easy in silico serotyping of Escherichia coli isolates by use of whole-genome sequencing data. *Journal of Clinical Microbiology*. 2015;53(8):2410-2426. doi:10.1128/JCM.00008-15

68. Page AJ, Cummins CA, Hunt M, et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*. 2015;31(22):3691-3693. doi:10.1093/bioinformatics/btv421

69. Price MN, Dehal PS, Arkin AP. Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*. 2009;26(7):1641-1650. doi:10.1093/molbev/msp077

70. Letunic I, Bork P. Interactive Tree of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Research*. 2019;47(W1). doi:10.1093/nar/gkz239

71. Page AJ, Taylor B, Delaney AJ, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microbial genomics*. 2016;2(4):e000056. doi:10.1099/mgen.0.000056

72. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications*. 2018;9(1):1-8. doi:10.1038/s41467-018-07641-9

73. Langmead B, Wilks C, Antonescu V, Charles R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics*. 2019;35(3):421-432. doi:10.1093/bioinformatics/bty648

74. Danecek P, McCarthy SA. BCFtools/csq: Haplotype-aware variant consequences. *Bioinformatics*. 2017;33(13):2037-2039. doi:10.1093/bioinformatics/btx100

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

76. Paradis E, Schliep K. Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*. 2019;35(3):526-528. doi:10.1093/bioinformatics/bty633

77. Poulsen LK, Licht TR, Rang C, Krogfelt KA, Molin S. Physiological state of Escherichia coli BJ4 growing in the large intestines of streptomycin-treated mice. *Journal of Bacteriology*. 1995;177(20):5840-5845. doi:10.1128/jb.177.20.5840-5845.1995

78. Rang CU, Licht TR, Midtvedt T, et al. Estimation of growth rates of Escherichia coli BJ4 in streptomycin- treated and previously germfree mice by in situ rRNA hybridization. *Clinical and Diagnostic Laboratory Immunology*. 1999;6(3):434-436. doi:10.1128/cdli.6.3.434-436.1999

79. Felsenstein J. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*. 1989;5:164-166.

80. Wilson DJ. GenomegaMap: Within-Species Genome-Wide dN=dS Estimation from over 10,000 Genomes. *Molecular Biology and Evolution*. 2021;37(8):2450-2460. doi:10.1093/MOLBEV/MSAA069

81. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012;28(23):3150-3152. doi:10.1093/bioinformatics/bts565

82. Dixon P. VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science*. 2003;14(6):927-930. doi:10.1111/j.1654-1103.2003.tb02228.x

83. Götz S, García-Gómez JM, Terol J, et al. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*. 2008;36(10):3420-3435. doi:10.1093/nar/gkn176

84. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. Gibas C, ed. *PLoS ONE*. 2011;6(7):e21800. doi:10.1371/journal.pone.0021800

85. Shannon P, Markiel A, Ozier O, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*. 2003;13(11):2498-2504. doi:10.1101/gr.1239303

86. Madeira F, Park YM, Lee J, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*. 2019;47(W1):W636-W641. doi:10.1093/NAR/GKZ268

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

88. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications [version 1; referees: 2 approved]. *Wellcome Open Research*. 2018;3. doi:10.12688/wellcomeopenres.14826.1

89. Antipov D, Hartwick N, Shen M, Raiko M, Lapidus A, Pevzner PA. plasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics*. 2016;32(22):btw493. doi:10.1093/bioinformatics/btw493

90. Rozov R, Brown Kav A, Bogumil D, et al. Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics (Oxford, England)*. 2017;33(4):475-482. doi:10.1093/bioinformatics/btw651

91. Carattoli A, Zankari E, García-Fernández A, et al. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrobial agents and chemotherapy*. 2014;58(7):3895-3903. doi:10.1128/AAC.02412-14

92. McGinnis S, Madden TL. BLAST: At the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*. 2004;32(WEB SERVER ISS.):W256-W259. doi:10.1093/nar/gkh435

93. Huerta-Cepas J, Szklarczyk D, Heller D, et al. EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*. 2019;47(D1):D309-D314. doi:10.1093/nar/gky1085

94. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic acids research*. 2006;34(Database issue):D32-D36. doi:10.1093/nar/gkj014

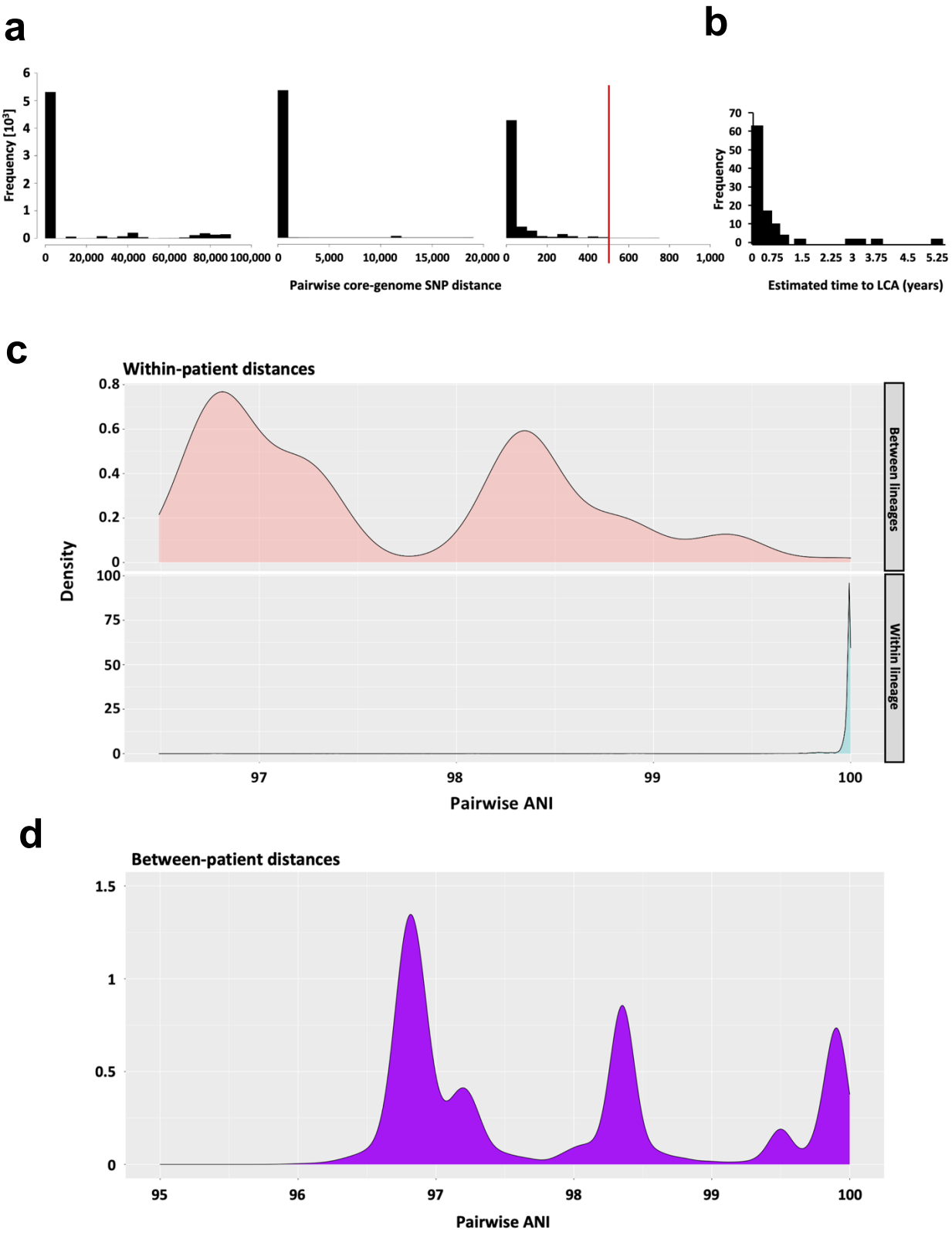

Figure 4.1 UPEC Lineage definition. (a) Histogram of *E. coli* pairwise within-patient core-genome SNP distances. Panels from left to right depict the same data using sequentially shorter x-axis ranges. Red line indicates cutoff used to define lineages. (**b)**

Histogram of time to last common ancestor for UPEC lineages applying a 500 core-genome SNP cutoff to define lineages. **(c)** Pairwise ANI values between same-patient isolates of different (top) and the same (bottom) *E. coli* lineage applying a 500 core-genome SNP cutoff to define lineages. **(d)** Pairwise ANI values between different-patient isolates.

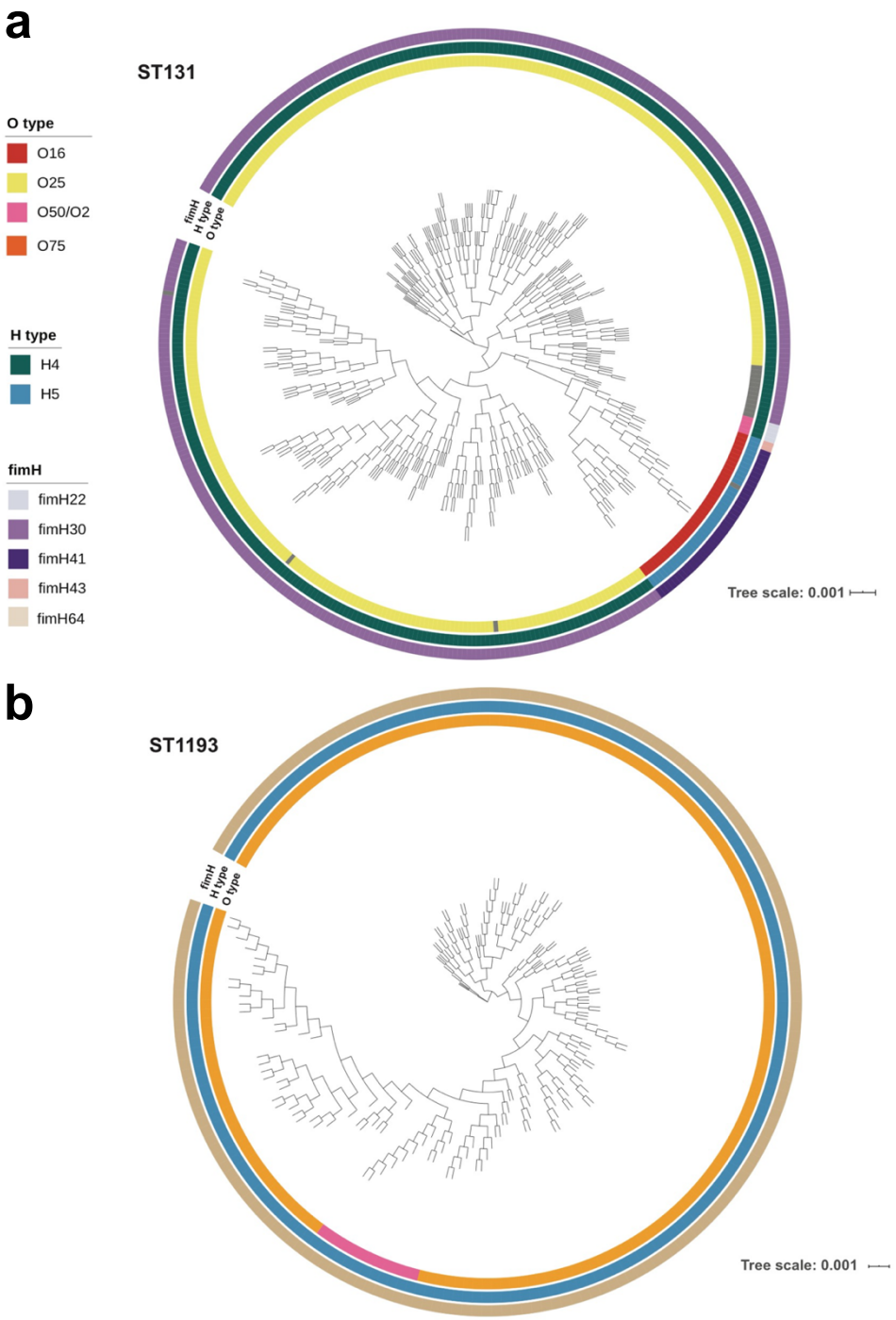

Figure 4.2 Phylogenetic analysis of ST131 and ST1193. Unrooted core genome phylogeny of *E. coli* **(a)** ST131 and **(b)** ST1193. The outer rings annotate the O-type, H-type, and *fimH*-type of each isolate.

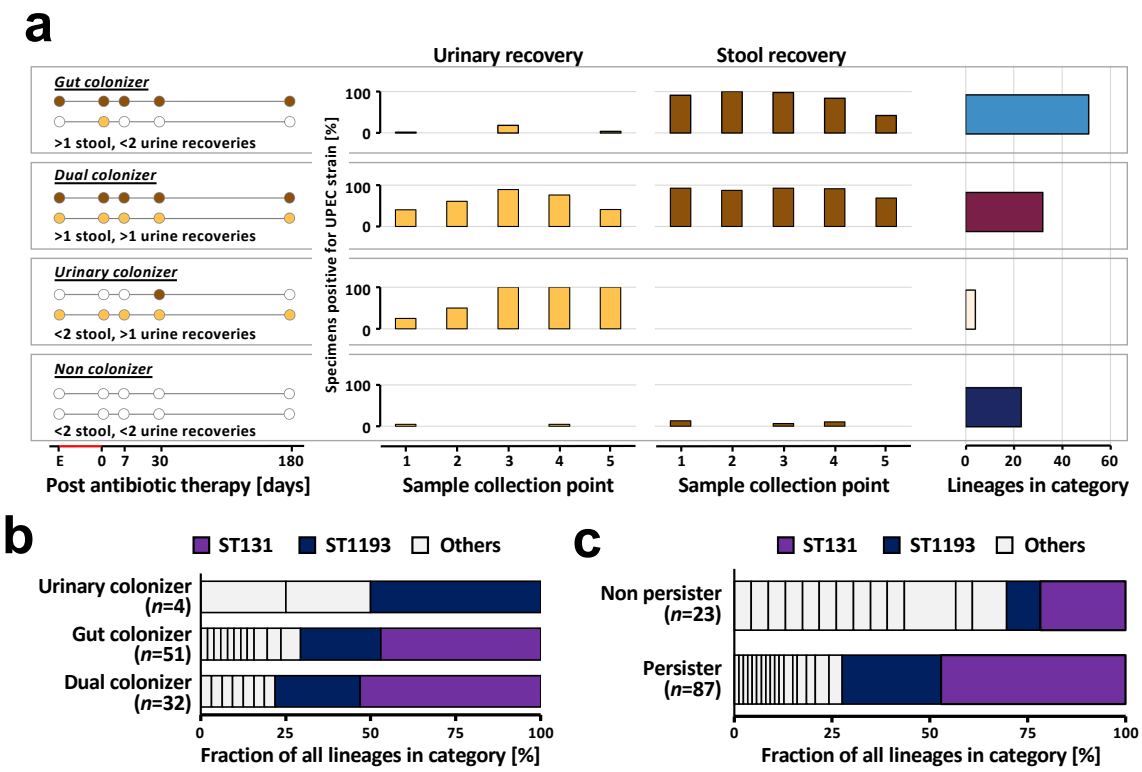

Figure 4.3 Persistent UPEC lineages group into distinct colonization patterns. (a) Schematic representation of UPEC colonization patterns (Left) as determined by recovery from stool (brown circles) and urine (yellow circles) from UTI patients with available DxU isolates. The definition for each colonization type is given below the schematic. UPEC lineages (*n*=119) are classified into four persistence types: gut colonizer, dual colonizer, urinary colonizer, and non colonizer. (Middle) UPEC lineage presence at follow-up sample collection points as determined by whole genome sequencing of isolates (Key: *1*: enrollment; *2*: 0-3 days post-antibiotic treatment (pAT); *3*: 7-14 days pAT; *4*: 30-60 days pAT; *5*: 150-180 days pAT). Bars indicate the fraction of patient's urine (yellow) and stool (brown) specimens positive for the disease causing UPEC lineage at each sampling point. Patients are grouped by UPEC lineage persistence type. Only data from the first episode caused by a UPEC lineage is shown. (Right) Number of UPEC lineages falling into each colonization category (gut colonizer=51, dual colonizer=32, urinary colonizer=4, and non colonizer=23). Boxes group together panels showing data of the same persistence type. **(b)** Sequence types (ST) are evenly distributed between UPEC persistence types. Prevalence of the two dominant STs, ST131 and ST1193, is color highlighted. **(c)** ST composition varies significantly between persisting and non-persisting lineages (n=110 lineages, Fisher's exact test, $P<0.001$). ST131 (light purple) and ST1193 (dark purple) are significantly underrepresented in the set of non-persisting UPEC lineages (*n*=110 lineages, Fisher's exact test, $P<0.001$). Prevalence of the two dominant STs, ST131 and ST1193, is color highlighted.

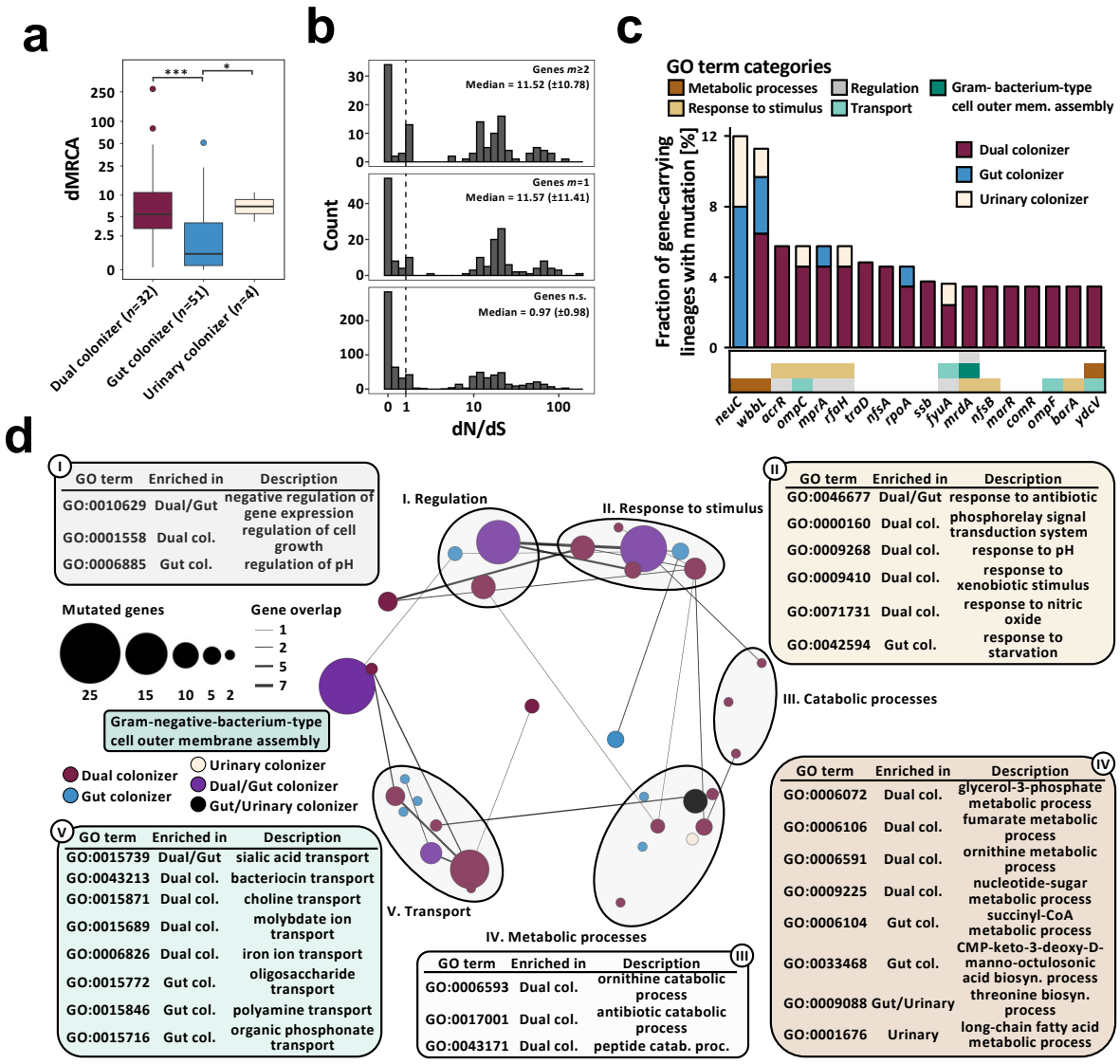

Figure 4.4 Niche-specific adaptation shapes UPEC within-host adaptation. **(a)** Boxplot of lineage dMRCA values (n=87 lineages, Kruskal-Wallis P=1.38e^{-05}, Dunn post-hoc test gut vs dual colonizer P=2.39e^{-05}, gut vs urinary colonizer P=3.32e^{-02}). Outliers (outside 1.5x interquartile range) are depicted as points. Whiskers represent 1.5x interquartile range. Upper, middle, and lower box lines indicate 75th, 50th, and 25th percentiles, respectively. **(b)** Histogram of gene-wise dN/dS values with signatures of non-random mutation (Permutation test, P<0.05) mutated in parallel across more than two lineages (m≥2, top) or in one lineage (m=1, middle), and in genes non-significant in permutation test (bottom). Median and median absolute deviation (MAD) are given for both gene groups. Dashed vertical line indicates neutral selection at dN/dS=1. **(d)** Genes found to be mutated in parallel in ≥3 lineages, normalized by the total number of gene-carrying lineages. Hypothetical genes are not shown. Color of the bar corresponds to colonization type in which mutations were found (gut colonizer - blue, dual colonizer - maroon,

urinary colonizer - light yellow). Color bar below the histogram provides GO category (as shown in Figure 4.4d) for all genes with GO terms annotation found to be significantly enriched in a colonization type. **(d)** Network visualization of GO terms significantly overrepresented in the pool of genes with non-random signature of selection within-lineages as defined by the permutation test. Bubble size represents number of mutations in genes categorized into each GO term. Color of bubbles corresponds to colonization type GO terms were enriched in (*gut colonizer*: blue; *dual colonizer*: maroon; *urinary colonizer*: light yellow; *gut/dual colonizer*: purple; *gut/urinary colonizer*: black). GO terms were clustered semantically into the 2D space using REVIGO. Circles group together semantically related GO terms.

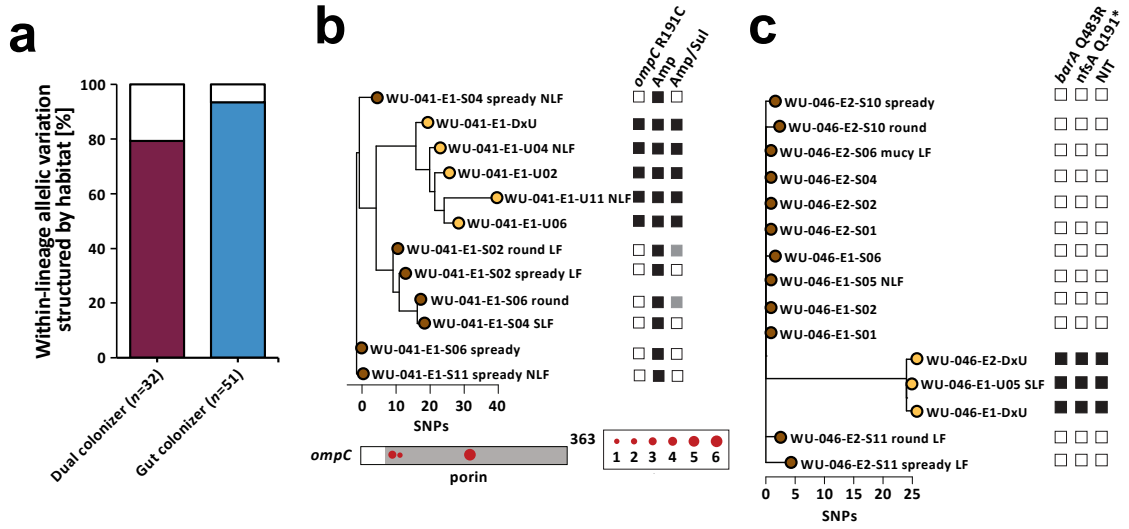

Figure 4.5 UPEC niche-specific adaptation impacts antibiotic resistance phenotypes. **(a)** The majority of allelic diversity in genes found to be mutated in parallel within gut and dual colonizers is structured by habitat (Fisher's exact test P=0.001). Color of the bar corresponds to either dual colonizer (maroon) or gut colonizers (blue). **(b)** (Top) Phylogeny of lineage WU-041_1 with annotated nonsynonymous *ompC* mutation and corresponding phenotypic resistance to ampicillin/sulbactam. Black squares denote gene presence or antibiotic resistance. White squares indicate gene absence or drug susceptibility. Grey squares indicate intermediate drug susceptibility. Phylogeny is unrooted based on SNP distances. (Bottom) SNP locations on the *ompC* gene. The porin domain is annotated in grey. Circle size corresponds to number of isolates carrying that mutation. **(c)** Lineage WU-046_2 exhibited nonsynonymous *barA* and *nfsA* mutations in urinary isolates only, corresponding to phenotypic resistance to nitrofurantoin. Phylogeny is unrooted based on SNP distances.

a

ompC

Isolate ID		37		Study PMID
UTI89	23	EVYNKDGNKLDLYGKVDGLHYFSDDKSVD	51	RefSeq
WU-018_1	23	EVYNKDGNKLDLYGEVDGLHYFSDDKSVD	51	This study
PN-029_1	23	EVYNKDGNKLDLYGEVDGLHYFSDDKSVD	51	This study
HVH_100	23	EVYNKDGNKLDLYGEVDGLHYFSDDKSVD	51	22571989
HVH_158	23	EVYNKDGNKLDLYGEVDGLHYFSDDKSVD	51	22571989
RT_ABU_295	23	EVYNKDGNKLDLYGEVDGLHYFSDNKSED	51	33235212
MVAST0093	23	EVYNKDGNKLDLYGEVDGLHYFSDNKSED	51	30304506
LtABU12	23	EVYNKDGNKLDLYGEVDGLHYFSDNKSED	51	30304506

b

ompC

Isolate ID		191		Study PMID
UTI89	176	SVSGEGMTNNGRGALRQNGDGVGGSITYDYE	206	RefSeq
WU-041_1	176	SVSGEGMTNNGRGALCQNGDGVGGSITYDYE	206	This study
PN-004_1	176	SVSGEGMTNNGRGALCQNGDGVGGSITYDYE	206	This study
LtABU15	176	SVDGEGMTNNGRGALCQNGDGVGGSITYDYE	206	30304506

c

nfsA

Isolate ID		191		Study PMID
UTI89	176	VHENSYQPLDKDALAQYDEQLAEYYLTRGSN	206	RefSeq
WU-046_2	176	VHENSYQPLDKDALA*----------		This study
ABU_9	176	VHENSYQPLDKDALA*----------		33235212

Figure 4.6 Multiple sequence alignment of variable regions in *ompC* and *nfsA*. (a) Multiple sequence alignment of region of *ompC* region 23-51 between lineages with the 37 K->E found in this study and previously published genomes. **(b)** Multiple sequence alignment of region of *ompC* region 176-206 between lineages with the 191 R->S found in this study and previously published genomes. **(c)** Multiple sequence alignment of region of *nfsA* region 176-206 between lineages with the 191 Q->* found in this study and previously published genomes. UTI89 sequence is added as a reference in all panels. Study PMID for published genomes are provided.

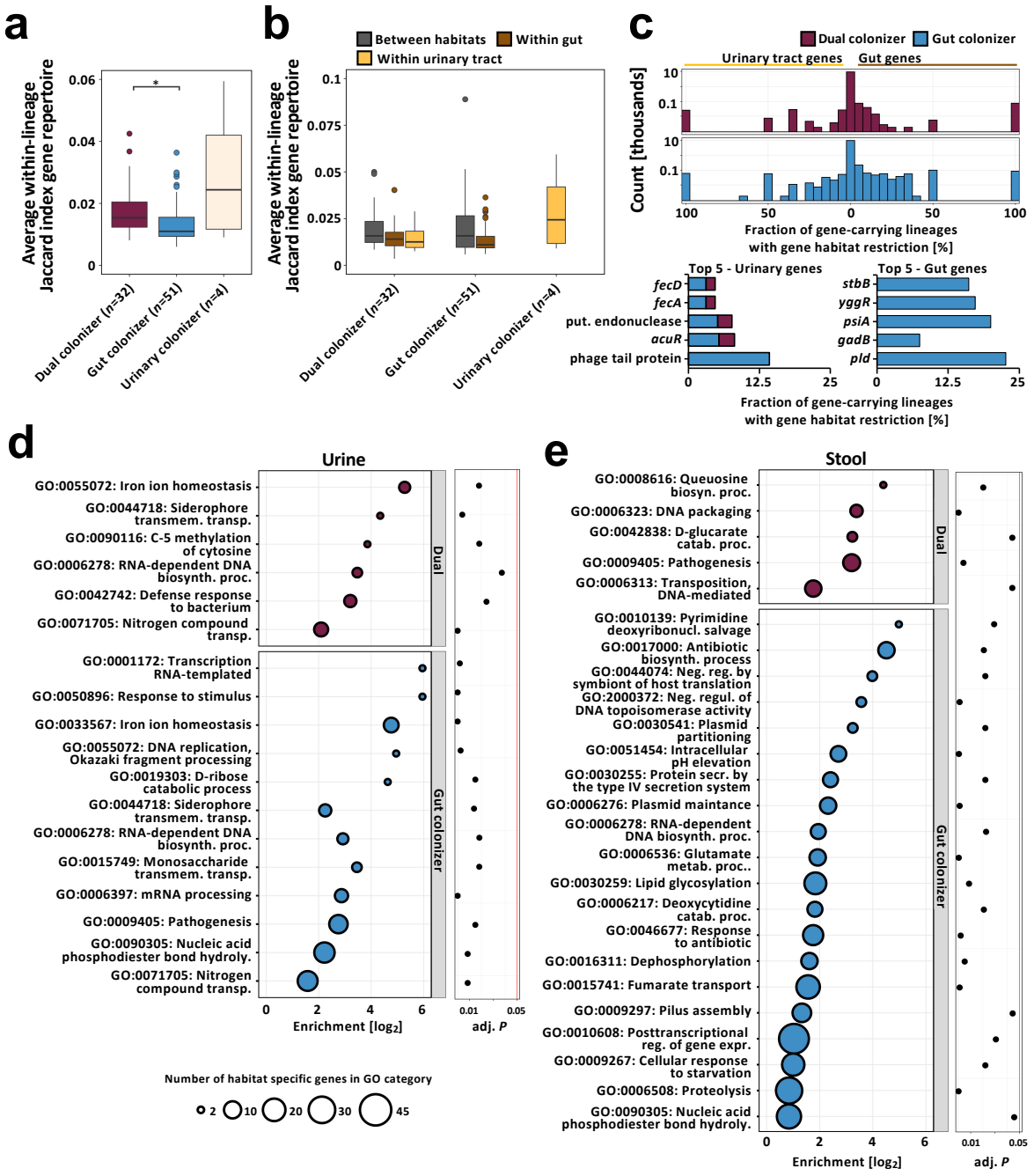

Figure 4.7 Persisting UPEC lineages exhibit niche-specific genomic plasticity. (a) Boxplot of average within-lineage Jaccard distances based on gene presence/absence data (n=87 lineages, Kruskal-Wallis test P=0.009, Dunn post-hoc test gut vs dual colonizer P=0.012). Outliers (outside 1.5x interquartile range) are depicted as points. Whiskers represent 1.5x interquartile range. Upper, middle, and lower box lines indicate 75th, 50th, and 25th percentiles, respectively. **(b)** Average between- and within-habitat lineage Jaccard distances based on gene presence/absence data of same-lineage isolates by colonization type (n=87 lineages, Two-way ANOVA, habitat P=5.94e^{-4}, colonization type

P >0.05). Outliers (outside 1.5x interquartile range) are depicted as points. Whiskers represent 1.5x interquartile range. Upper, middle, and lower box lines indicate 75th, 50th, and 25th percentiles, respectively. Colors correspond to within-lineage comparison (*between habitats*: grey; *within gut*: brown; *within urinary tract*: yellow). **(c)** (Top) Two-sided histogram of within-lineage habitat-specific genes of dual (maroon) and gut (blue) colonizers. Urinary-specific genes are shown towards the left. Gut-specific genes are shown towards the right. (Bottom) Genes most frequently found to be urine (left) or gut (right) specific across lineages, normalized by the total number of gene-carrying lineages. Bar color corresponds to the colonization type a gene was found in as habitat specific. Hypothetical genes are not shown**. (d)** Overrepresented GO terms associated with urine specific genes of dual (top - maroon) or gut colonizers (bottom - blue). Bubble size corresponds to the number of habitat-specific genes in each GO term. **(e)** Overrepresented GO terms associated with stool specific genes**.**

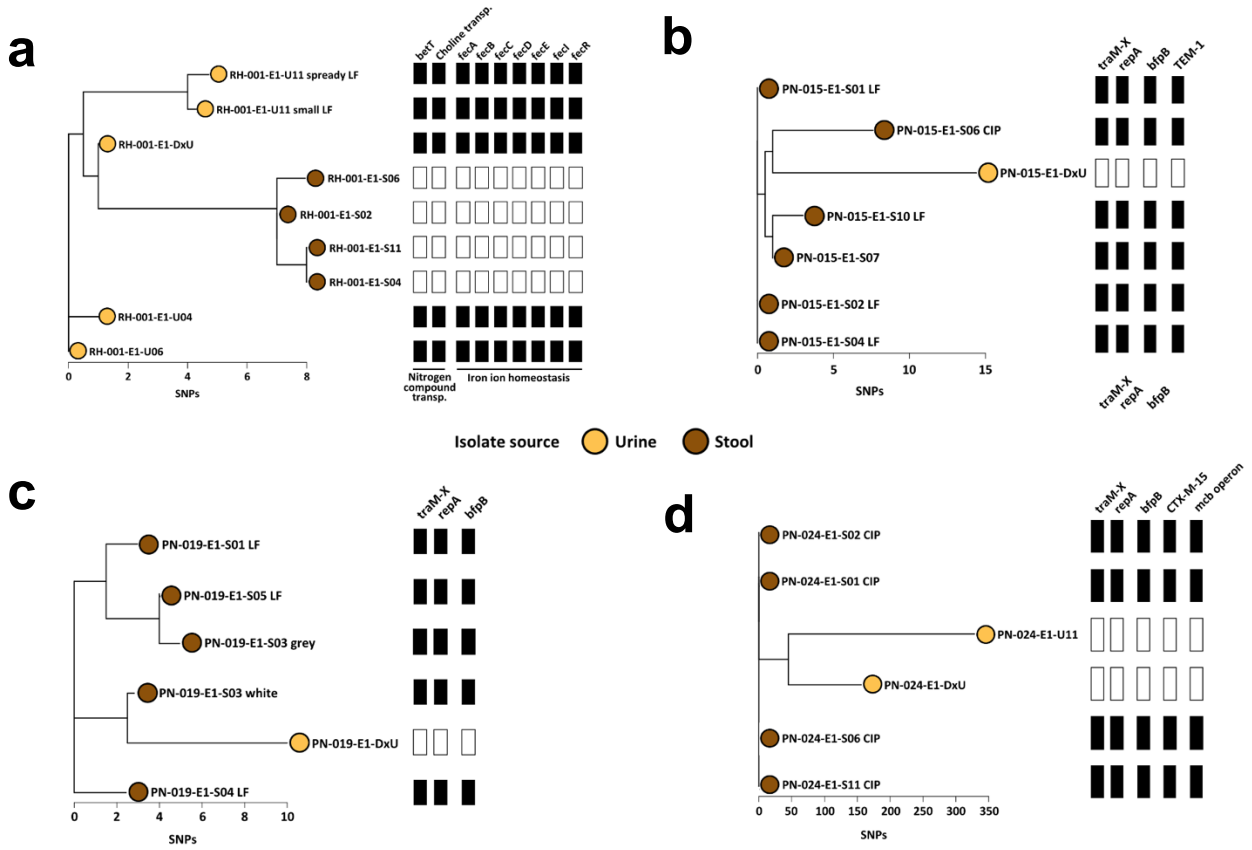

Figure 4.8 A set of virulence and resistance genes is habitat-specific in persisting UPEC lineages. (a) Unrooted phylogeny of lineage RH-001_1 based on SNP distances annotated with selected habitat specific genes. **(b)** Unrooted phylogeny of lineage PN-015_1 based on SNP distances annotated with selected habitat specific genes. **(c)** Unrooted phylogeny of lineage PN-019_2 based on SNP distances annotated with selected habitat specific genes. **(d)** Unrooted phylogeny of lineage PN-024_1 based on SNP distances annotated with selected habitat specific genes.

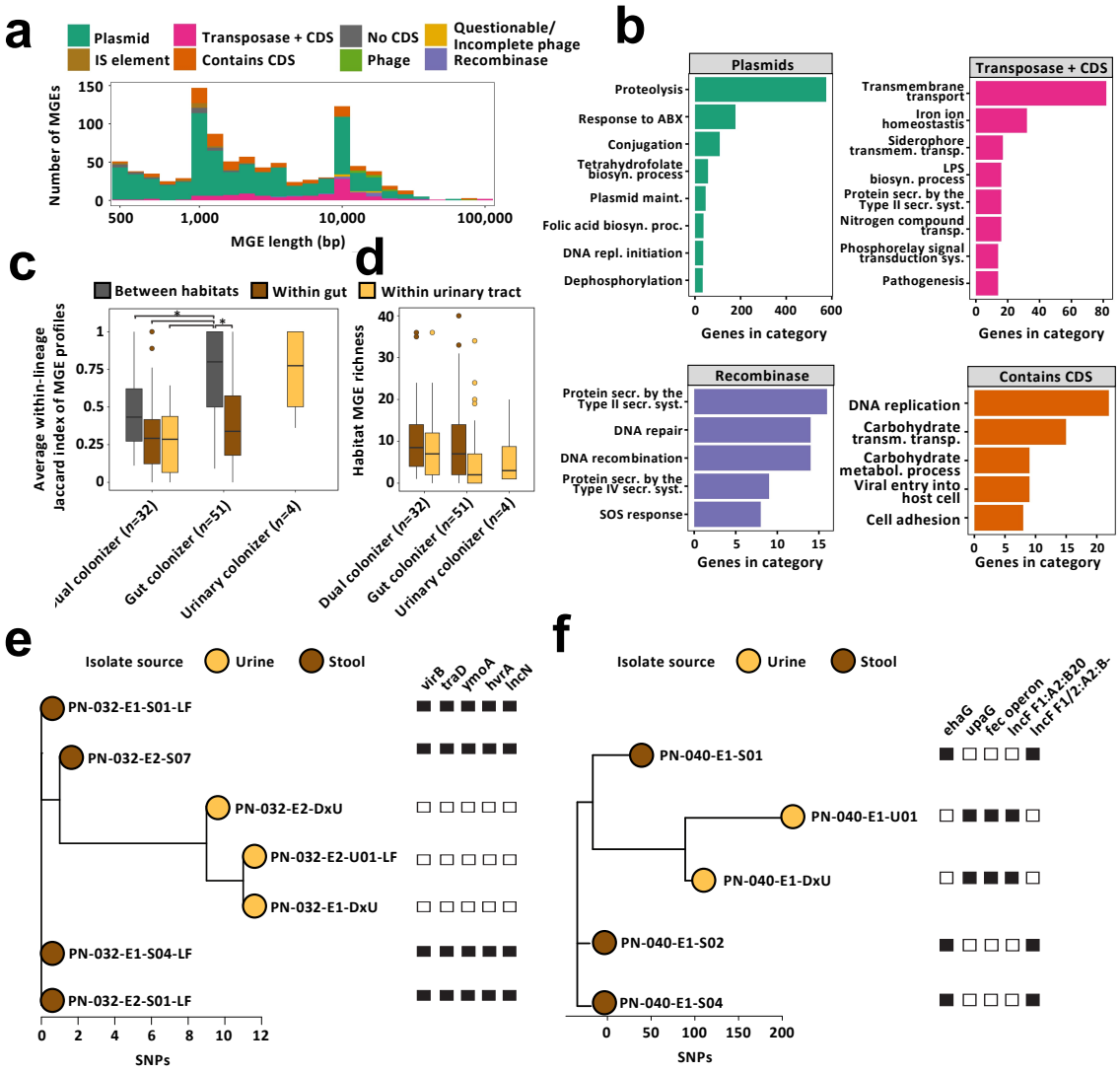

Figure 4.9 Mobile genetic elements drive niche-specific genomic plasticity of UPEC. **(a)** Visualization of within-lineage MGEs. Element length (log-scale) is plotted against element count. IS, insertion sequence; CDS, coding sequence. **(b)** GO terms overrepresented in selected MGE subclasses. **(c)** Box plot of average within-lineage Jaccard distance based on MGE presence/absence data of same-lineage isolates between habitats (grey), within gut (brown), and within urine (yellow) grouped by colonization type. All comparisons are statistically significant (n=87 lineages, Two-way ANOVA $P \leq 1.57e^{-05}$, Tukey post-hoc gut colonizer within-gut vs between habitats $P<0.001$, gut colonizer between habitat vs dual colonizer between habitat $P=0.014$). **(d)** MGE richness is larger in gut compared to urine isolates (n=87 lineages, Two-way ANOVA $P=0.042$). Outliers (outside 1.5x interquartile range) are depicted as points. Whiskers represent 1.5x interquartile range. Upper, middle, and lower box lines indicate 75th, 50th, and 25th percentiles, respectively. **(e)** Unrooted phylogeny of lineage PN-040_1 based on SNP

distances annotated with selected habitat-specific genes. Relative short-read coverage over selected, habitat-specific MGEs harboring depicted genes is shown. **(f)** Unrooted phylogeny of lineage PN-004_1 based on SNP distances annotated with selected habitat-specific genes. Relative short-read coverage over selected, habitat-specific MGEs harboring depicted genes is shown.

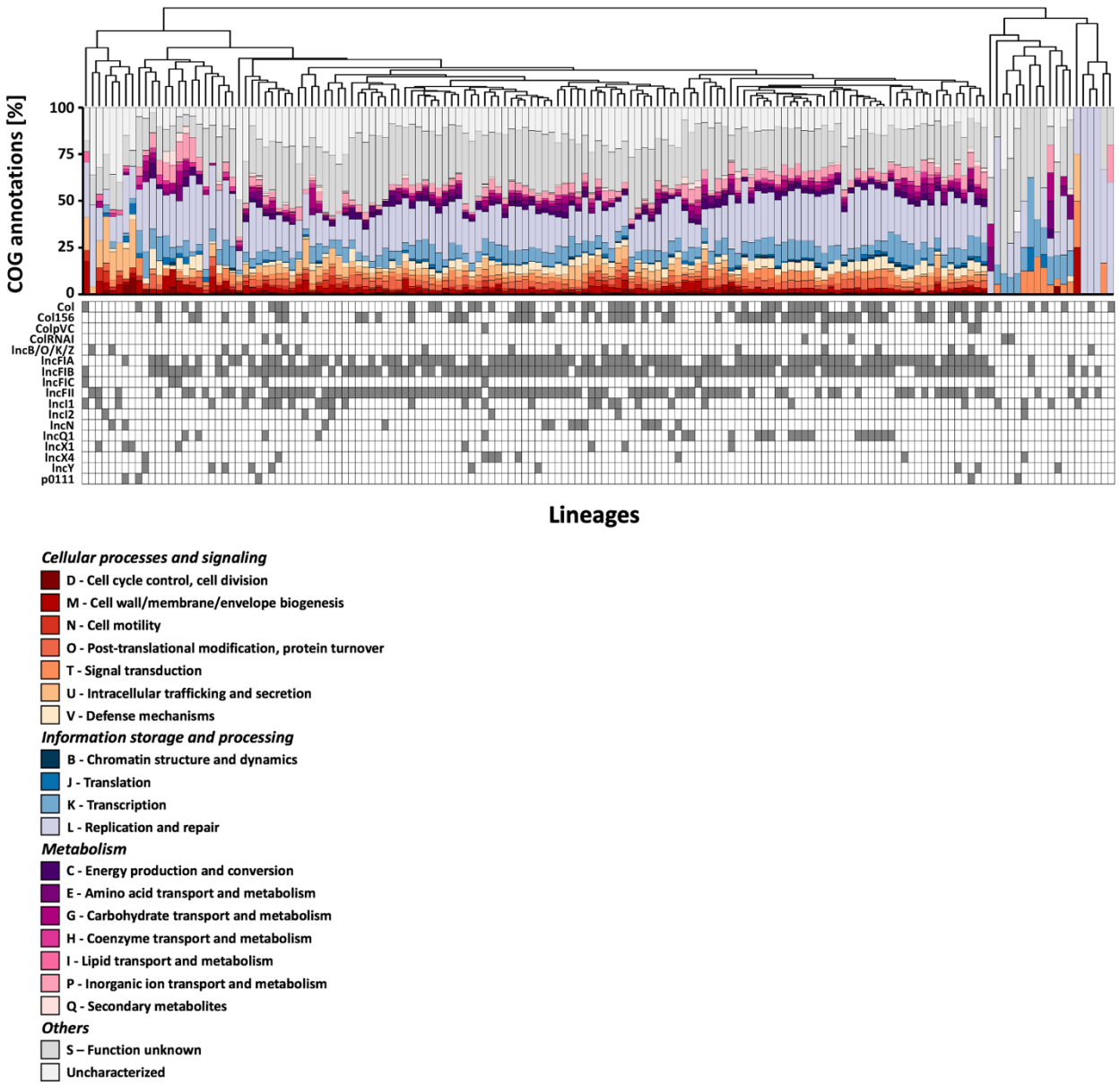

Figure 4.10 The predicted lineage-specific plasmid repertoire of AR *E. coli* differs. (Top) Lineage-specific GO-term annotation of coding sequences on contigs identified *in silico* to be of putative plasmidic origin. Only lineages with predicted plasmidic contigs are shown. (Bottom) Corresponding lineage-specific replicon-repertoire as determined using plasmidFinder.

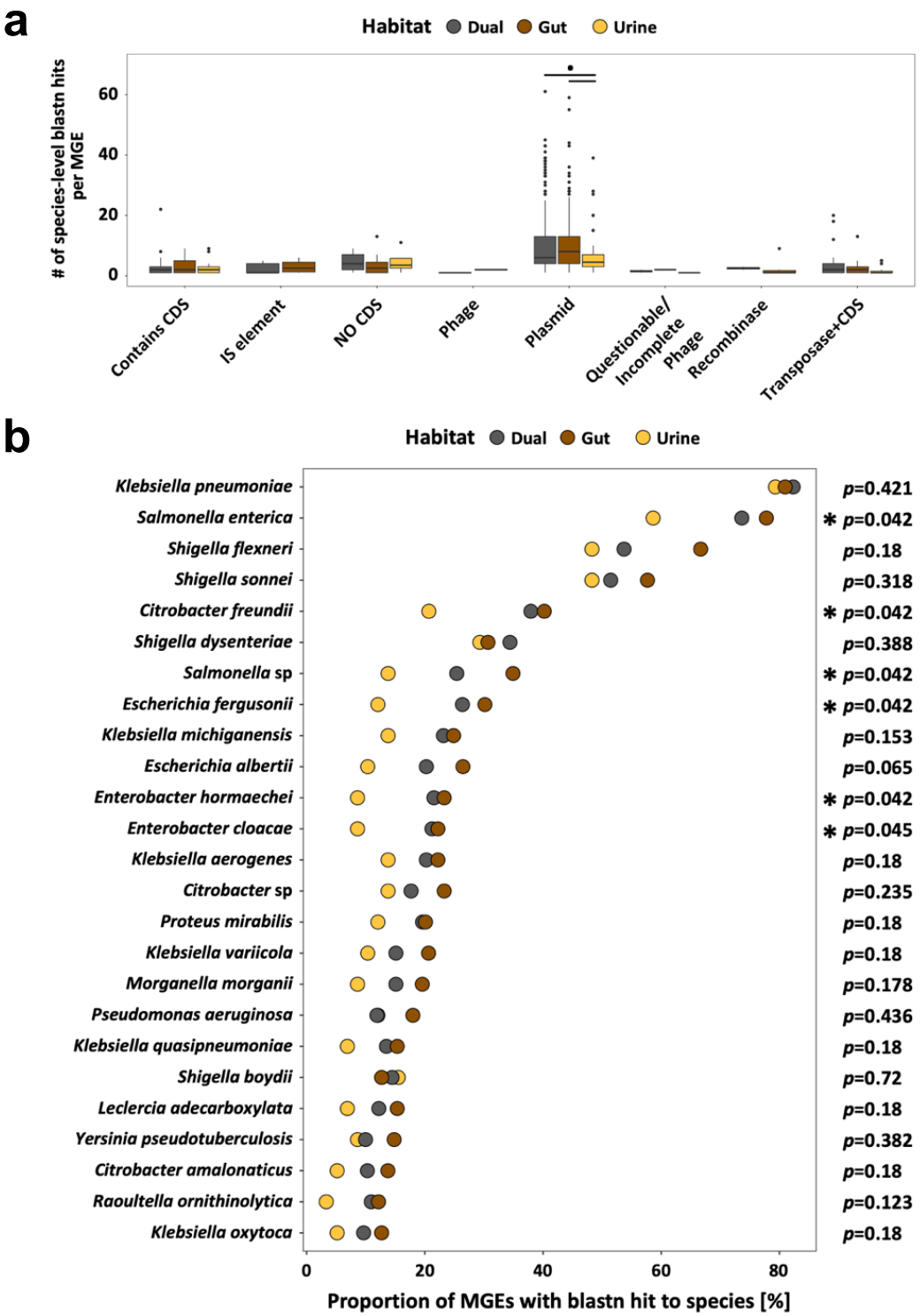

Figure 4.11 Predicted host-range of putative MGEs. (a) UPEC putative plasmidic MGEs are commonly found in other species. Blastn results of putative MGEs classified as plasmidic against the NCBI nucleotide database (>95% identity, >95% query coverage. Urinary plasmidic MGEs were found in significantly less species compared to contigs

present in stool or across habitats (Two-way ANOVA $P \leq 1.57e^{-05}$, Tukey post-hoc $P<0.001$ and $P=0.014$, respectively) **(b)** Percentage of plasmidic MGE sharing between UPEC and the 25 species found to share the most plasmidic contigs with UPEC. *P*-values indicate significance values for the underrepresentation of species in the pool of urinary MGEs compared to the combined stool/dual plasmidic MGE pool as determined using Fisher's exact test. *P*-values are FDR corrected.

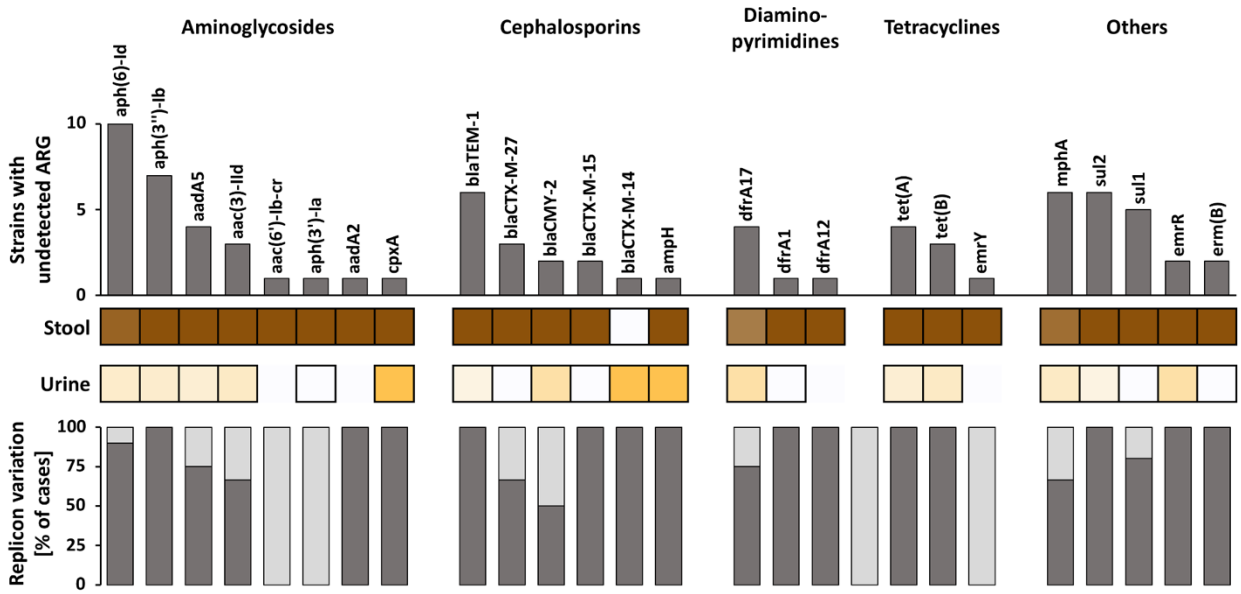

Figure 4.12 Intestinally persistent UPEC are a reservoir for ARGs. (Top) Number of lineages with 'hidden' ARGs grouped by resistance class (see Results). (Middle) Heatmap indicating the percentage of 'hidden' ARG cases where the ARG is found in an asymptomatic isolate recovered from urine (yellow) or stool (brown). (Bottom) Percentage of cases where 'hidden' ARGs are accompanied by variation in the replicon repertoire of the isolate carrying the compared to the DxU isolate.

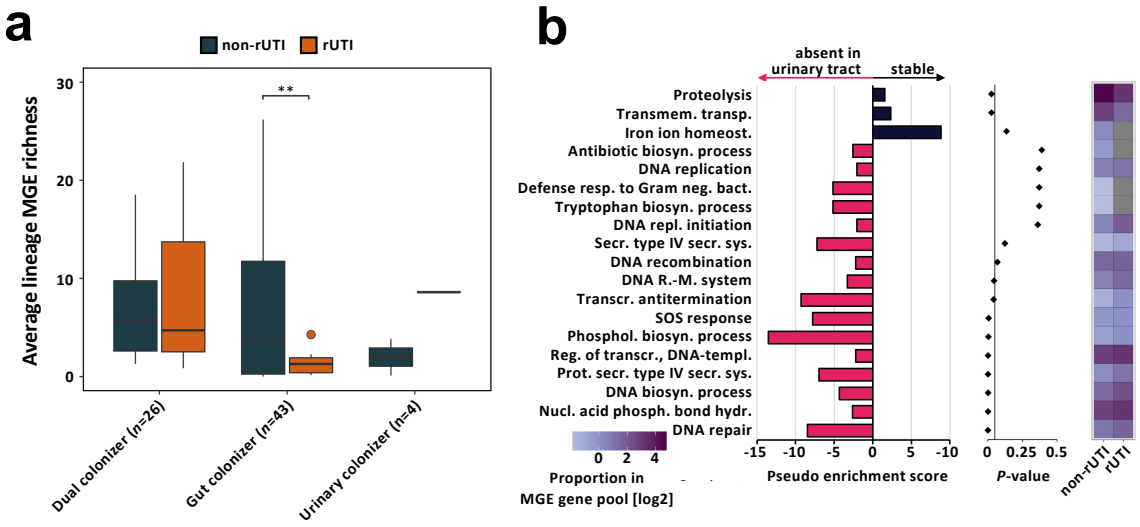

Figure 4.13 Gut colonizing UPEC lineages causing rUTI exhibit decreased MGE richness. (a) MGE richness of lineages causing rUTI during the follow-up period and non-rUTI lineages parsed by colonization type (n=73 lineages, Welch's t-test, FDR corrected gut colonizer P=0.001, dual and urinary colonizer FDR corrected P>0.05). Outliers (outside 1.5x interquartile range) are depicted as points. Whiskers represent 1.5x interquartile range. Upper, middle, and lower box lines indicate 75th, 50th, and 25th percentiles, respectively. **(b)** (Left) Pseudo enrichment score of GO terms in the pool of MGEs absent or stable in urinary isolates of gut colonizing UPEC lineages. Top 19 GO categories by P-value are visualized. Pink bars indicate gene associated GO terms overrepresented in the urine instable MGE pool, black bars indicate GO terms enriched in the pool of MGEs stable in urinary isolates. Pseudo enrichment score was calculated by adding one count to all GO categories. (Middle) P-values for each GO category determined from overrepresentation analysis using hypergeometric distribution. (Right) Proportion of each visualized GO term in the MGE associated gene pool of rUTI and non-rUTI causing lineages of gut colonizing UPEC. Grey tiles indicate absence of a GO term in the MGE gene pool.

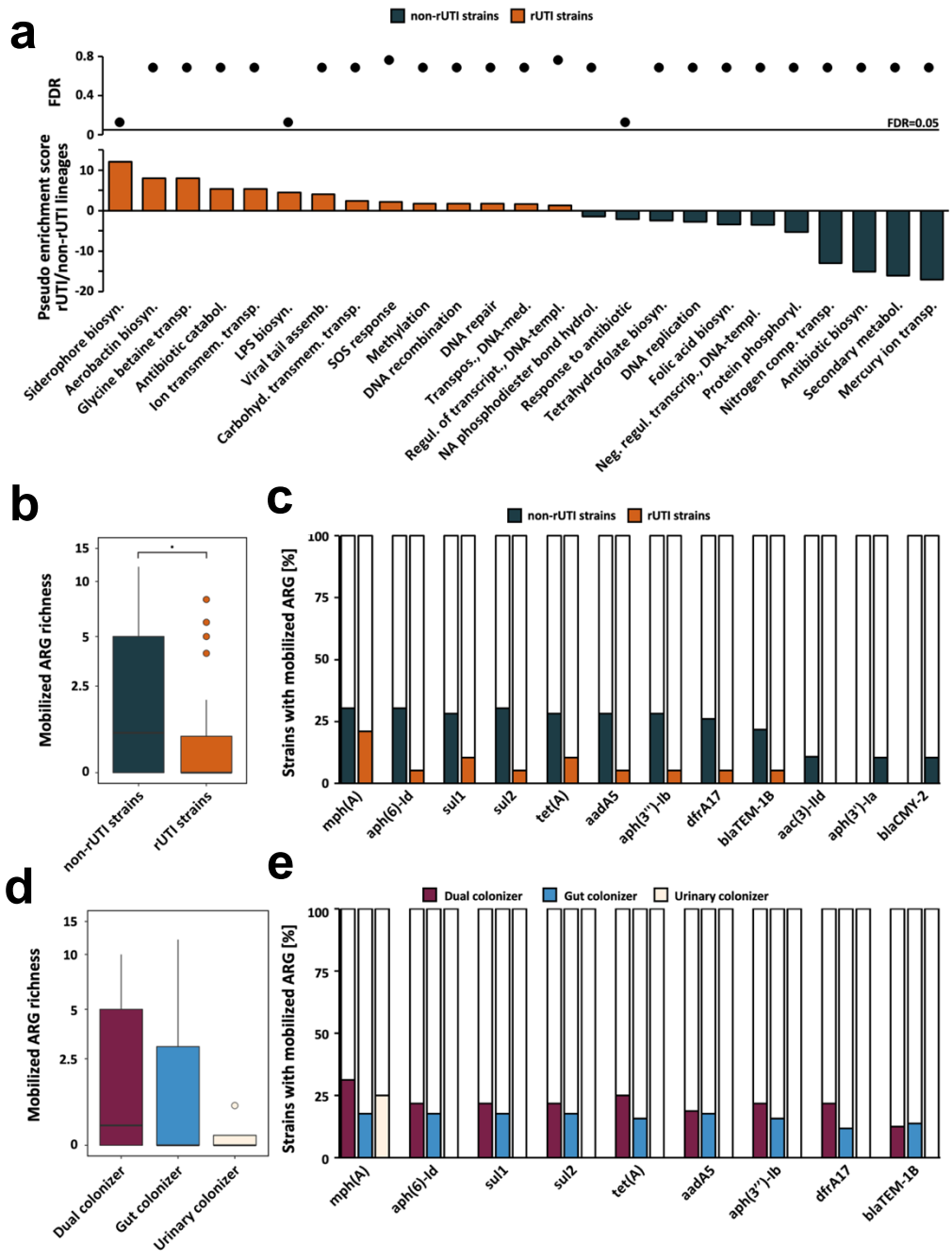

Figure 4.14 Enrichment of MGE GO terms and mobilized ARGs by lineage recurrence status and persistence type. (a) Despite variability no GO terms are over- or underrepresented in the mobilized gene pool of rUTI (orange) and non-rUTI (green) UPEC lineages (n=69 lineages, Fisher's exact test, all FDR corrected P-values >0.05). GO term overrepresentation was assessed using Fisher's exact test. P-values were FDR corrected. Pseudo enrichment scores were calculated comparing observed GO term abundances between compared groups adding the minimal value in the array as a pseudo-count. **(b)** Mobilized ARG richness between rUTI (orange) and non-rUTI (green) lineages (n=69 lineages, Wilcoxon rank-sum test P=0.055). **(c)** Prevalence of specific

mobilized ARGs did not vary significantly between rUTI (orange) and non rUTI lineages (green, n=69 lineages, Fisher's exact test, all FDR corrected P-values >0.05). **(d)** Mobilized ARG richness did not differe significantly between dual colonizers (maroon), gut colonizers (blue) and urinary colonizing lineages (light yellow, n=87 lineages, Kruskal-Wallis P=0.231). **(e)** Prevalence of specific mobilized ARGs did not vary significantly between dual colonizers (maroon), gut colonizers (blue) and urinary colonizing lineages (light yellow, Fisher's exact test, all FDR corrected P-values >0.05).

Table 4.1 UPEC sequence type (ST) distribution

Phylogroup (Prevalence %)	Clonal groups (Prevalence %)	*fimH* type	Dual colonizer (*n*=32)	Gut colonizer (*n*=51)	Urinary colonizer (*n*=4)
A (2.3%)	410 (1.1%)	24	0	1 (1.9%)	0
	744 (1.1%)	54	1 (3.1%)	0	0
B2 (75.9%)	73 (1.1%)	103	0	0	1 (25%)
	95 (1.1%)	27	1 (3.1%)	0	0
	131 (47.1%)	30	14 (43.75%)	23 (45.1%)	0
		41	3 (9.4%)	0	0
		undefined	0	1 (1.9%)	0
	636 (1.1%)	undefined	0	1 (1.9%)	0
	1193 (25.3%)	64	8 (25%)	12 (23.5%)	2 (50%)
C (1.1%)	10 (1.1%)	171	0	1 (1.9%)	0
D (13.8%)	38 (2.3%)	5	1 (3.1%)	0	0
		65	0	1 (1.9%)	0
	69 (3.4%)	27	1 (3.1%)	2 (3.9%)	0
	70 (1.1%)	65	0	0	1 (25%)
	405 (3.4%)	27	0	3 (5.9%)	0
	501 (1.1%)	undefined	1 (3.1%)	0	0
	1177 (1.1%)	65	0	1 (1.9%)	0
	2003 (1.1%)	65	1 (3.1%)	0	0
F (5.7%)	354 (2.3%)	58	0	2 (3.9%)	0
	648 (2.3%)	29	1 (3.1%)	0	0
		undefined	0	1 (1.9%)	0
	6870 (1.1%)	undefined	0	1 (1.9%)	0
Unknown (1.1%)	2006 (1.1%)	61	0	1 (1.9%)	0

Table 4.2 Permutation analysis results (>1 lineages)

Gene	#lineages	SNPs	Indels	*P*-value	Analysis Group
hemX	4	0	4	0	Gut colonizer all isos
angR	4	3	1	0.175	Gut colonizer all isos
neuC	3	2	1	0	Gut colonizer all isos
cadC	3	2	1	0.009	Gut colonizer all isos
sat	3	2	1	0.1	Gut colonizer all isos
gltC	3	2	1	0.134	Gut colonizer all isos
umuC	3	2	2	0.039	Gut colonizer all isos
traI	3	2	1	0.567	Gut colonizer all isos
traD	3	2	1	0.035	Gut colonizer all isos
gspA	3	3	0	0.177	Gut colonizer all isos
nfsA	3	3	0	0.001	Gut colonizer all isos
degS	3	3	0	0.002	Gut colonizer all isos
dmlR	3	3	0	0.239	Gut colonizer all isos
infB	3	3	0	0.044	Gut colonizer all isos
wbbL	2	1	1	0.001	Gut colonizer all isos
mltD	2	1	1	0.033	Gut colonizer all isos
kfoC	2	1	1	0	Gut colonizer all isos
dinG	2	0	2	0.443	Gut colonizer all isos
evgS	2	1	1	0.33	Gut colonizer all isos
lacY	2	1	1	0.03	Gut colonizer all isos
yraI	2	1	1	0.001	Gut colonizer all isos
nupC	2	1	1	0.033	Gut colonizer all isos

rfaL	2	1	1	0.019	Gut colonizer all isos
acrR	2	2	1	0.045	Gut colonizer all isos
ptrA	2	1	1	0.24	Gut colonizer all isos
ecpC	2	1	1	0.16	Gut colonizer all isos
recF	2	2	0	0.286	Gut colonizer all isos
dgoT	2	2	0	0.047	Gut colonizer all isos
topA	2	2	0	0.3	Gut colonizer all isos
nanC	2	2	0	0.025	Gut colonizer all isos
fecA	2	2	0	0.419	Gut colonizer all isos
prfA	2	2	0	0.066	Gut colonizer all isos
sppA	2	2	0	0.469	Gut colonizer all isos
yhdJ	2	2	0	0.099	Gut colonizer all isos
lptB	2	2	0	0.057	Gut colonizer all isos
metE	2	2	0	0.132	Gut colonizer all isos
ompC	2	2	0	0.025	Gut colonizer all isos
barA	2	2	0	0.211	Gut colonizer all isos
rfaY	2	2	0	0.006	Gut colonizer all isos
yahB	2	2	0	0.217	Gut colonizer all isos
ydhC	2	2	0	0.031	Gut colonizer all isos
narX	2	2	0	0.094	Gut colonizer all isos
prs	2	2	0	0.016	Gut colonizer all isos
marA	2	2	0	0.001	Gut colonizer all isos
spoT	2	2	0	0.146	Gut colonizer all isos

gyrA	2	2	0	0.205	Gut colonizer all isos
pka	2	2	0	0.181	Gut colonizer all isos
mprA	2	2	0	0.02	Gut colonizer all isos
kdpB	2	2	0	0.116	Gut colonizer all isos
fabB	2	2	0	0.031	Gut colonizer all isos
crfC	2	2	0	0.438	Gut colonizer all isos
finO	2	2	0	0.003	Gut colonizer all isos
klcA	2	2	0	0.068	Gut colonizer all isos
dtpD	2	2	0	0.051	Gut colonizer all isos
focA	2	2	0	0.03	Gut colonizer all isos
caiT	2	2	0	0.13	Gut colonizer all isos
sucC	2	2	0	0.103	Gut colonizer all isos
typA	2	2	0	0.089	Gut colonizer all isos
gntT	2	2	0	0.043	Gut colonizer all isos
spuE	2	2	0	0.032	Gut colonizer all isos
yqcE	2	2	0	0.031	Gut colonizer all isos
rbsA	2	2	0	0.594	Gut colonizer all isos
wecA	2	2	0	0.021	Gut colonizer all isos
btsT	2	2	0	0.125	Gut colonizer all isos
soxR	2	2	0	0.002	Gut colonizer all isos
arnT	2	2	0	0.081	Gut colonizer all isos
neuC	2	1	1	0	Gut colonizer gut isos
degS	2	2	0	0.001	Gut colonizer gut isos

wbbL	2	1	1	0.002	Gut colonizer gut isos
nanC	2	1	0	0.004	Gut colonizer gut isos
rpoS	2	2	0	0.006	Gut colonizer gut isos
yqcE	2	2	0	0.006	Gut colonizer gut isos
typA	2	2	0	0.01	Gut colonizer gut isos
spuE	2	2	0	0.011	Gut colonizer gut isos
prfA	2	2	0	0.015	Gut colonizer gut isos
dcp	2	2	0	0.015	Gut colonizer gut isos
metE	2	2	0	0.016	Gut colonizer gut isos
sucC	2	1	0	0.022	Gut colonizer gut isos
umuC	2	2	1	0.025	Gut colonizer gut isos
acrR	5	3	2	0	dual colonizer
ompC	4	4	0	0	dual colonizer
wbbL	4	3	1	0	dual colonizer
mprA	4	3	1	0	dual colonizer
nfsA	4	1	3	0	dual colonizer
rfaH	4	3	1	0	dual colonizer
ssb	3	3	0	0.003	dual colonizer
mrdA	3	3	0	0.011	dual colonizer
nfsB	3	2	2	0.001	dual colonizer
marR	3	2	2	0	dual colonizer
comR	3	1	2	0	dual colonizer
ompF	3	1	2	0.002	dual colonizer
barA	3	3	0	0.037	dual colonizer
ydcV	3	3	0	0.031	dual colonizer
rpoA	3	3	0	0.003	dual colonizer
traD	3	1	2	0.009	dual colonizer
acrB	3	3	0	0.353	dual colonizer
ydfR	2	2	0	0.003	dual colonizer
ompW	2	2	0	0.003	dual colonizer
atzC	2	2	0	0.01	dual colonizer
rfbB	2	1	1	0.006	dual colonizer

ampC	2	2	0	0.022	dual colonizer
cadC	2	1	1	0.033	dual colonizer
focC	2	2	0	0.001	dual colonizer
dtpA	2	2	0	0.041	dual colonizer
astC	2	2	0	0.018	dual colonizer
yibH	2	2	0	0.012	dual colonizer
alsT	2	2	0	0.032	dual colonizer
prfC	2	2	0	0.038	dual colonizer
oppA	2	2	0	0.046	dual colonizer
glpD	2	2	0	0.038	dual colonizer
murP	2	2	0	0.023	dual colonizer
moaC	2	2	0	0.001	dual colonizer
kdgT	2	2	0	0.01	dual colonizer
cpxA	2	2	0	0.024	dual colonizer
pepQ	2	2	0	0.019	dual colonizer
fyuA	2	2	0	0.041	dual colonizer
narU	2	1	1	0.03	dual colonizer
frc	2	1	1	0.017	dual colonizer
recG	2	2	0	0.079	dual colonizer
parE	2	2	0	0.065	dual colonizer
ftsI	2	2	0	0.053	dual colonizer
yrfF	2	2	0	0.072	dual colonizer
clpV1	2	2	0	0.137	dual colonizer
poxB	2	2	0	0.052	dual colonizer
aspS	2	2	0	0.059	dual colonizer
fliI	2	2	0	0.097	dual colonizer
evgS	2	2	0	0.251	dual colonizer
rpoC	2	2	0	0.319	dual colonizer
iutA	2	2	0	0.068	dual colonizer
yccS	2	1	1	0.326	dual colonizer
gltC	2	1	1	0.171	dual colonizer
lapB	2	2	0	0.368	dual colonizer
sppA	2	2	0	0.315	dual colonizer
traI	2	2	0	0.53	dual colonizer
klcA	2	2	0	0.056	dual colonizer
cirA	2	2	0	0.663	dual colonizer
srlR	2	2	0	0.389	dual colonizer
codB	2	2	0	0.072	dual colonizer
lpxL	2	2	0	0	urinary colonizer

Table 4.3 Reference *E. coli* genomes

Species	Strain	GenBank assembly
Escherichia coli	101-1	GCA_000168095.1
Escherichia coli	11128	GCA_000010765.1
Escherichia coli	12009	GCA_000010745.1
Escherichia coli	APEC O78	GCA_000332755.1
Escherichia coli	B7A	GCA_000725265.1
Escherichia coli	DH1	GCA_000023365
Escherichia coli	DH10B	GCA_006352235.1
Escherichia coli	E110019	GCA_000167875.1
Escherichia coli	E22	GCA_000167855.1
Escherichia coli	EC4115	GCA_000021125.1
Escherichia coli	ED1A	GCA_000026305.1
Escherichia coli	EDL933	GCA_000732965.1
Escherichia coli	F11	GCA_000167835.1
Escherichia coli	IAI1	GCA_000026265.1
Escherichia coli	IAI39	GCA_000026345.1
Escherichia coli	K-12 substr. MG1655	GCA_000005845.2
Escherichia coli	REL606	GCA_000017985.1
Escherichia coli	S88	GCA_000026285.2
Escherichia coli	SE11	GCA_000010385.1
Escherichia coli	SE15	GCA_000010485.1
Escherichia coli	UMN026	GCA_000026325.2
Escherichia coli	UMNK88	GCA_000212715.2
Escherichia coli	UTI89	GCA_000013265.1
Escherichia coli	W	GCA_000184185.1

Chapter 5

Clinical Risk Factors and Gut Microbiome Correlates of Recurrent Urinary Tract Infection

5.1 Abstract

The cycle of antimicrobial treatment and recurrent UTI (rUTI) is thought to be facilitated by the gut reservoir of uropathogenic *Escherichia coli* (UPEC). 125 participants with UTI were enrolled in a longitudinal, multi-center cohort study investigating the gut microbiome and clinical risk factors for recurrence. 644 stool samples and 895 UPEC isolates were interrogated for taxonomic composition, antimicrobial resistance genes, and phenotypic resistance. Antimicrobial treatment in the 6 months prior to UTI was associated with elevated risk of recurrence, while more than 7 days of antimicrobial treatment, antimicrobials after index UTI, and treatment with trimethoprim (TMP) and/or sulfamethoxazole (SMX) were associated with reduced risk. The UTI microbiome was distinct from healthy reference microbiomes in both taxonomic composition and antimicrobial resistance gene (ARG) burden. rUTI and non-rUTI samples in the cohort did not significantly differ, but gut microbiomes from urinary tract colonized participants were elevated in *E. coli* abundance at post-antimicrobial days 7 and 14. Corresponding UPEC gut isolates from urinary tract colonizing lineages showed increased phenotypic resistance against 11 of 23 tested drugs compared to non-colonizers. These findings demonstrate that UPEC can asymptomatically colonize the gut and urinary tract, and post-antimicrobial blooms of gut *E. coli* among urinary tract colonized participants suggest that cross-habitat migration of UPEC is an important mechanism of rUTI. Treatment timing and asymptomatic colonization should be considered in treating rUTI and developing novel therapeutics.

5.2 Introduction

Urinary tract infections (UTIs) are estimated to affect 250 million people worldwide each year[1]. In the United States (US) alone, 13.7% of men and 60% of women experience a UTI in their lifetime[2,3], and 24% of women with UTI experience a recurrent UTI (rUTI) within 6 months of the initial episode[4]. As UTIs are typically treated with antimicrobials, the cycle of treatment and recurrence is fertile ground for selection of antimicrobial resistance (AR)[5]. Uropathogenic *Escherichia coli* (UPEC) are the most common causative agents of UTI[6], and comparative genomic analyses of UPEC have established that the cycle of recurrence is fueled by at least three independent pathways: urinary persistence, reinfection from external sources, and gastrointestinal colonization[7–9]. The gut in particular is a known reservoir for UPEC, from which multiple episodes of UTI can be seeded[7,10,11].

In healthy individuals, commensal microbiota populating the gut can provide colonization resistance against pathogenic Enterobacterales through competitive exclusion or by modulating host immunity[12]. A disrupted gut microbiome state has been implicated in a number of chronic and recurrent conditions, including *Clostridioides difficile* infection (CDI)[13] and inflammatory bowel disease (IBD)[14]. Similarly, the history of repeated antimicrobial exposures in rUTI may render patients more susceptible to colonization with UPEC[11]. One recent study comparing the gut microbiomes of 15 women with a history of rUTI and 16 healthy controls reported depleted richness in the gut microbiome in women with rUTI, including depleted richness and reduced abundance of butyrate producers[15]. However, our understanding of UPEC's role in the

gut microbiome and which factors drive some UTI patients towards recurrence is incomplete. The purpose of this 125-patient, multicenter, prospective cohort study was to determine clinical risk factors for recurrence among patients with antimicrobial resistant UTI, and to investigate the relationship between urinary tract colonization, gut microbiota, and rUTI.

5.3 Results

5.3.1 Cohort description

A total of 125 patients were enrolled in the study from the four participating sites (Table 5.1, Figure 5.1). Forty-seven (37.6%) patients experienced rUTI within 6 months. 12/38 (31.6%) patients who continued in the study after their first recurrence experienced a second recurrence, and 7/12 (58.3%) of those who continued in the study after their second recurrence experienced a third recurrence. Most patients were female (93.6%) with a median age of 58 years (interquartile range 42–71). 92.8% of first UTI episodes were caused by *E. coli.* The most common symptoms of UTI episodes were pain or burning during urination and cloudy urine (>40% of patients experienced each of these symptoms; Table 5.2). The most common antimicrobials used to treat UTI episodes were nitrofurantoin (44.6%) and cephalosporin or a penicillin (30.3%).

5.3.2 Treatment History and Urinary Tract Colonization are Associated with Risk of Recurrence

Clinical variables associated with increased risk of recurrence included steroid use in the 6 months prior to or at enrollment, antimicrobial use in the 6 months prior to or at enrollment, and UTI episode treatment with a cephalosporin or a penicillin (Table 5.3; Table 5.4). Variables associated with decreased risk of recurrence included antimicrobial use (other than for UTI treatment) during the UTI episode prior to censor date (HR = 0.33; 95% CI 0.12 – 0.90) and UTI episode treatment with trimethoprim-sulfamethoxazole (TMP-SMX; HR = 0.36; 95% CI 0.14 – 0.90). Intestinal colonization with the UTI episode-causing organism was not a significant risk factor for recurrence, but asymptomatic urinary tract colonization by UPEC was of borderline significance (HR=1.58; 95% CI 0.94 – 2.66).

Clinical risk factors and protective factors independently associated with rUTI are shown in Table 5.3. History of prior antimicrobial use (HR = 2.20; 95% CI 1.03 – 4.70) and steroid use (HR = 2.35; 95% CI 1.28 – 4.31) were associated with increased risk of recurrence. Variables associated with decreased risk of recurrence included duration of UTI episode antimicrobial treatment >7 days (HR = 0.55; 95% CI 0.32 – 0.95), antimicrobial use (other than for UTI treatment) during the UTI episode prior to censor date (HR = 0.32; 95% CI 0.12 – 0.90), and UTI treatment with TMP-SMX (HR = 0.39; 95% CI 0.15 – 0.95).

5.3.3 The gut microbiome in UTI patients is distinct from that of healthy individuals

To characterize the gut microbiome, 644 stool samples from 106 patients with available stool were sequenced. Forty-three (40.6%) of these patients experienced 45 episodes of

rUTI during the study period, and 63 did not (59.4%; non-rUTI). In total, 331 rUTI samples and 313 non-rUTI stool samples were subject to whole metagenome sequencing. The enrollment samples from this cohort (E1-S1) were grouped together with 15 published rUTI samples from the UMB study (See Methods) as "UTI". Microbiome samples from healthy adults (20 HH, 16 UMB) were included as a "Healthy" comparison group (Table 5.5).

Species richness was lower among UTI samples compared to healthy controls, though not reaching significance (Kruskal-Wallis, P=0.055 Figure 5.2a). Pairwise microbiome dissimilarity (Bray-Curtis) was measured, and even after accounting for differences among studies (PERMANOVA, P=0.001, Figure 5.2b), there were significant differences in species-level microbiota composition between UTI and healthy samples (PERMANOVA, P=0.043, Figure 5.2c).

Using linear mixed-effect models (MaAsLin2)[16], 11 differentially abundant intestinal taxa were identified at the genus level (False Discovery Rate; FDR< 0.25) between UTI samples and healthy controls, of which 9 were depleted in UTI samples (Figure 5.2d). Genera depleted in UTI samples included *Parasutterella*, *Akkermansia*, and *Bilophila*. The healthy samples were enriched in commensal Firmicutes *Ruminococcus*, *Roseburia*, and *Eubacterium*.

We hypothesized the UTI gut microbiome may be enriched for antimicrobial resistance genes (ARGs) compared to the healthy microbiome, due to a history of UTI treatment-related antimicrobial exposure. The abundance of identified ARGs (as measured in units of Reads Per Kilobase of reference sequence per Million sample

reads; RPKM) was significantly higher among UTI samples (Kruskal-Wallis, *P*=0.002, Figure 5.2e), but not their richness (Kruskal-Wallis, *P*=0.09, Figure 5.2f) or diversity (Kruskal-Wallis, *P*=0.53).

5.3.4 The gut microbiomes of patients with rUTI and those without (non-rUTI) are similar

The gut microbiomes of all 480 samples from each patient's first UTI episode were compared (including S1) to query differences between the rUTI and non-rUTI microbiome. Neither richness (Kruskal-Wallis, *P*=0.37) nor Shannon diversity (Kruskal-Wallis, *P*=0.24, Figure 5.3a-b) differed between groups. Patient ID was the greatest source of microbiome variation (PERMANOVA, *P*=0.001), but not rUTI status (*P*>0.05, Figure 5.3c). When the analysis was repeated with just one representative taxonomic profile per patient (average relative abundance of each species across all samples per patient), rUTI status was again not a significant variable explaining microbiome composition (*P*=0.35, Figure 5.3d).

5.3.5 Urinary tract colonized patients have increased gut **E.coli** *at 7 to 14 days post-antimicrobials*

Gut microbiome species richness was significantly depleted during and after antibiotic therapy (enrollment, day 3), but increased significantly by days 7 to 14 post-antimicrobial treatment (Wilcoxon, BH-adjusted *P* <0.05, Figure 5.4a). Moreover, antimicrobials differentially impacted microbiome richness at earlier timepoints

(Ertapenem and Amoxicillin/Clavulanic acid with lowest richness, Kruskal-Wallis, Dunn post-hoc BH-adjusted P <0.05, Figure 5.4b), but these differences were non-significant by days 7 to 14 (Kruskal-Wallis, BH-adjusted P < 0.05, Figure 5.4b). This observation prompted us to investigate the microbiome at specific timepoints. Urinary tract colonized patients (as defined in the Methods) had distinct gut microbiomes from non-urinary tract colonized patients at days 7 to 14 post-antimicrobials (PERMANOVA P <0.05, Figure 5.5a), even after adjusting for age and UTI treatment antimicrobial type. The gut microbiome at no other timepoint differed significantly in taxonomic structure by recurrence, urinary tract colonization, or gut colonization.

E. coli and *Paraprevotella xylaniphila* were the only two intestinal taxa significantly enriched in urinary tract colonized patients (MaAsLin2 FDR<0.25, Figure 5.5b-c). These cohort-level observations were also quantifiable at the individual scale: Patient WU-16 exhibited a 44-fold increase of intestinal *E. coli* from day 3 to day 7, and a 6-fold increase from day 7 to day 14 (Figure 5.5d).

Among the urinary tract colonized patients, 54.5% (18/33) experienced rUTI during the follow-up period. These patients exhibited depleted gut *Bacteroides xylanisolvens* abundance compared to non-rUTI patients, and this was the singular distinguishing taxon observed (MaAsLin2 FDR<0.25, Figure 5.5e).

5.3.6 Intestinal **E. coli** *from urinary tract colonized individuals exhibit heightened phenotypic resistance*

Gut E. coli from urinary tract colonizing lineages were enriched in resistance against 11/23 drugs: ceftriaxone, ceftazidime, cefotetan, cefazolin, ampicillin, TMP-SMX, ampicillin-sulbactam, ciprofloxacin, levofloxacin, aztreonam, and nitrofurantoin (Fisher's exact test, BH-adjusted $P < 0.05$, Table 5.6, Figure 5.5f). Non-urinary tract colonizing lineages were enriched in resistance against meropenem and imipenem (BH-adjusted $P < 0.05$). Gut E.coli from urinary tract colonizing lineages were elevated in overall AST score (Kruskal-Wallis, $P < 0.001$, Figure 5.5g). Corresponding urinary isolates from urinary tract colonizing lineages were not significantly elevated in AST score (Kruskal-Wallis, P=0.13, Figure 5.5h).

5.4 Discussion

We enrolled a prospective cohort of 125 patients with UTI to investigate both clinical and metagenomic risk factors for recurrence. Antimicrobial use in the prior 6 months was associated with elevated risk of recurrence and may be a correlate of a disrupted microbiome state, increasing the risk for opportunistic infection. Use of steroids was also associated with increased risk, potentially due to their immunosuppressive effect[17]. TMP-SMX was associated with decreased risk of recurrence, though phenotypic resistance to TMP-SMX was elevated among gut isolates from urinary tract colonizing lineages. Together these findings suggest that although TMP-SMX is generally efficacious, resistance is still selected for among persistent lineages. Antimicrobial use after UTI start time, and more than 7 days of antimicrobial treatment were associated with reduced risk of recurrence, potentially reflecting the effects of continued control or

eradication of otherwise persisting UPEC populations in the urinary or gastrointestinal tract. Further investigation of clinical risk factors for rUTI is needed to independently replicate these findings in larger cohorts.

We utilized metagenomics to investigate the gut-bladder axis. Here we show that the gut microbiome in people with UTI is distinct from that of healthy individuals, reaffirming the role of gut microbiome dysbiosis in UTI[11,15,18]. In particular, the genera *Parasutterella*, *Akkermansia,* and *Bilophila* were depleted in intestinal samples of subjects with UTI in our cohort, consistent with previous findings[15]. However, when we compared UTI patients in our cohort with recurrence during the study period and those without, we found no significant gut microbiome differences. Instead, our findings point to asymptomatic colonization of the urinary tract as a significant distinguishing factor among gut microbiomes. Patients with urinary tract colonization displayed elevated gut *E. coli* abundance at post-antimicrobial, asymptomatic timepoints. This finding of *E. coli* blooms in the gut has been previously observed[19], though importantly, the previous study utilized culture-based quantification while our metagenomic observations are limited in sub-species taxonomic resolution. Further subsetting the urinary colonized group into recurrence and non-recurrence samples found *B. xyalnisolvens* to be the singular taxon significantly elevated in the non-recurrent group, indicating the lack of broad taxonomic differences. Nevertheless, *Bacteroides* are commensals whose member species are under active investigation for probiotic development[20]. Their elevated presence may reflect a protective effect via competition in the gut microbiome[21], despite urinary tract colonization by UPEC.

Urinary tract colonization was associated with elevated phenotypic resistance among gut isolates, but not urinary isolates. This finding underlines the gut microbiome's role in selection for specific resistance types during UTI, as reflected in elevated ARG abundance, but not Shannon index, compared to healthy controls. A previous study of this cohort demonstrated the presence of 'hidden' ARGs among UPEC lineages which appeared after the diagnostic isolate, likely gained through mobile genetic elements enriched in the gut microbiome[18]. While urinary isolates belonging to the same lineage as the causative pathogen do not appear to maintain high resistance profiles during asymptomatic colonization[22], it is plausible for a highly resistant gut isolate to migrate and cause recurrence in the urinary tract. Further research is needed to elucidate the migratory dynamics of UPEC in the host.

Together our findings suggest that antimicrobial treatment type, history, and duration are associated with differential risk of rUTI. The gut-bladder axis plays an important role in rUTI, but not all patients follow the same patterns of asymptomatic colonization. Altogether these patient and case-specific characteristics should be considered to effectively combat rUTI.

5.5 Methods

5.5.1 Study population

Participants for this prospective, multi-center cohort study were recruited between July 2016 and May 2019 among patients with positive urine cultures at Barnes-Jewish Hospital/Washington University in St. Louis (WU), St. Louis, Missouri, Duke

University Hospital (DK), Durham, North Carolina, the Hospital of the University of Pennsylvania (PN), Philadelphia, Pennsylvania, and Rush University Medical Center (RH), Chicago, Illinois. This study was approved by the Washington University Human Research Protection Office as the single IRB. Local IRB approvals were obtained as necessary.

5.5.2 Inclusion/exclusion criteria

Patients with a symptomatic UTI diagnosed and treated by a physician and a urine culture that yielded Enterobacterales with one of the following resistances were included in the current analysis: (1) resistance to ciprofloxacin or levofloxacin, (2) resistance to any third generation cephalosporin, (3) resistance to ertapenem and susceptible to meropenem, imipenem, and/or doripenem, (4) resistance to >2 of the following antimicrobial classes: carbapenems, aminoglycosides, fluoroquinolones, fourth generation cephalosporins, piperacillin/tazobactam, or (5) identification of any of the following resistance mechanisms: ESBL, CRE, KPC, NDM-1, OXA-48, IMP, IMP-1, or VIM.

Patients were excluded if they had any of the following conditions: >1 organism in their urine, recurrent CDI, intra-abdominal devices, absolute neutrophil count [ANC] <500mm^3, intestinal mucosal disruption, unlikely to survive 6 months, pregnancy or unwilling/unable to use contraception, short gut syndrome, intestinal motility medication use, irritable bowel disease, recent abdominal surgery, active typhlitis or diverticulitis, current gastrointestinal graft-versus-host disease, HIV without

antiretroviral therapy, CD4 <200mm^3, peritoneal dialysis, cirrhosis with ascites, active intra-abdominal malignancy, chronic indwelling foley or suprapubic catheter, chronic ileal conduit, active hepatitis B or C, ureteral stent, or active kidney stone.

5.5.3 Enrollment

Eligible patients were contacted by study personnel by phone (if outpatient) or in person (if hospitalized) to verify that all inclusion/exclusion criteria were met. Written, informed consent was obtained from all patients. Once a patient was enrolled, study personnel interviewed the patient regarding their UTI symptoms, UTI antimicrobial treatment, and medical history. If available, study personnel also collected remnant urine from the patient's diagnostic urine culture from the clinical microbiology laboratory.

5.5.4 Episode and outcome definitions

The first UTI episode per patient was defined as starting on the date of study enrollment. UTI recurrence (rUTI) was defined as the diagnosis of a subsequent symptomatic UTI that required antimicrobial treatment during the six-month follow-up period with any uropathogen. All UTI diagnosis and treatment decisions were made by the patient's primary treatment provider. The recurrence date was assigned as the date of first symptom onset if known; otherwise, the antimicrobial treatment start date was used. If a patient continued in the study, the recurrence date served as both the end of follow-up for the episode and the start date for a new UTI episode. From episode 1

enrollment, a patient could continue in the study for to up to three total UTI episodes; patients with a fourth UTI were censored at that time. Patients who did not develop a rUTI were followed for up to 6 months.

5.5.5 Specimen and data collection

Patients submitted stool and urine specimens to the study team at enrollment (Sample 1), the end of UTI antimicrobial treatment (S2), and days 3 (S3), 7 (S4), 14 (S5), 30 (S6), 60 (S7), 90 (S8), 120 (S9), 150 (S10), and 180 (S11) post-antimicrobial treatment. If a patient had a recurrence and chose to continue in the study, the stool and urine specimen collection schedule restarted as a new episode (E1, E2, E3).

At each collection point, patients were provided with supplies for collecting their stool and urine, along with questionnaires about UTI symptoms, medications received, and changes in medical history. Stool/urine specimens and questionnaires were shipped to the study team by courier. Upon arrival in the laboratory, samples were immediately processed fir microbiologic culture or frozen at -80^{o}C. Stool and urine samples collected at sampling points S1, S2, S4, S6, and S11 were selectively cultured to assess uropathogen persistence. If a patient did not submit a specimen at a sampling point, the sample collected at the next closest time point was selected for analysis.

5.5.6 Selective culture

Approximately 1g of stool samples collected at enrollment and on days 0, 7, 30, and 180 post-antimicrobial treatment (pAT) were supplemented with an equal amount

(wt/vol) of phosphate-buffered saline (PBS) and vortexed to homogenize the samples. Ten 10-fold serial dilutions were prepared in PBS, and 10 µL of each of the first 10 dilutions was streaked onto selective agar (Hardy Diagnostics, Santa Maria, CA, USA) specific to each patient's identified ARO using a 10 µL calibrated loop. MacConkey (MAC) agar supplemented with ciprofloxacin (10 g/ml) was used for ciprofloxacin-resistant Enterobacteriaceae, while ESBL-producing Enterobacteriaceae were cultured on Hardy Diagnostics ESBL agar and MAC agar supplemented with cefotaxime (1 g/ml). Isolate species was confirmed using MALDI-TOF MS (VITEK MS, bioMérieux, Durham, NC, USA). Single colonies were diluted in TSB/glycerol and stored at -80°C for later analysis.

5.5.7 Antimicrobial susceptibility testing

Antimicrobial susceptibility testing (AST) of pathogens was performed on Mueller Hinton agar (Hardy Diagnostics, Santa Maria, CA, USA) using Kirby Bauer disk diffusion with antimicrobial disks purchased from Hardy Diagnostics (Santa Maria, CA, USA) and Becton Dickinson (Franklin Lakes, NJ, USA). Results were interpreted according to Clinical & Laboratory Standards Institute guidelines[23]. Fisher's exact tests were conducted to compare AST results between urinary tract colonizing and non-colonizing lineages of *E. coli*, where intermediate isolates were grouped together with susceptible isolates as 'non-resistant'. To calculate AST scores, the AST data were converted into a numeric matrix (0: susceptible, 0.5: intermediate, 1: resistant) and summed for each isolate.

5.5.8 UPEC Colonization

UPEC colonization definitions were retained from an earlier publication from this cohort[18]. Briefly, UTI episodes were categorized as colonized by UPEC if (1) the same *E. coli* lineage was recovered from a specimen type (stool/urine) at >1 asymptomatic sample, or (2) if all isolates recovered from a specimen type (stool/urine) from a UTI episode belonged to the same *E. coli* lineage. Ultimately, colonization for a UTI episode was dichotomized for analysis to represent urinary tract and gastrointestinal colonization any time during the follow-up period before the next recurrence or censor date. Colonization status was re-set at the start of any subsequent UTI episodes.

5.5.9 Statistical Analysis

We used univariate and multivariable Prentice, Williams, and Peterson (PWP) total time models—a conditional model extension of the Cox proportional hazards model that models the full time course of recurrent events—to examine risk factors for rUTI[24,25]. Potential clinical risk factors for rUTI were collected from baseline and post-UTI follow-up questionnaires. The proportional hazards assumption was assessed and confirmed for all potential variables via visualization of negative log of estimated survivor functions plots for each covariate and modeling time-dependent covariates using interaction terms. For multivariable models, backward selection was used with $P \leq 0.1$ as the cutoff for inclusion among variables with $P \leq 0.2$ in univariate analysis. Data management was performed using REDCap and SPSS v27 (IBM Corp., Armonk, NY),

and statistical analysis was performed using SAS version 9.4 (SAS Institute Inc., Cary, NC).

5.5.10 DNA extraction, sequencing and quality filtering

Metagenomic DNA for stool microbiome profiling was extracted from ~100mg of frozen stool using the DNeasy PowerSoil kit (Qiagen, Germantown, MD, USA). Sequencing libraries from fecal metagenomic DNA were prepared using the Nextera kit (Illumina, San Diego, CA, USA). Libraries were pooled and sequenced (2 x150 bp) to a depth of ~5 million reads (fecal metagenomes) on the NextSeq 500 HighOutput platform (Illumina, San Diego, CA, USA). The resulting reads were trimmed of adapters using Trimmomatic v.36 (parameters: LEADING:10 TRAILING:10 SLIDINGWINDOW:4:15 MINLEN:60) and depleted of human read contamination using DeconSeq v.4.3 (default parameters)[26,27].

5.5.11 Microbiome analysis

To assess differences in gut microbiota between participants with a history of UTI compared to a healthy population ("cross-cohort comparisons"), we downloaded two publicly-available metagenomic datasets from recent studies in the US: microbiomes from 20 healthy adults (PRJNA664754; "HH") as well as 31 microbiomes from a rUTI study ("UMB") comprising 15 rUTI (>2 episodes of UTI in past 12 months) and 16 healthy participants (<2 UTIs in lifetime; PRJNA400628; Table 5.5). The first available metagenomic stool sample from every individual was used. Both datasets

featured sequencing depth >2.5 million reads per sample, and the HH cohort utilized identical metagenomic DNA extraction and sequencing techniques as this study.

Paired-end metagenomic reads from all cohorts were used to access sample-specific microbial taxa relative abundance using MetaPhlAn3 v.3.1.0 (default parameters)[28]. Average taxonomic profiles for each patient were also generated by averaging the relative abundances of each taxon at the species level. This process was repeated to generate average taxonomic profiles per patient at specific timepoints, by averaging species abundance from samples corresponding to the relevant timepoints. Taxa were filtered for 10% prevalence prior to each analysis. Resistance gene abundance was determined using ShortBRED v.0.9.4[29] using marker sequences built on the CARD and NCBI AMR databases.

Statistical analysis and visualization of gut microbiome data from all cohorts were conducted in R v.3.6.3[30]. α- and β-microbiota diversity were calculated using vegan v2.5.7[25]. Repeat measures permutational analysis of variance (PERMANOVA) was implemented using the adonis function. Patient ID was included as a mandatory blocking factor in all repeat measure PERMANOVA. In cross-cohort comparisons, a unique study ID was assigned per cohort and included as the first PERMANOVA term. For within-cohort comparisons, age (18-64; 65-79; ≥80) and UTI treatment antimicrobial were included as categorical variables. Linear mixed-effects models (LMEs) were implemented using the MaAsLin2 package via arcsine square root transformation[16]. LMEs included study ID as a random effect in cross-cohort comparisons, and age and treatment drug as categorical random effects in within-cohort comparisons. The

phyloseq[31] package was used to calculate pairwise Bray-Curtis distance between samples and conduct ordination via principal coordinates analysis (PCoA) and canonical analysis of principal coordinates (CAP). Visualizations were created using ggplot2[25] and ggpubr[33].

5.6 Data sharing

Raw reads generated from this study are available on NCBI SRA under PRJNA682246.

5.7 References

1. Ronald AR, Nicolle LE, Stamm E, et al. Urinary tract infection in adults: research priorities and strategies. *International Journal of Antimicrobial Agents*. 2001;17(4):343-348. doi:10.1016/S0924-8579(01)00303-X

2. Griebling TL. Urologic diseases in america project: trends in resource use for urinary tract infections in men. *Journal of Urology*. 2005;173(4):1288-1294. doi:10.1097/01.ju.0000155595.98120.8e

3. Foxman B. Urinary tract infection syndromes. Occurrence, recurrence, bacteriology, risk factors, and disease burden. *Infectious Disease Clinics of North America*. 2014;28(1):1-13. doi:10.1016/j.idc.2013.09.003

4. Foxman B, Gillespie B, Koopman J, et al. Risk Factors for Second Urinary Tract Infection among College Women. *American Journal of Epidemiology*. 2000;151(12):1194-1205. doi:10.1093/oxfordjournals.aje.a010170

5. Klein RD, Hultgren SJ. Urinary tract infections: microbial pathogenesis, host-pathogen interactions and new treatment strategies. *Nat Rev Microbiol*. 2020;18(4):211-226. doi:10.1038/s41579-020-0324-0

6. Flores-Mireles AL, Walker JN, Caparon M, Hultgren SJ. Urinary tract infections: epidemiology, mechanisms of infection and treatment options. *Nature Reviews Microbiology*. 2015;13(5):269-284. doi:10.1038/nrmicro3432

7. Yamamoto S, Tsukamoto T, Terai A, Kurazono H, Takeda Y, Yoshida O. Genetic Evidence Supporting the Fecal-Perineal-Urethral Hypothesis in Cystitis Caused by Escherichia Coli. *Journal of Urology*. 1997;157(3):1127-1129. doi:10.1016/S0022-5347(01)65154-1

8. Chen SL, Wu M, Henderson JP, et al. Genomic diversity and fitness of E. coli strains recovered from the intestinal and urinary tracts of women with recurrent urinary tract infection. *Science translational medicine*. 2013;5(184):184ra60. doi:10.1126/scitranslmed.3005497

9. Thänert R, Reske KA, Hink T, et al. Comparative Genomics of Antibiotic-Resistant Uropathogens Implicates Three Routes for Recurrence of Urinary Tract Infections. Parkhill J, ed. *mBio*. 2019;10(4). doi:10.1128/mBio.01977-19

10. Nielsen KL, Dynesen P, Larsen P, Frimodt-Møller N 2014. Faecal Escherichia coli from patients with E. coli urinary tract infection and healthy controls who have never had a urinary tract infection. *Journal of Medical Microbiology*. 63(4):582-589. doi:10.1099/jmm.0.068783-0

11. Worby CJ, Olson BS, Dodson KW, Earl AM, Hultgren SJ. Establishing the role of the gut microbiota in susceptibility to recurrent urinary tract infections. *J Clin Invest*. 2022;132(5). doi:10.1172/JCI158497

12. Buffie CG, Pamer EG. Microbiota-mediated colonization resistance against intestinal pathogens. *Nature Reviews Immunology*. 2013;13(11):790-801. doi:10.1038/nri3535

13. Zhang L, Dong D, Jiang C, Li Z, Wang X, Peng Y. Insight into alteration of gut microbiota in Clostridium difficile infection and asymptomatic C. difficile colonization. *Anaerobe*. 2015;34:1-7. doi:10.1016/j.anaerobe.2015.03.008

14. Halfvarson J, Brislawn CJ, Lamendella R, et al. Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat Microbiol*. 2017;2(5):1-7. doi:10.1038/nmicrobiol.2017.4

15. Worby CJ, Schreiber HL, Straub TJ, et al. Longitudinal multi-omics analyses link gut microbiome dysbiosis with recurrent urinary tract infections in women. *Nat Microbiol*. 2022;7(5):630-639. doi:10.1038/s41564-022-01107-x

16. Mallick H, Rahnavard A, McIver LJ, et al. Multivariable association discovery in population-scale meta-omics studies. *PLOS Computational Biology*. 2021;17(11):e1009442. doi:10.1371/journal.pcbi.1009442

17. Steiner RW, Awdishu L. Steroids in kidney transplant patients. *Semin Immunopathol*. 2011;33(2):157-167. doi:10.1007/s00281-011-0259-7

18. Thänert R, Choi J, Reske KA, et al. Persisting uropathogenic Escherichia coli lineages show signatures of niche-specific within-host adaptation mediated by mobile genetic elements. *Cell Host & Microbe*. Published online May 10, 2022. doi:10.1016/j.chom.2022.04.008

19. Thanert R, Reske K, Hink T, et al. Comparative Genomics and Clonal Tracking of Multi-drug-Resistant Uropathogens Implicates the Fecal Microbiome as a Potential Reservoir for Recurrent Urinary Tract Infections. *Open Forum Infect Dis*. 2019;6(Suppl 2):S70-S71. doi:10.1093/ofid/ofz359.152

20. Tan H, Zhai Q, Chen W. Investigations of Bacteroides spp. towards next-generation probiotics. *Food Research International*. 2019;116:637-644. doi:10.1016/j.foodres.2018.08.088

21. Deng H, Yang S, Zhang Y, et al. Bacteroides fragilis Prevents Clostridium difficile Infection in a Mouse Model by Restoring Gut Barrier and Microbiome Regulation. *Frontiers in*

Microbiology. 2018;9. Accessed January 27, 2023. https://www.frontiersin.org/articles/10.3389/fmicb.2018.02976

22. Stracy M, Snitser O, Yelin I, et al. Minimizing treatment-induced emergence of antibiotic resistance in bacterial infections. *Science*. 2022;375(6583):889-894. doi:10.1126/science.abg9868

23. Melvin P. Weinstein M. *M100Ed29 | Performance Standards for Antimicrobial Susceptibility Testing, 29th Edition*. 29th ed.; 2018.

24. Prentice RL, Williams BJ, Peterson AV. On the Regression Analysis of Multivariate Failure Time Data. *Biometrika*. 1981;68(2):373-379. doi:10.2307/2335582

25. Amorim LDAF, Cai J. Modelling recurrent events: a tutorial for analysis in epidemiology. *Int J Epidemiol*. 2015;44(1):324-333. doi:10.1093/ije/dyu222

26. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-2120. doi:10.1093/bioinformatics/btu170

27. Schmieder R, Edwards R. Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets. Rodriguez-Valera F, ed. *PLoS ONE*. 2011;6(3):e17288. doi:10.1371/journal.pone.0017288

28. Beghini F, McIver LJ, Blanco-Míguez A, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. eLife. doi:10.7554/eLife.65088

29. Kaminski J, Gibson MK, Franzosa EA, Segata N, Dantas G, Huttenhower C. High-Specificity Targeted Functional Profiling in Microbial Communities with ShortBRED. Noble WS, ed. *PLOS Computational Biology*. 2015;11(12):e1004557. doi:10.1371/journal.pcbi.1004557

30. R Core Team. R: A Language and Environment for Statistical Computing. Published online 2020. www.R-project.org

31. McMurdie PJ, Holmes S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. Watson M, ed. *PLoS ONE*. 2013;8(4):e61217. doi:10.1371/journal.pone.0061217

32. Hadley W. *Ggplot2: Elegant Graphics for Data Analysis*. Springer International Publishing; 2016. doi:10.1007/978-3-319-24277-4

33. Kassambara A. ggpubr: "ggplot2" Based Publication Ready Plots. Published online November 16, 2022. Accessed November 28, 2022. https://CRAN.R-project.org/package=ggpubr

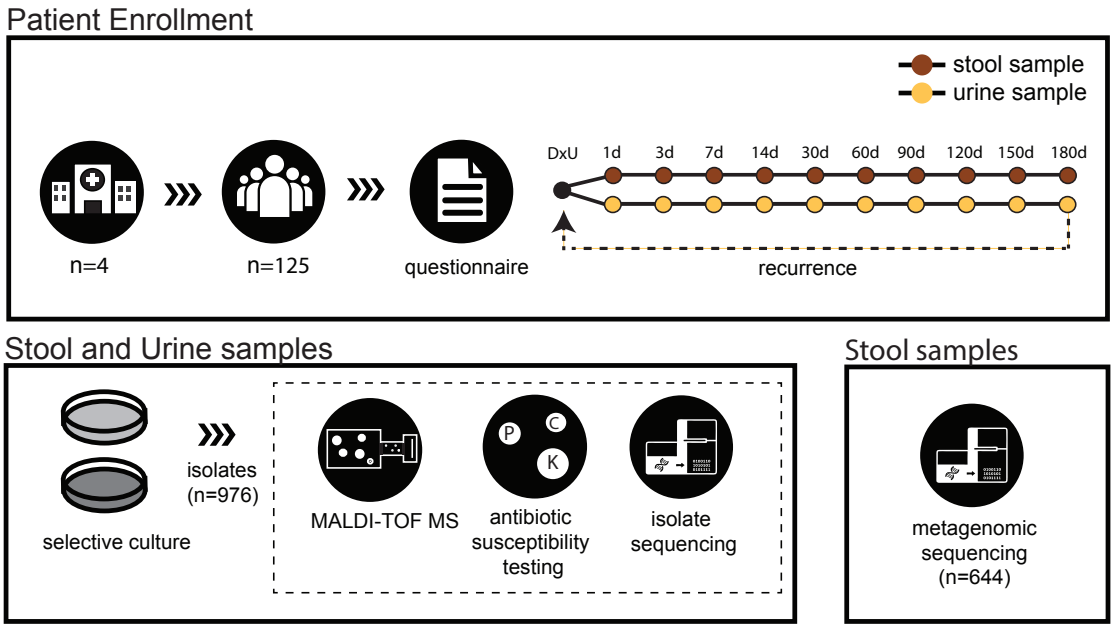

Figure 5.1 Study Overview. A cohort of 125 patients with UTI were enrolled from four hospital centers in the US. Questionnaires regarding UTI symptoms were collected at time of hospital visit. Stool and urine samples were collected from diagnosis (DxU) to 6 months after enrollment. Patients experiencing multiple episodes of UTI (recurrence) re-started the follow-up period beginning with another DxU sample. Stool and Urine samples were plated for selective culture, sequenced, and tested for antibiotic susceptibility. Results of the comparative genomic analyses of the 976 isolates are presented in Chapter 4. 644 stool samples from 106 patients were further subject to metagenomic sequencing.

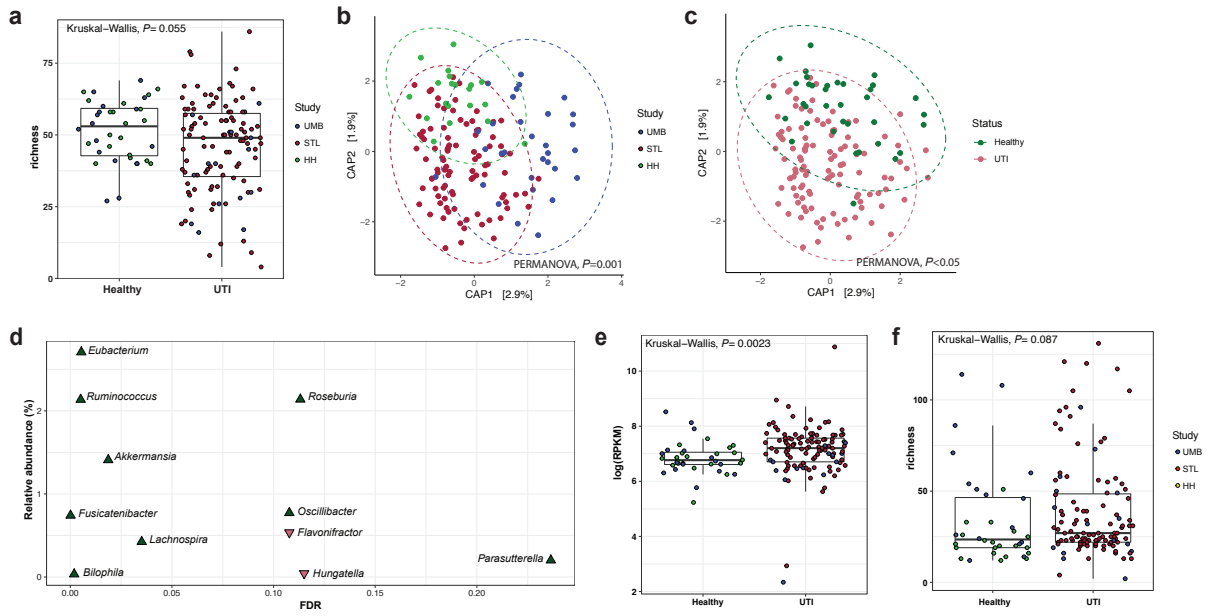

Figure 5.2 Comparison of microbiomes between healthy and UTI participants. 31 published microbiomes from a healthy humans study (HH), and 20 published microbiomes from an rUTI study (UMB) were included for cross-cohort comparisons with our samples (STL). **(a)** Richness is higher in healthy microbiomes compared to UTI (Kruskal-Wallis, *P*=0.055) **(b)** Microbiomes were significantly different by study (PERMAOMVA, *P*=0.001) but **(c)** Healthy and UTI microbiomes were significantly different even after accounting for study effect (PERMANOVA, *P*=0.043). **(d)** Differentially abundant taxa at the genus level were identified using MaAsLin2. Green and upwards pointing triangles signify taxa enriched in healthy microbiomes, while green and downwards pointing arrows signify taxa enriched in UTI individuals. X-axis denotes the false discovery rate (FDR), and Y-axis shows relative abundance. **(e)** UTI microbiomes had higher numbers of antibiotic resistance genes (ARGs) as identified by ShortBRED (Kruskal-Wallis, *P*=0.0023). X-axis shows healthy or UTI groups, while Y-axis indicates the number of ARG hits as measured by Reads Per Kilobase of reference sequence per Million sample reads (RPKM). **(f)** Richness of ARGs was not significantly different between the two groups (Kruskal-Wallis, *P*=0.087).

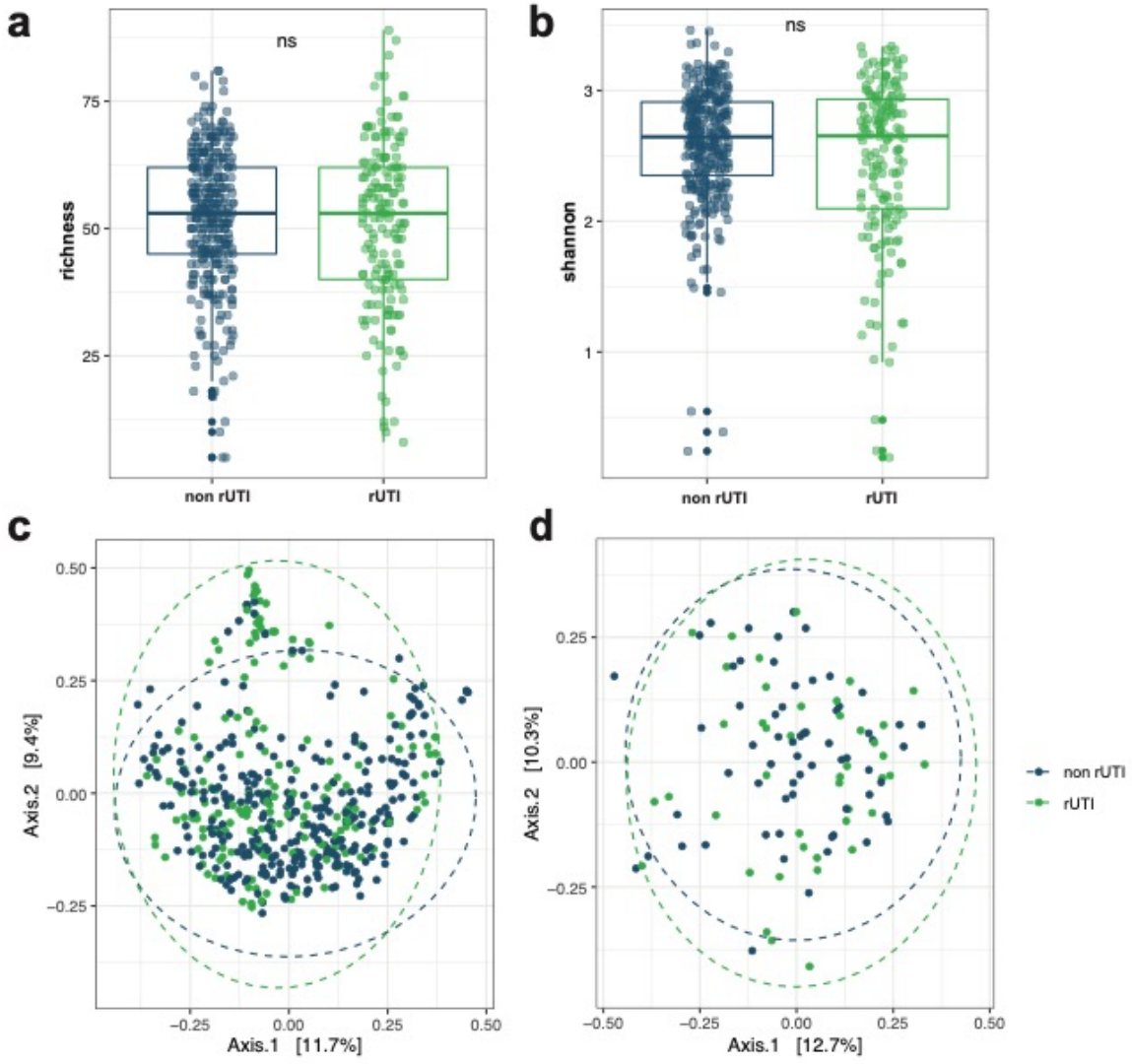

Figure 5.3 Microbiomes of rUTI and non-rUTI patients do not differ. There were no significant differences in **(a)** taxonomic richness, **(b)** Shannon index (Kruskal-Wallis), or (**c**) Overall β-diversity as tested by PERMANOVA, after accounting for repeated measures by patient (*n*=480 index episode samples). **(d)** Taxonomic profiles were averaged by patient (*n*=106), but no overall differences were observed between the two groups.

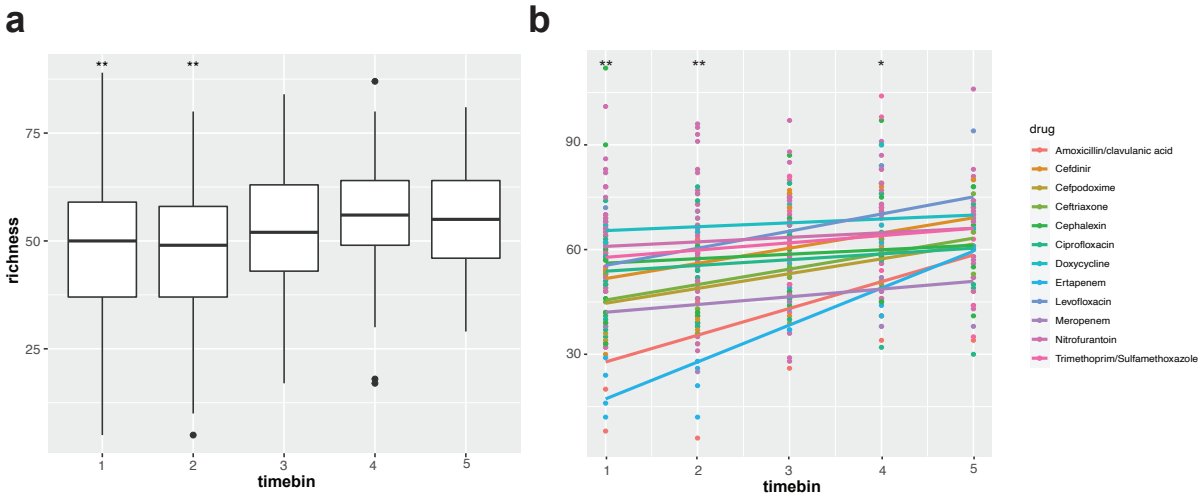

Figure 5.4 Timepoint and taxonomic richness. (a) Richness across samples was binned by timepoint (timebin 1: enrollment (*n*=96); timebin 2: end of abx- day3 post abx (n=141); timebin 3: day 7-14 (n=149); timebin 4: day 30-60 (*n*=100); timebin 5: day 90-180 (*n*=70)) for pairwise comparisons (Kruskal-Wallis). Figure shows significant BH-adjusted *P*-values in pairwise comparison against timebin 5, since this was the farthest timebin from time of treatment antibiotic. Timebins 1 and 2 are significantly depleted in richness, but not timebins 4 and 5. **(b)** Richness by timebin, stratified by treatment drug. Comparisons were made between drugs at each timebin (Wilcoxon signed-rank test, BH-adjusted). There were significant differences in richness by drug at timebins 1 and 2, but not in timebins 3, 4, or 5. (*: P <0.05; **: P <0.01; ***: P < 0.001)

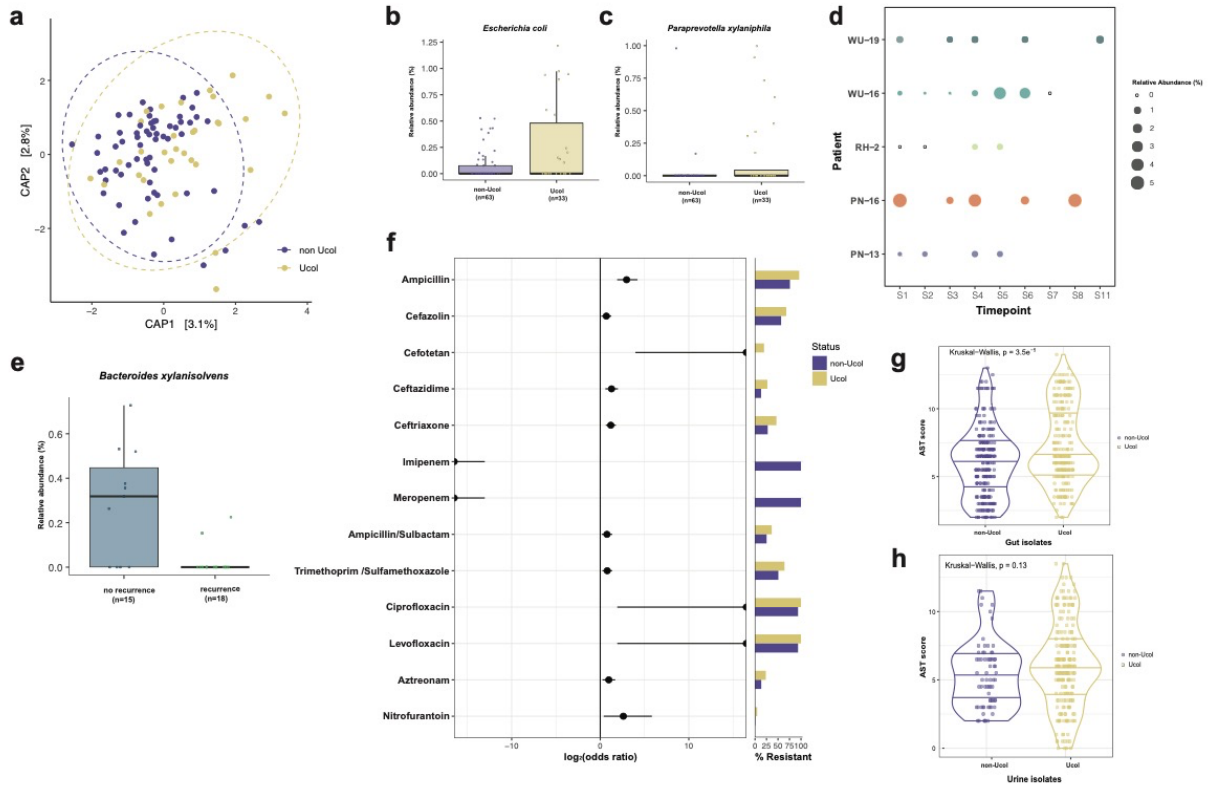

Figure 5.5 Urinary tract colonization corresponds to significant differences in gut microbiome at days 7-14 post-abx. (a) Taxonomic compositions of microbiome samples from days 7-14 post-abx were significantly different between urinary tract colonized (Ucol) and non-colonized patients (non-Ucol), even after accounting for age and treatment drug (PERMANOVA, $P < 0.05$). **(b-c)** MaAsLin2 identified two taxa to be differentially abundant in Ucol patients: *Escherichia coli* and *Paraprevotella xylaniphila*. **(d)** Ucol patients experience *E. coli* "blooms' in gut as measured by relative abundance. X-axis corresponds to sampling timepoint (S1: enrollment; S2: end of abx; S3: day3 post-abx; S4: day7; S5: day14; S6: day30; S7: day60; S8: day90; S11: day180). Y-axis rows and bubble colors correspond to patient ID, bubble size denotes relative abundance. Empty circles show 0.00% relative abundance in a sequenced sample. **(e)** *Bacteroides xylanisolvens* was the singular differentiating taxon between Ucol patients with recurrence, and Ucol patients without. **(f)** Fisher's exact tests of AST results found gut isolates from Ucol lineages to be enriched in resistance for 11 of 23 tested drugs. Gut isolates from non-Ucol lineages were enriched in resistance to imipenem and meropenem. Circles indicate the odds ratio, while lines show the 95% confidence interval. Bars on the right show the percent of isolates from each group that are resistant to each drug. **(g)** Ucol gut isolates were significantly higher in AST score compared to non-Ucol gut isolates. Lines in violin plots show quartiles of distribution for each group. **(h)** Corresponding urinary isolates were not significantly different in AST score between Ucol and non-Ucol groups.

Table 5.1 Characteristics of 125 patients with a multidrug-resistant organism (MDRO) urinary tract infection (UTI)[1]

Factor	Value	N (%)
Demographics		
Female		117 (93.6)
Age, median (IQR)		58 (42, 71)
Race	White	73 (58.4)
	African-American	47 (37.6)
	Other	5 (4.0)
Hispanic		5 (4.0)
Study site	1	41 (32.8)
	2	23 (18.4)
	3	13 (10.4)
	4	48 (38.4)
Number of recurrences per patient	1	47/125 (37.6)
	2	12/38 (31.6)
	3	7/12 (58.3)
Comorbidities		
Body mass index	Normal/underweight	32 (25.6)
	Overweight	39 (31.2)
	Obese	54 (43.2)
Cancer		22 (17.6)
Cardiovascular disease (MI, CHF, and PVD)		22 (17.6)
Cerebrovascular disease		2 (1.6)
Chronic obstructive pulmonary disease		2 (1.6)
Chronic renal failure		8 (6.4)
Diabetes mellitus		32 (25.6)
Leukemia or lymphoma		2 (1.6)
Other comorbidity[2]		59 (47.2)
Rheumatologic disease		10 (8.0)

[1] For patients with >1 UTI episode, information from the first episode is reported.
[2]Any other medical condition noted by participant.

Abbreviations: CHF, congestive heart failure; IQR, interquartile range; MI, myocardial infarction; PVD, peripheral vascular disease.

Table 5.2 Characteristics of 175 urinary tract infection episodes

Factor	N (%)
UTI antibiotic treatment[1]	
Carbapenem	10 (5.7)
Cephalosporin or a penicillin	53 (30.3)
Doxycycline	4 (2.3)
Nitrofurantoin	78 (44.6)
Quinolone	26 (14.9)
TMP, SMX, or TMP-SMX	28 (16.0)
UTI antibiotic treatment duration >7 days	85 (48.6)
Characteristics of UTI	
Organism, first episode per person (*n*=125)	
Citrobacter freundii	1 (0.8)
Escherichia coli	116 (92.8)
Klebsiella pneumoniae	5 (4.0)
Proteus mirabilis	3 (2.4)
UTI symptoms	
Bladder pain	45 (25.7)
Bladder not emptying	48 (27.4)
Blood in urine	21 (12.0)
Burning during urination	84 (48.0)
Chills	29 (16.6)
Cloudy urine	71 (40.6)
Fever	30 (17.1)
Flank pain	49 (28.0)
Other UTI and/or non-specific symptoms	20 (11.4)
Pain during urination	81 (46.3)
Urinary hesitancy	58 (33.1)
Urine odor	67 (38.3)

Abbreviations: SMX, sulfamethoxazole; TMP, trimethoprim, UTI, urinary tract infection.

[1] Treatment antibiotics are not mutually exclusive, >1 antibiotic was reported for 23 (13.1%) episodes.

Table 5.3 Univariate and multivariable risk factors for recurrence after urinary tract infection (UTI), clinical model (N=175)[1]

Factor	Value	Univariate HR and 95% CI	Multivariable HR and 95% CI
Steroids in 6 months before /at enrollment		1.89 (1.08, 3.31)	2.35 (1.28, 4.31)
Any antibiotics in 6 months before /at enrollment		2.03 (0.96, 4.28)	2.20 (1.03, 4.70)
New antibiotic after episode start and within -60 to -1 days from the censor date		0.33 (0.12, 0.90)	0.32 (0.12, 0.90)
Urinary tract colonization	Not colonized	Ref.	
	Colonized	1.58 (0.94, 2.66)	
	Unknown	0.68 (0.26, 1.78)	
Bladder not emptying		0.65 (0.36, 1.17)	
Other UTI symptoms		0.46 (0.17, 1.27)	
Urinary hesitancy		0.67 (0.38, 1.16)	
UTI treatment with cephalosporin or a penicillin		1.67 (1.01, 2.76)	
UTI treatment with TMP-SMX		0.36 (0.14, 0.90)	0.39 (0.15, 0.99)
UTI antibiotic treatment duration >7 days		0.71 (0.43, 1.17)	0.55 (0.32, 0.95)

[1] Leukemia/lymphoma and post-index steroids were significant in univariate analysis, but not entered into the model because of cell sizes of 1.

Table 5.4 Univariate clinical risk factors for recurrence after UTI[1]

Factor	Value	N and rate of recurrence per 10K days	Univariate HR and 95% CI
Demographics			
Age	18–64 years	43 (30.9)	Ref.
	65–79 years	17 (32.1)	0.95 (0.54, 1.67)
	≥ 80 years	6 (44.9)	1.21 (0.51, 2.90)
Gender	Male	4 (27.0)	Ref.
	Female	62 (32.5)	1.26 (0.45, 3.51)
Hispanic	Yes	2 (24.3)	0.68 (0.16, 2.91)
	No	64 (32.4)	Ref.
Race	White	39 (33.9)	Ref.
	African-American	25 (30.3)	1.01 (0.60, 1.70)
	Other	2 (24.1)	0.79 (0.19, 3.31)
Study site	1	22 (31.6)	Ref.
	2	9 (22.2)	0.67 (0.30, 1.46)
	3	5 (24.0)	0.69 (0.26, 1.85)
	4	30 (40.2)	1.19 (0.68, 2.08)
Comorbidities			
Body mass index	Normal/underweight	15 (26.6)	Ref.
	Overweight	21 (34.0)	1.31 (0.67, 2.56)
	Obese	30 (34.2)	1.30 (0.69, 2.44)
Cancer	Yes	14 (37.2)	1.18 (0.64, 2.17)
	No	52 (30.9)	Ref.
Cardiovascular disease (MI, CHF, and PVD)	Yes	9 (22.9)	0.70 (0.34, 1.43)
	No	57 (34.2)	Ref.
Cerebrovascular disease	Yes	1 (25.2)	0.82 (0.11, 5.98)
	No	65 (32.2)	Ref.
COPD	Yes	1 (39.7)	1.18 (0.16, 8.58)

Factor	Value	N and rate of recurrence per 10K days	Univariate HR and 95% CI
	No	65 (32.0)	Ref.
Chronic renal failure	Yes	5 (36.1)	1.28 (0.50, 3.24)
	No	61 (31.8)	Ref.
Diabetes mellitus	Yes	16 (29.5)	0.92 (0.52, 1.62)
	No	50 (33.0)	Ref.
Leukemia or lymphoma	Yes	3 (152.3)	4.49 (1.31, 15.36)
	No	63 (30.9)	Ref.
Other comorbidity	Yes	27 (26.9)	0.77 (0.47, 1.27)
	No	39 (37.0)	Ref.
Rheumatologic disease	Yes	5 (28.7)	0.86 (0.34, 2.16)
	No	61 (32.4)	Ref.
Baseline infection and hospitalization history			
Hospital admission within 60 days prior to UTI onset	Yes	6 (21.9)	0.66 (0.28, 1.54)
	No	60 (33.7)	Ref.
Any infections in previous 12 months	Yes	7 (27.3)	0.78 (0.35, 1.73)
	No	59 (32.8)	Ref.
UTI clinical characteristics			
Hospitalized for current infection	Yes	9 (45.4)	1.33 (0.64, 2.77)
	No	57 (30.7)	Ref.
History of any previous UTIs (at time E1)	Yes	52 (35.0)	1.32 (0.73, 2.41)
	No	14 (24.4)	Ref.
History of any previous UTIs, categorical (at time E1)	0	14 (24.4)	Ref.
	1	21 (31.4)	1.29 (0.65, 2.54)
	2+	31 (38.1)	1.35 (0.71, 2.58)

Factor	Value	N and rate of recurrence per 10K days	Univariate HR and 95% CI
UTI in previous 60 days	Yes	21 (36.6)	1.00 (0.53, 1.89)
	No	45 (30.3)	Ref.
Medications in 6 months before/at enrollment			
Bladder or urinary tract medication	Yes	10 (41.0)	1.21 (0.61, 2.40)
	No	56 (30.9)	Ref.
Hormonal birth control	Yes	1 (22.5)	0.67 (0.09, 4.87)
	No	65 (32.3)	Ref.
Hormone replacement therapy	Yes	6 (29.1)	0.67 (0.28, 1.64)
	No	60 (32.4)	Ref.
Immunosuppressant	Yes	14 (40.0)	1.38 (0.76, 2.51)
	No	52 (30.5)	Ref.
Prostate medications	Yes	5 (41.0)	1.28 (0.51, 3.20)
	No	61 (31.5)	Ref.
Steroids	Yes	17 (52.3)	1.89 (1.08, 3.31)
	No	49 (28.3)	Ref.
Antibiotics in 6 months before/at enrollment			
Any antibiotic	Yes	58 (36.2)	2.03 (0.96, 4.28)
	No	8 (17.6)	Ref.
Cephalosporins	Yes	26 (40.6)	1.39 (0.84, 2.30)
	No	40 (28.2)	Ref.
Nitrofurantoin	Yes	31 (32.9)	1.11 (0.67, 1.83)
	No	35 (31.4)	Ref.
Quinolones	Yes	15 (27.8)	0.88 (0.49, 1.58)
	No	51 (33.6)	Ref.
TMP-SMX	Yes	15 (25.7)	0.82 (0.46, 1.47)

Factor	Value	N and rate of recurrence per 10K days	Univariate HR and 95% CI
	No	51 (34.6)	Ref.
Post-index medications[2]			
New bladder/urinary tract medication	Yes	1 (15.2)	0.46 (0.06, 3.43)
	No	65 (32.6)	Ref.
New hormone replacement therapy	Yes	1 (10.7)	0.32 (0.04, 2.36)
	No	65 (33.1)	Ref.
New steroid	Yes	1 (7.8)	0.22 (0.03, 1.59)
	No	65 (33.7)	Ref.
New antibiotic	Yes	4 (11.9)	0.33 (0.12, 0.90)
	No	62 (36.0)	Ref.
Colonization[3]			
Intestinal colonization	Not colonized	19 (32.5)	Ref.
	Colonized	42 (35.0)	1.11 (0.63, 1.96)
	Unknown	5 (18.3)	0.59 (0.22, 1.63)
Urinary tract colonization	Not colonized	34 (28.1)	Ref.
	Colonized	27 (46.3)	1.58 (0.94, 2.66)
	Unknown	5 (18.8)	0.68 (0.26, 1.78)
UTI symptoms			
Bladder pain	Yes	14 (25.7)	0.74 (0.41, 1.35)
	No	52 (34.4)	Ref.
Bladder not emptying	Yes	15 (23.8)	0.65 (0.36, 1.17)
	No	51 (35.8)	Ref.
Blood in urine	Yes	6 (22.6)	0.71 (0.30, 1.65)
	No	60 (33.5)	Ref.
Burning during urination	Yes	30 (31.6)	1.01 (0.61, 1.67)
	No	36 (32.5)	Ref.

Factor	Value	N and rate of recurrence per 10K days	Univariate HR and 95% CI
Chills	Yes	10 (26.5)	0.77 (0.39, 1.53)
	No	56 (33.3)	Ref.
Cloudy urine	Yes	27 (33.2)	1.09 (0.66, 1.80)
	No	39 (31.4)	Ref.
Fever	Yes	11 (26.8)	0.84 (0.43, 1.63)
	No	55 (33.4)	Ref.
Flank pain	Yes	15 (26.2)	0.77 (0.43, 1.37)
	No	51 (34.4)	Ref.
Other UTI symptoms	Yes	4 (16.6)	0.46 (0.17, 1.27)
	No	62 (34.1)	Ref.
Pain during urination	Yes	31 (31.6)	0.99 (0.60, 1.62)
	No	35 (32.5)	Ref.
Urinary hesitancy	Yes	19 (25.4)	0.67 (0.38, 1.16)
	No	47 (35.9)	Ref.
Urine odor	Yes	23 (29.1)	0.91 (0.54, 1.53)
	No	43 (34.0)	Ref.
UTI antibiotic treatment[4]			
Carbapenem	Yes	5 (56.4)	1.74 (0.69, 4.39)
	No	61 (31.0)	Ref.
Cephalosporin or a penicillin	Yes	26 (48.3)	1.67 (1.01, 2.76)
	No	40 (26.3)	Ref.
Doxycycline	Yes	1 (14.9)	0.50 (0.07, 3.68)
	No	65 (32.7)	Ref.
Nitrofurantoin	Yes	27 (27.8)	0.79 (0.48, 1.31)
	No	39 (35.9)	Ref.
Quinolone	Yes	8 (25.5)	0.85 (0.40, 1.79)
	No	58 (33.3)	Ref.

Factor	**Value**	**N and rate of recurrence per 10K days**	**Univariate HR and 95% CI**
TMP-SMX	Yes	5 (12.5)	0.36 (0.14, 0.90)
	No	61 (36.8)	Ref.
UTI antibiotic treatment duration	≤ 7 days	38 (39.3)	Ref.
	>7 days	28 (25.7)	0.71 (0.43, 1.17)
>1 type of UTI antibiotic treatment	Yes	6 (18.7)	0.59 (0.25, 1.37)
	No	60 (34.6)	Ref.

Abbreviations: COPD, chronic obstructive pulmonary disease; CI, confidence interval; CHF, congestive heart failure; E1, episode 1; HR, hazard ratio; MI, myocardial infarction; PVD, peripheral vascular disease; SMX, sulfamethoxazole; TMP, trimethoprim; UTI, urinary tract infection.

[1] The following variables were evaluated but not included in the table due to zero cells: post-index new birth control medication, post-index new immunosuppressant medication, post-index new prostate medication.
[2] Post-index medications were captured at the episode-level after the start of the episode but before the censor date, i.e., the earliest of recurrence/last date of follow up. Post-index antibiotics were further restricted to those within -60 to -1 days from the censor date.
[3] Colonization was defined as colonization at any time after the start of the episode and before the censor date, i.e., the earliest of recurrence/last date of follow up
[4] UTI antibiotic treatment not mutually exclusive.

Table 5.5 Reference microbiomes

Accession Number	Sample_ID	Patient_ID	Site	group	Study
SRS8744671	01-S01	HH-1	HH	Control	HH
SRS8744590	02-S01	HH-2	HH	Control	HH
SRS8744580	03-S01	HH-3	HH	Control	HH
SRS8744563	04-S01	HH-4	HH	Control	HH
SRS8744804	05-S01	HH-5	HH	Control	HH
SRS8744789	06-S01	HH-6	HH	Control	HH
SRS8744769	07-S01	HH-7	HH	Control	HH
SRS8744752	08-S01	HH-8	HH	Control	HH
SRS8744642	09-S01	HH-9	HH	Control	HH
SRS8744627	10-S01	HH-10	HH	Control	HH
SRS8744610	11-S01	HH-11	HH	Control	HH
SRS8744594	12-S01	HH-12	HH	Control	HH
SRS8744828	13-S01	HH-13	HH	Control	HH
SRS8744816	14-S01	HH-14	HH	Control	HH
SRS8744735	15-S01	HH-15	HH	Control	HH
SRS8744719	16-S01	HH-16	HH	Control	HH
SRS8744703	17-S01	HH-17	HH	Control	HH
SRS8744687	18-S01	HH-18	HH	Control	HH
SRS8744661	19-S01	HH-19	HH	Control	HH
SRS8744842	20-S01	HH-20	HH	Control	HH
SRR14881730	UMB01_00	UMB01	UMB	Control	UMB
SRR14881720	UMB02_00	UMB02	UMB	Control	UMB
SRR14881706	UMB03_00	UMB03	UMB	Control	UMB
SRR14882081	UMB04_00	UMB04	UMB	rUTI	UMB
SRR14882036	UMB05_00	UMB05	UMB	rUTI	UMB
SRR14882025	UMB06_00	UMB06	UMB	Control	UMB
SRR14882018	UMB07_00	UMB07	UMB	Control	UMB
SRR14882008	UMB08_01.1	UMB08	UMB	rUTI	UMB
SRR14882064	UMB09_01	UMB09	UMB	Control	UMB
SRR14882060	UMB10_01	UMB10	UMB	rUTI	UMB
SRR14882047	UMB11_01	UMB11	UMB	rUTI	UMB
SRR14882041	UMB12_04	UMB12	UMB	rUTI	UMB
SRR14882002	UMB13_00	UMB13	UMB	Control	UMB
SRR14881993	UMB14_00	UMB14	UMB	Control	UMB
SRR14881990	UMB15_04	UMB15	UMB	rUTI	UMB

SRR14881985	UMB16_01	UMB16	UMB	Control	UMB
SRR14881982	UMB17_01	UMB17	UMB	rUTI	UMB
SRR14881968	UMB18_00.1	UMB18	UMB	rUTI	UMB
SRR14881956	UMB19_01	UMB19	UMB	Control	UMB
SRR14881943	UMB20_01	UMB20	UMB	rUTI	UMB
SRR14881929	UMB21_04	UMB21	UMB	Control	UMB
SRR14881922	UMB22_01	UMB22	UMB	rUTI	UMB
SRR14881909	UMB23_02	UMB23	UMB	rUTI	UMB
SRR14881899	UMB24_02	UMB24	UMB	rUTI	UMB
SRR14881885	UMB25_01	UMB25	UMB	rUTI	UMB
SRR14881872	UMB26_02	UMB26	UMB	Control	UMB
SRR14881863	UMB27_00	UMB27	UMB	Control	UMB
SRR14881850	UMB28_00	UMB28	UMB	Control	UMB
SRR14881838	UMB29_01	UMB29	UMB	Control	UMB
SRR14881827	UMB30_01	UMB30	UMB	rUTI	UMB
SRR14881824	UMB31_00	UMB31	UMB	Control	UMB

Table 5.6 AST Fisher's exact test results

Drug	Ucol_R (# isos)	Ucol_S (# isos)	noncol_R (# isos)	noncol_S (# isos)	pval	BH-adj pval
Imipenem	0	234	250	0	7.19E-145	8.27E-144
Meropenem	0	234	250	0	7.19E-145	8.27E-144
Cefotetan	46	188	0	250	2.41E-16	1.84E-15
Ampicillin	225	9	190	60	4.08E-11	2.35E-10
Ceftriaxone	109	125	69	181	2.04E-05	8.87E-05
Ciprofloxacin	234	0	234	16	2.70E-05	8.87E-05
Levofloxacin	234	0	234	16	2.70E-05	8.87E-05
Ceftazidime	63	171	33	217	0.00016152	0.000464
Trimethoprim.sulfa	150	84	127	123	0.00331171	0.008463
Aztreonam	55	179	34	216	0.00670758	0.015427
Ampicillin.Sulbactam	85	149	63	187	0.01015424	1.95E-02
Nitrofurantoin	11	223	2	248	0.0097128	1.95E-02
Cefazolin	159	75	142	108	0.01464205	2.59E-02
Gentamicin	29	205	45	205	0.10049489	1.65E-01
Amikacin	0	234	3	242	0.24882732	3.58E-01
Pipercillin.Tazobactam	0	234	3	247	0.24926852	3.58E-01
Cefepime	31	203	29	221	0.58473917	0.791118
Doxycycline	79	155	89	161	0.70282738	0.898057
Fosfomycin	7	227	6	244	0.78215797	0.946823
Ceftazidime.Avibactam	0	234	1	249	1	1.00E+00
Ceftolozane.Tazobactam	1	233	2	248	1	1
Minocycline	17	217	18	232	1	1
Tigecycline	0	234	0	250	1	1

Chapter 6

6.1 Conclusion

Throughout this Thesis I have studied seemingly unrelated topics: The results of a clinical trial using FMT to treat recurrent *Clostridium difficile* infection (Chapter 2), in-host adaptation of *Mycobacterium abscessus* (Chapter 3), recurrent urinary tract infections and the uropathogenic *Escherichia coli* which cause them (Chapter 4), clinical risk factor models and the gut microbiome (Chapter 5). All these chapters, however, carry the common thread of exploring the nuanced relationship between human health and the microbes in our bodies. By understanding how pathogens adapt from the environment to the human host, the mechanisms that facilitate their survival, and the functional consequences of those adaptive behaviors, we can reinvison our approach to treatment.

As the threat of multidrug resistant superbugs looms near, there is a pressing need to steer away from broad-spectrum antimicrobials which select for further resistance. Instead, novel developments such as FMT, prebiotics, and therapeutics that limit pathogen adherence factors are increasingly informed by our understanding of pathogen within-host activity and aim to specifically target the source of dysbiosis.

Future studies will continue to enhance our understanding of human-microbe interactions through various technological advances: accurate in-depth profiling of individual strains of microbes from metagenomic sequencing data; inclusion of viruses,

fungi, protists, and archaea to examine cross-kingdom interactions in the microbiome; and expanded study of other sites such as the urine, lung, skin, and oral microbiome.

This Thesis is but a droplet in the ocean of our collective knowledge of microbiology, yet there is much more work to be done. I am excited to see what the future holds: for both my microbes and me.