Summer 8-15-2015

Characterizing and Modeling Antibiotic Resistance Dynamics in Diverse Microbial Communities

Molly Krisann Gibson
Washington University in St. Louis

Follow this and additional works at: http://openscholarship.wustl.edu/art_sci_etds
Part of the Biology Commons

Recommended Citation
http://openscholarship.wustl.edu/art_sci_etds/572

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

by

Molly K. Gibson

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

August 2015
St. Louis, Missouri
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contents</td>
<td>iii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>vii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>ix</td>
</tr>
<tr>
<td>List of Abbreviations</td>
<td>xi</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>xiii</td>
</tr>
<tr>
<td>Abstract of the Dissertation</td>
<td>xvii</td>
</tr>
</tbody>
</table>

## 1 Introduction

1.1 The threat of antibiotic resistance ........................................... 1

1.1.1 Traditional mechanisms of antibiotic resistance .......................... 2

1.1.2 Strategies for combatting antibiotic resistance ............................ 4

1.1.3 Reservoirs of transferable antibiotic resistance ........................... 6

1.1.4 The commensal resistome as a resilience factor ............................. 8

1.2 Challenges in studying diverse reservoirs of antibiotic resistance .... 10

1.2.1 Functional metagenomic selections ............................................. 10

1.3 Antibiotics and the developing infant gut microbiota and resistome ...... 14

1.3.1 Normal infant gut microbiota and resistome development .................. 16

1.3.2 Early-life antibiotics and the human microbiota ........................... 18

1.3.3 Long-term effects of early-life antibiotic therapy ........................ 20

1.3.4 Enrichment of the infant antibiotic resistome ............................... 22

1.3.5 Future Directions ................................................................. 23
1.4 Scope of thesis ................................................................. 24
1.5 Acknowledgements ......................................................... 26

2 Improved Annotation of Antibiotic Resistance Determinants Reveals Microbial Resistomes Cluster by Ecology 27
  2.1 Abstract & Introduction .................................................. 27
  2.2 Results & Discussion .................................................... 29
    2.2.1 Optimization of antibiotic resistance profile hidden Markov models (Resfams) .................................................. 29
    2.2.2 Resfams accurately predicts novel resistance functions from sequence alone .................................................. 30
    2.2.3 Antibiotic resistomes cluster by ecology ......................... 32
    2.2.4 Antibiotic resistance potential encoded in microbial genomes .... 35
    2.2.5 Discussion ................................................................. 40
  2.3 Materials & Methods ..................................................... 41
    2.3.1 Training and Benchmarking of Resfams pHMMs .................. 41
    2.3.2 Resfams pHMM Precision and Recall Optimization Overview .... 42
    2.3.3 Protein annotation using Resfams pHMMs or BLAST to AR specific databases .................................................. 43
    2.3.4 Resistome analysis using functional metagenomic selections .......... 44
    2.3.5 Summary of microbial genomes used in analysis ................. 46
    2.3.6 Resfams pHMM Annotation from Microbial Sequence Alone .......... 46
    2.3.7 Comparison of Resfams pHMMs to BLAST to AR gene databases .... 47
    2.3.8 Generation of binary heatmaps for genome comparisons .......... 48
  2.4 Acknowledgments .......................................................... 49
  2.5 Supplementary Figures .................................................. 50

3 Fast and accurate metagenomic characterization of antibiotic resistance reservoirs 57
  3.1 Abstract & Introduction .................................................. 57
  3.2 Results & Discussion .................................................... 60
    3.2.1 Profiling protein families in metagenomes using representative marker sequences .................................................. 62
    3.2.2 ShortBRED achieves better specificity than mapping reads to centroids .................................................. 63
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2.3 ShortBRED in analyses of the 10%- and 25%-spiked metagenomes</td>
<td>66</td>
</tr>
<tr>
<td>3.2.4 ShortBRED is faster than centroid-based profiling</td>
<td>67</td>
</tr>
<tr>
<td>3.2.5 Antibiotic resistance in the human gut microbiome worldwide</td>
<td>68</td>
</tr>
<tr>
<td>3.2.6 Connecting antibiotic resistance to phylogeny</td>
<td>70</td>
</tr>
<tr>
<td>3.2.7 Predicting antibiotic resistance profiles from community composition</td>
<td>73</td>
</tr>
<tr>
<td>3.2.8 Discussion</td>
<td>75</td>
</tr>
<tr>
<td>3.3 Materials &amp; Methods</td>
<td>79</td>
</tr>
<tr>
<td>3.3.1 Creating protein family-specific marker sequences with ShortBRED-Identify</td>
<td>79</td>
</tr>
<tr>
<td>3.3.2 Profiling protein family metagenomic abundance with ShortBRED-Quantify</td>
<td>82</td>
</tr>
<tr>
<td>3.3.3 Creation of synthetic spiked metagenomes</td>
<td>83</td>
</tr>
<tr>
<td>3.3.4 Application of ShortBRED to human gut metagenomes</td>
<td>84</td>
</tr>
<tr>
<td>3.3.5 Application of ShortBRED to bacterial isolate genomes</td>
<td>85</td>
</tr>
<tr>
<td>3.4 Acknowledgements</td>
<td>85</td>
</tr>
<tr>
<td>3.5 Supplementary Tables &amp; Figures</td>
<td>87</td>
</tr>
<tr>
<td>4 Response of preterm infant gut microbiota and resistome to early life antibiotics</td>
<td>91</td>
</tr>
<tr>
<td>4.1 Abstract &amp; Introduction</td>
<td>91</td>
</tr>
<tr>
<td>4.2 Results &amp; Discussion</td>
<td>93</td>
</tr>
<tr>
<td>4.2.1 Cohort of 84 preterm infants sampled longitudinally for first 2.5 months of life</td>
<td>93</td>
</tr>
<tr>
<td>4.2.2 Impact of early life antibiotics on preterm infant gut microbiota development</td>
<td>94</td>
</tr>
<tr>
<td>4.2.3 Functional profiling of the preterm infant gut resistome</td>
<td>98</td>
</tr>
<tr>
<td>4.2.4 Preterm infant gut microbiota and resistome response to specific antibiotic therapies</td>
<td>103</td>
</tr>
<tr>
<td>4.2.5 Discussion</td>
<td>106</td>
</tr>
<tr>
<td>4.3 Materials &amp; Methods</td>
<td>108</td>
</tr>
<tr>
<td>4.3.1 Sample collection</td>
<td>108</td>
</tr>
<tr>
<td>4.3.2 Metagenomic DNA Extraction</td>
<td>109</td>
</tr>
<tr>
<td>4.3.3 Comparison of term to preterm gut microbiota using 16S rRNA marker gene sequencing</td>
<td>110</td>
</tr>
</tbody>
</table>
List of Figures

1-1 The commensal resistome as a resilience factor ........................................ 8
1-2 Functional metagenomic selections for antibiotic resistance .................... 12
1-3 Antibiotic resistomes are under sampled in uncultured microbial commu-
nities .............................................................................................................. 13

2-1 Annotation of functional metagenomic selections using Resfams .............. 30
2-2 Resfams improves the resolution of annotation of antibiotic resistance genes 32
2-3 Resistomes differ by ecology ................................................................. 34
2-4 Resistomes annotated by Resfams across phylogeny and habitats ............ 37
2-5 Annotation of antibiotic resistance across habitats using Resfams compared to pairwise sequence alignment .................................................. 39
2-S1 Resfams family precision and recall distributions ..................................... 50
2-S2 Phenotypic resistance profiles across ecologies ....................................... 51
2-S3 Principal coordinates (PCoA) plots of functional metagenomic selections . 52
2-S4 Distribution of total AR genome composition across habitat and bacteria phyla ..................................................................................................... 53
2-S5 Enrichment of AR mechanisms in functional metagenomic selections ...... 54
2-S6 Enrichment of antibiotic resistance functions in pathogens ..................... 55
2-S7 Resfams pHMM annotation of 540,732 proteins in the UniProtKB/Swiss-
Prot curated database ..................................................................................... 56

3-1 The ShortBRED algorithm ........................................................................ 61
3-2 Accuracy of ShortBRED and centroid-based profiling .............................. 65
3-3 Speed of execution: ShortBRED versus centroid-based profiling .............. 67
3-4 Antibiotic resistance in the human gut microbiome ................................... 69
3-5 Prevalence of antibiotic resistance across bacterial isolate genomes .......... 72
3-S1 Accuracy of ShortBRED and centroid-based mapping in applications involving 454 sequencing data ......................................................... 87
3-S2 Training/testing of ShortBRED’s default parameters ..................... 88
3-S3 Prevalence of all AR classes across bacterial isolate genomes ........ 89
4-1 Composition and development of the preterm infant gut microbiota . . 97
4-2 Antibiotic treatment significantly alters preterm infant gut microbiota development ...................................................................................... 99
4-3 Functional metagenomic selections identify novel antibiotic resistance gene clusters in Proteobacterial species in the preterm infant gut .... 102
4-4 Species and antibiotic resistance gene enrichment following specific antibiotic treatments ............................................................... 105
4-5 Predictable response of preterm infant gut microbiota species richness to gentamicin and vancomycin combination treatment ................. 107
4-S1 The preterm infant gut microbiota is lower in species richness and compositionally distinct from age-matched term infants ......................... 128
4-S2 84 preterm infants and 401 samples included in this study with antibiotic treatments .................................................................................. 129
4-S3 84 preterm infants and 401 samples included in this study with feeding . 130
4-S4 Functional selections of 21 preterm infant gut metagenomes for resistance against 16 antibiotics .................................................................... 131
4-S5 Resistome development in preterm infants early in life .................. 132
4-S6 Rarefaction curves for species identified by sequencing depth (after quality-filtering and removal of human-sequence reads) ............... 133
List of Tables

3-1 Characteristics of ShortBRED markers for profiling synthetic metagenomes 64
3-2 Characteristics of ShortBRED markers used to profile metagenomes and Bacterial Genomes ......................................................... 77
3-S1 Performance of ShortBRED and centroid method .......................... 87
4-1 Details of preterm infant cohort analyzed in this study ....................... 95
4-S1 Antibiotic exposure for only those infants who received a particular antibiotic 134
4-S2 Metagenomic libraries constructed from preterm infant stool samples ... 135
4-S3 Antibiotics and minimum inhibitory concentration used for functional selections ................................................................. 136
List of Abbreviations

AR ............ antibiotic resistance
ARDB .......... antibiotic resistance database
CARD .......... comprehensive antibiotic resistance database
FPR .......... false positive rate
HGT .......... horizontal gene transfer
IMG .......... integrated microbial genomes (database)
IQR .......... interquartile range
MDR .......... multidrug resistant
MSA .......... multiple sequence alignment
NICU .......... neonatal intensive care unit
ORF .......... open reading frame
PCoA .......... principal coordinates analysis
pHMM .......... profile hidden Markov model
TPR .......... true positive rate
WMS .......... whole metagenome shotgun
Acknowledgments

This work was funded in part by a National Science Foundation graduate research fellowship (award number DGE-1143954) and through support from The Mr. & Mrs. Spencer T. Olin Fellowship program at Washington University.

I am eternally grateful for the incredible friends and colleagues I have had the privilege of being influenced by in the Center for Genome Sciences, the Division of Biology and Biomedical Sciences, and the greater Washington University and Saint Louis communities. I will forever be changed by the incredible kindness, generosity, and sheer brilliance of these people.

I would like to thank my advisor, Gautam Dantas, for his guidance, encouragement, and most of all his confidence in me as an independent scientist and leader. Gautam’s mentorship and genuine care for the people around him has resulted in a uniquely incredible place to work and learn. Gautam has taught me to not shy from the hard problems, continually have high expectations, and to never stop taking risks.

I would like to thank my thesis committee for their advice and critical feedback. In particular, I would like to thank Barak Cohen for continually challenging me from day one to think about the big picture, frame questions with clear scientific insight, and present my research in a meaningful way. A big thank you to Jeff Gordon for his support, expertise, and display of scientific rigor. His commitment to improving the human condition and advancing science is truly impressive. I would also like to thank Jim Skeath for his gracious
I have had the privilege of collaborating with some incredible scientists and clinicians during my thesis work. I would like to thank Barb Warner for her mentorship and genuine support of me as a scientist and a person and for her invaluable clinical insight. I would like to thank Phil Tarr for his passion and excitement for my work and for always challenging me to go a step further.

My colleagues in the Dantas Lab have set the bar, for which I will measure all future work environments against, extremely high. I could not have asked for a better group of people to grow and learn beside. In particular, I would like to thank Mitch Pesesky for his contagious optimism, yet critical scientific mind, as well as for his company during the many hours setting up a gnotobiotic mouse facility and collecting endless fecal samples from that inevitable stubborn mouse. I would also like to thank Kevin Forsberg for his excitement and passion and for all the insanely in-depth scientific or theoretical discussions, Erica Pehrsson for always being willing to provide another perspective and/or emotional support, Pablo Tsukayama for being a patient and fun bay mate and for his invaluable experimental insight, and Pat Gonzales for always sharing his incredible bank of antibiotic knowledge at the drop of a hat. Finally, my work would not have been possible without the amazing technical expertise and support of Bin Wang.

A sincere thank you to all members of the Center for Genome Sciences for demonstrating the true meaning of community. Thank you to Brian Koebbe and Eric Martin for their technical support. My work would not have been possible without their flexibility, patience, and endless willingness to help. Thank you to the members of the Gordon Lab and morning coffee time buddies for both the scientific as well as random conversations.

Thank you to all of my friends, near and far, for your support and friendship through the good times and the hard times. In particular, to Laura & Jacob Blanton and Kurt & Marysa Warnhoff for being there from day one and ensuring graduate school was an
unforgettable experience and to Jamie Kwasnieski for your friendship and for forcing me to make the hard decisions.

A huge thank you to our Saint Louis Ultimate Frisbee community for keeping me sane outside of lab. You showed me the true joy of Ultimate Frisbee and allowed me to grow and challenge myself in a different capacity in a unbelievably supportive environment. In particular, a big thank you to Alex Morales-Heil for being the best friend, neighbor, and workout partner anyone could ask for.

I cannot thank my family enough for their support. My parents have always been my biggest advocates and have shown true selflessness in their support of my career goals. My in-laws have been enthusiastic cheerleaders throughout my journey and helped me in ways I cannot even describe.
Above all else, I owe all that I am and all that I have to my husband, Ben.

To your endless encouragement and faith.

To your unconditional love.

To making me a better person everyday.
Complex microbial communities colonize every habitat investigated to date, including soil, animals, water, and humans, as well as the structures we live in. It has been hypothesized that there is a network of exchange allowing both bacterial organisms and functions to seamlessly move between and within these environments. These microbial communities often serve essential and beneficial functions for the host organism or environment. This is particularly evident in the human gut, where microbial communities consistently provide a set of services to its human host, including protecting against enteric pathogens, liberating nutrients from food, and signaling immune system regulation. However, these communities can also serve as reservoirs of antibiotic resistance (AR) genes readily available for transfer to pathogenic bacteria, compromising treatment of infectious disease.

The goal of my thesis was to (1) understand how AR genes move through this potential network of exchange between environments and (2) how therapeutic levels of antibiotics effect a dynamically developing microbial community and its encoded reservoir of AR genes within the human gut environment. In order to accomplish this goal, I developed and optimized two computational methods, Resfams and ShortBRED, to study antibiotic resistance in diverse and dynamic microbial communities using high-throughput shotgun sequencing methods. I employed Resfams, an AR-specific profile hidden Markov model
(pHMM) database, to demonstrate that despite the apparent environmental source of clinical AR genes, the AR functions encoded in soil microbial communities are significantly distinct from clinical, human-associated microbial communities. I was also able to show that using a consensus and probabilistic model approach, like Resfams, significantly reduces bias when studying diverse microbial communities over the current best practice of pairwise-sequence alignment (BLAST). To study how therapeutic levels of specific antibiotics modulate microbial communities and encoded AR functions, I studied the developing preterm infant gut microbiota, a population which almost universally receives early and prolonged antibiotic therapy at birth. By analyzing 401 fecal samples from 84 longitudinally-sampled preterm infants, I show that meropenem, cefotaxime, and ticarcillin-clavulanate significantly reduce species richness, and significantly enrich or deplete specific bacterial species and AR genes. In contrast, vancomycin and gentamicin, two of the antibiotics most commonly administered to preterm infants, result in a varied response in species richness. I show this response is predictable with 85% accuracy based on the relative abundance of only two bacterial species and two AR genes prior to treatment. Further, I show that AR genes enriched following specific antibiotic treatments are generally unique to the specific treatment and highly correlated with the abundance of a single bacterial species. In addition to AR genes relevant to specific antibiotics likely providing a protective function, I show that all antibiotic treatments also result in widespread collateral microbiome impact, through enrichment of AR genes with no known direct activity against the specific antibiotic treatment. Together, these results demonstrate that while movement of AR functions between environmental and human-associated communities is limited, antibiotic pressure applied to dynamic microbial communities can result in significant and predictable responses. Further work to interrogate the rates of AR movement between alternative nodes in the network of gene exchange as well as animal models to confirm our predicted response of developing microbial communities to antibiotic treatment is warranted.
Chapter 1

Introduction

1.1 The threat of antibiotic resistance

It is estimated that infection with antibiotic resistant pathogens incurs over $55 billion in societal and direct healthcare costs annually in the United States [34]. More than 70% of hospital-acquired bacterial infections are currently resistant to at least one of the major antibiotics used as standard treatment [201], and patient deaths as a result of hospital-acquired infections have increased by over 675% during the past 20 years [156]. As the incidence and spectrum of AR increases [14], antibiotic development has slowed to a trickle [197], further hindering the effectiveness of antibiotics to treat infectious disease. The majority of acquired AR genes in bacterial pathogens are obtained via horizontal gene transfer (HGT) [158], likely with environmental origins [24, 222, 52]. Accordingly, an increasing impetus has been placed on cataloging AR reservoirs [5, 6], determining how AR genes are enriched and most readily transferred to the clinic [141, 191], and identifying resistance mechanisms heretofore unseen in clinical settings [105, 203].
1.1.1 Traditional mechanisms of antibiotic resistance

An AR phenotype can be either an acquired trait or intrinsic to the bacterium in question. In the case of intrinsic resistance, antibiotic therapy is rendered ineffective due to a pre-existing physiological trait of the species, such as reduced accessibility to or absence of a drug target [4, 143, 50, 48]. Examples of intrinsic AR include vancomycin tolerance in Gram-negative bacteria (the outer cell membrane reduces access to the peptidoglycan target) [15] and biofilm formation in numerous organisms (perhaps providing resistance via reduced drug penetration and/or altered microenvironments) [200]. Conversely, acquired AR stems from the expression of a specific resistance gene and is commonly the result of de novo mutation or the acquisition of resistance-conferring genes on mobile genetic elements (e.g., plasmids, transposons, integrons) [214]. The AR genes present in a microbial community that are capable of transfer to a new host are collectively referred to as the “transferable resistome”. Intrinsic resistance is, by definition, limited to the context of the parent organism, whereas acquired resistance represents a more flexible phenotype, and its prevalence is more immediately responsive to selection pressure [143]. As nearly all infectious bacteria were antibiotic-susceptible prior to the introduction of antibiotic therapy [95, 50], the exceeding majority of resistance in human pathogens is acquired, either through mutation or HGT [4]. This resistance represents a diversity of biochemical mechanisms that break down into three general categories [213, 214]: (i) inactivation of the antibiotic, (ii) reducing intracellular antibiotic concentration through efflux or permeability barriers, and (iii) altering the cellular target of the antibiotic, reducing their association.

Perhaps the most intuitive of resistance mechanisms, antibiotic inactivation, is subcategorized into two groups: enzymes that inactivate drugs via degradation (e.g., the beta-lactamases) versus those that function via chemical modification. The beta-lactamases are characterized by their ability to cleave the four-membered ring present in all beta-
lactam antibiotics and are some of the best-studied and widely-distributed AR genes (for review, see [103]). These enzymes confer high-level AR and are found associated with mobile DNA elements and integrated into bacterial chromosomes. Beta-lactamases function via either a serine active site or metal cation cofactor [103] and can be found across bacterial phyla. Antibiotic-modifying enzymes are also phylogenetically widespread, as well as mechanistically diverse. These enzymes can confer tolerance towards numerous drugs, including the aminoglycoside [51], tetracycline [223], amphenicol [183], and macrolide-lincosamide-streptogramin [219] antibiotics, typically functioning via covalent modification of the drug with some functional moiety (e.g., acetyl, phosphoryl, nucleotidyl, glycosyl, and hydroxyl groups) [4].

The intracellular concentration of any given antibiotic can be reduced by either efflux mechanisms to remove the drug from the cytosol or permeability barriers that limit the drug's uptake. Many antibiotics have poor activity against Gram-negative pathogens due to efflux systems [127], most notably the RND superfamily transporters [129, 130]. Other major families of efflux systems include the MFS, SMR, and ABC superfamily transporters, which are present in both Gram-negative and -positive organisms [129, 130]. Although commonly chromosomal, many efflux systems are found on plasmids and other mobile elements and can confer drug-specific, class-specific, or multidrug resistance [167]. Some permeability barriers, such as the Gram-negative outer membrane [15], represent intrinsic AR, while in other instances, permeability barriers are acquired. Examples include multidrug-resistance via the altered expression of Gram-negative porin proteins (e.g., OmpF in Escherichia coli and OprD in Pseudomonas) [54] and glycopeptide resistance due to thickened Gram-positive cell walls [46].

Antibiotic resistance via cellular target modification often occurs via chromosomal mutation and represents a common means by which the fluoroquinolone, sulfonamide, and trimethoprim antibiotics, among others, are tolerated [4]. In the case of the fluoroquinolones, mutations to a variety of residues within the quinolone-resistance-determining-
region (QRDR) of the DNA gyrase GyrA or topoisomerase IV ParC/GrlA prevent the interaction of the synthetic antibiotic with its target, facilitating resistance [94]. Importantly, resistance to fluoroquinolones via mutation is typically recessive, suppressing the acquisition of a resistant phenotype in the presence of wild-type GyrA or ParC/GrlA and thus preventing widespread horizontal dissemination of resistant gene variants [221, 196]. However, a plasmid-borne GyrA protection protein, Qnr, has been discovered that confers low-level fluoroquinolone resistance [209] and can potentiate the incidence of QRDR mutations, which combined provide high levels of resistance [103]. Both the sulfonamides and trimethroprim competitively inhibit enzymes within the folate biosynthesis pathway: mutations to the dihydropteroate synthase and dihydrofolate reductase enzymes can reduce affinity for sulfonamides and trimethroprim, respectively, and provide tolerance to the antibiotics [100]. In addition to arising de novo, both sulfonamide- and trimethroprim-resistant enzymes are present on mobile DNA elements, providing resistance to numerous bacteria via HGT [4]. Other highly mobile mechanisms of target-modification include vancomycin resistance via the modification of peptidoglycan precursors [30], aminoglycoside resistance via methylation of the 16S rRNA subunit [77], and macrolide resistance from 23S rRNA methylases [225].

1.1.2 Strategies for combatting antibiotic resistance

The rapid evolution and expansion of AR in pathogenic bacteria has made treating infectious disease, while allowing growth of healthy commensal microbes, particularly challenging. Pathogenic bacteria have proven to be exceptionally resilient, continually evolving resistance to every antibiotic that has been deployed against them within a short period after introduction [34]. For a brief period the rise in AR was matched by the development of new antibiotics, but in recent decades antibiotic development has not kept pace [91], resulting in the need for new interventions for treatment of infectious disease.
Three major strategies have been proposed to keep our current arsenal of antibiotics relevant: (i) synthetic tailoring of antibiotic side groups \([110, 214]\), (iii) antibiotic combinations \([136, 92, 19]\), and (iii) antibiotic cycling \([102, 124]\). Each of these strategies is applied to eliminate pathogens, but they often have high collateral damage, disturbing the entire gut microbiota community. Synthetic tailoring is the modification of side groups in an antibiotic molecule to extend its effectiveness or circumvent AR, while maintaining the core antibiotic mechanism \([214]\). For example, the beta-lactams are a class of antibiotics that have undergone several levels of modification since penicillin was first discovered. The effectiveness of penicillin is confined to specific Gram-negative bacteria, but synthetic tailoring generated new antibiotics like piperacillin and methicillin that have expanded activity \([110]\). The major limitation of synthetic tailoring is that it does not change the fundamental mechanism of the antibiotic. While synthetic tailoring can bypass some types of AR, for others a single AR gene can give resistance to an entire class of antibiotics, and the latter form of resistance gene are increasing in prevalence. For instance, the recently discovered, plasmid-born \(kpc\) and \(ndm\) genes encode for enzymes that can degrade all types of beta-lactams, and prevalence of these genes continues to increase in hospitals worldwide \([108, 212]\).

Antibiotic combinations have been used to treat organisms for which a single antibiotic treatment is insufficient, and successful combinations often exhibit synergy between the constituents. In a synergistic interaction, the effectiveness of the combination of drugs at a given concentration is greater than the effectiveness of either antibiotic on its own at that concentration. The major advantage of synergistic combinations is that they lower the total drug concentrations needed for killing, which can reduce the toxicity of the treatment to human cells \([136, 19]\). Unfortunately, while lower total concentrations are an advantage in terms of toxicity, they are a disadvantage in terms of evolution of resistance, as bacteria evolve resistance more quickly to synergistic drug combinations than to the drugs used singly \([92]\). This occurs because individual drugs are dosed at sub-therapeutic concentra-
tions, resulting in bacteria facing a lower evolutionary barrier to become resistant to each component of a combination. Once a bacterium evolves resistance to one component, the synergy is broken and other components are no longer at killing concentrations, overall increasing rates of evolution of resistance. This downside has been explicitly demonstrated for two-drug combinations [92], but the theoretical principle may also apply for higher-order synergistic compound combinations.

While antibiotic cycling is an established concept [60, 83, 173], the idea has received renewed recent interest in the context of the phenomenon of collateral antibiotic sensitivity. An antibiotic is considered to confer collateral sensitivity if resistance evolved to that antibiotic makes a bacterium more susceptible to another antibiotic, compared to the wild-type population [124]. In some cases, two antibiotics can be reciprocally collaterally sensitive, where resistance evolved to either antibiotic increases susceptibility to the other. In this case it has been proposed one antibiotic could be applied until resistance to that antibiotic is manifested and then treatment switched to the other [102]. This cycling process could be repeated until the infection is cleared, and since exposure to each antibiotic selects for susceptibility to the other, there would be no net evolution of resistance. This procedure holds much promise, but cycling based on reciprocal collateral sensitivity has yet to be implemented clinically, and it is not known how generalizable collateral sensitivities are between species or even strains.

1.1.3 Reservoirs of transferable antibiotic resistance

Research on AR over the past seventy years has focused on traditionally pathogenic bacteria isolated in a clinical setting and the role of AR genes already present in those species [195, 50]. The resistance phenotype, however, is an ancient function of environmental bacteria [52], despite being largely absent from human pathogens prior to the antibiotic age [97, 95], with estimates that beta-lactamases have existed for over 2
billion years [87, 88]. Moreover, diverse mechanisms of AR have been discovered in nearly all environments [53, 6, 50, 222]; seemingly each new metagenomic study uncovers numerous examples of resistance genes previously unreported in public databases [5, 195, 62, 71]. In short, the diversity and abundance of AR in commensal microbiota and environmental settings dwarfs that which is seen in the context of human pathogens.

Importantly, environmental AR is not only widespread, but also represents the likely origins of the resistance seen in human pathogens. It has been known for 40 years that environmental bacteria share the same resistance mechanisms as those seen in pathogens [24], with documented examples of environmental resistance genes moving from natural settings into human pathogens [165, 166]. It is becoming increasingly evident that human pathogens and environmental organisms share AR genes; a recent study described seven resistance genes from non-pathogenic soil organisms, conferring tolerance to five antibiotic classes, with perfect identity to genes from phylogenetically and geographically diverse pathogens [71]. Given the staggering diversity of environmental resistance, high adaptability of bacteria, and strong selection pressure for AR, the question of AR is “not a matter of if, but only a matter of when” [213].

To understand when novel resistance will appear in human pathogens and perhaps diminish its impact, one must understand how new resistance genes are most frequently acquired by pathogenic bacteria. Since the answer is, most commonly, via HGT [97, 158, 4], understanding the complement of resistance genes most likely to be transferred to pathogens is crucial to predicting resistance acquisition. Although cataloging the repertoire of resistance genes on Earth remains a prohibitively large undertaking, techniques for interrogating the resistance properties of complex microbial communities exist and are being applied towards the identification of diverse and novel resistance from numerous settings. Importantly, these studies are focused not only on environmental locales, but also on the resistomes associated with human and animal microbiota [189, 193, 194]. Although AR may have its origins in the environment [24, 222, 52], the commensal resistome
Figure 1-1: Antibiotic therapy increases prevalence of AR available to pathogens harbored in the commensal microbiota. (a) Human intestinal microbiota is invaded by a pathogen. (b) With antibiotic treatment, the pathogen is eliminated along with much of the commensal community. Several strains survive by acquiring a multi-drug resistance plasmid from another community member. (c) Pathogen invades recovered community. (d) During subsequent antibiotic treatment the pathogen has an increased likelihood of acquiring the multidrug resistance plasmid.

shares many resistance genes with both pathogens [193] and environmental organisms [71] and represents a likely route through which these populations exchange resistance genes [191, 194]. Focused efforts to understand how pathogens acquire AR will require a greater appreciation for the diversity of resistance genes and in which environments and under what conditions these resistance genes are most accessible to human pathogens.

1.1.4 The commensal resistome as a resilience factor

Each of the antibiotic treatment strategies outlined above provide benefits for treating pathogens individually, but they do not adequately address the additional resilience fac-
tors available for transfer to pathogens with access to the commensal microbiome. HGT between bacteria has made a major contribution to the increase in AR, as evidenced by the worldwide spread of specific beta-lactamases [91]. The various mechanisms of HGT have been extensively reviewed elsewhere [206, 121], but one of their cumulative effects has been to greatly expand the diversity of genetic material available to pathogens, well beyond what is found in any single genome. The effects of HGT on clinical outcomes can be seen in studies of the spread of particular resistance genes, such as the CTX-M beta-lactamases [99]. The complete set of AR genes present in a microbial community is known as the ‘resistome’ and previous research has shown that the resistome of both human adult and pediatric intestinal microbiota is far more diverse than what has been seen in pathogens [195, 151]. It has also been shown the collection of resistance genes accessible to pathogens extends well beyond the human microbiota, including animal [1] and soil [71] environments. Antibiotic therapy, therefore, does not only select for increased AR in pathogens, but also increased prevalence of AR genes in microbial communities available for transfer to pathogens (Figure 1-1).

Resistance acquired through HGT poses an immense challenge to treatment of infectious disease because it decouples phylogeny from AR profile. Tests for taxonomic identity of a pathogen can return results 1-2 days faster than tests for AR in rapidly growing pathogens. For vertically inherited resistance mechanisms, taxonomic identity can provide insight into the susceptibility and treatment options, however horizontally transferred resistance genes can distort this inference. Our current antibiotic treatment strategies are not adequate to combat pathogens with access to the commensal resistome through HGT. Synthetic tailoring and combination therapy provide for selection of increased AR in the entire microbiota, and thus increase the size and diversity of resistomes. Antibiotic cycling with reciprocally collaterally sensitive antibiotics shows promise in slowing evolution by mutation, but it remains to be seen whether collateral sensitivity cycling is robust to resistance acquired by HGT. Continued research into new treatments, especially treat-
ments that can eliminate pathogens without increasing the resistome, therefore remain a pressing need.

1.2 Challenges in studying diverse reservoirs of antibiotic resistance

Traditionally, either culture-based [53] or PCR-based approaches [163] have been used to study AR properties from microbial communities. While both techniques have led to major discoveries [76], both have inherent limitations that have contributed to an under-sampling of resistance genes from diverse microbial habitats. The majority of bacteria remain recalcitrant to culturing [47] and are therefore not interrogated when culture-dependent techniques are employed. Additionally, linking a resistance phenotype to a causal genotype is a time-consuming process, often necessitating experimental scope to be limited to a small number of organisms, rather than whole communities. PCR screens are an effective means to identify or quantify resistance genes of known sequence, circumventing the need for culture [116], but are only able to detect previously described genes and often require expression cloning and subsequent experimentation to verify function. In addition, annotation of AR genes in shotgun-sequenced microbial communities has proven challenging, as homology-based functional gene comparisons often fail due to low sequence similarity to previously discovered resistance genes, and in silico analyses are unable to confirm resistance function.

1.2.1 Functional metagenomic selections

In contrast with standard techniques, functional metagenomics is a culture- and sequence-independent means of identifying transferrable AR in complex metagenomes. This method
(Figure 1-2) involves shotgun-cloning total community DNA into an expression vector and transforming the library into an indicator host (commonly the model organism *Escherichia coli*). The resulting transformants are then selected for the desired function (e.g., AR), and metagenomic DNA fragments are sequenced and annotated to identify causal resistance-conferring genes [5, 195]. Functional metagenomics offers three classical advantages for the unbiased interrogation of complex resistomes [47, 194]: (i) no need to culture organisms, (ii) no required knowledge of resistance gene sequence, and (iii) direct association between a genotype and a demonstrated resistance phenotype. Additionally, functional metagenomic selections specifically identify those genes within a metagenome capable of conferring antibiotic tolerance to the indicator host when expressed exogenously (i.e., they distinguish transferrable resistance from intrinsic resistance) [48]. Recent improvements to the throughput of functional metagenomics [71] unlock the potential for the experiments of scale needed to identify the specific sequences, and environments, most readily able to confer resistance to human pathogens, frequently represented by the opportunistic pathogen *E. coli*.

A major limitation of sequence-based metagenomics is its restriction to AR genes that are recognizable as members of a previously characterized class. Pair-wise sequence alignment or BLAST, the most commonly used annotation tool, requires a threshold of shared sequence identity for a gene to be considered a member of an established gene class. Therefore, by definition, this method limits the ability to identify novel AR genes. This creates a significant challenge in studying AR in uncultured microbiota, as previous studies have demonstrated that most reservoirs of antibiotic resistance are highly under sampled [195, 71, 72] (Figure 1-3).

Functional metagenomics identifies resistance-conferring elements without prior knowledge of the sequence, circumventing this limitation. For instance, in a selection for tetracycline resistance encoded by the human oral metagenome, Diaz-Torres et al. discovered a novel tetracycline resistance gene, *tet(37)* [57]. *In vitro* analysis indicated that *tet(37)* in-
Total metagenomic DNA is extracted from a microbial community sample, sheared, and ligated into an expression vector and is subsequently transformed into a suitable library host to create a metagenomic library. The library is then plated on media containing antibiotics inhibitory to the wild-type host to select for metagenomic fragments conferring antibiotic resistance. Metagenomic fragments present in colonies growing on antibiotic selection media are then PCR-amplified and sequenced using either traditional Sanger sequencing or next-generation sequencing methods. Finally, reads are assembled and annotated in order to identify the causative antibiotic resistance genes.
activates tetracycline, making it one of only three tetracycline resistance proteins to utilize this mechanism [205]. *tet(37)* resembles flavoproteins, oxidoreductases, and NAD(P)-requiring enzymes in sequence and conserved motifs, but has no identity to *tet(X)*, the first tetracycline-inactivating gene identified. It is therefore unlikely that computational annotation would have identified it as an AR gene [57].

Similarly, a study of soil microbiota identified a dihydrofolate-reducing gene, *Tm8-3*, that resembles 3-oxoacyl-(acyl-carrier-protein) reductases but not dihydrofolate reductase, the target of trimethoprim [208]. A screen of an activated sludge metagenome also identified six resistance genes that appear to inactivate chloramphenicol through enzymatic modification, but which share no significant identity with known chloramphenicol acetyltransferases [161]. Although these genes may become important for clinical resistance in the future, their existence would likely have been overlooked without high-throughput functional selections independent of previous sequence knowledge.

**Figure 1-3: Antibiotic resistomes are under sampled in uncultured microbial communities.** Percent amino acid identity of all functionally selected proteins and their top hit in the NCBI protein database for the soil (median identity: 50.8%; green dashed line) [72] and the adult gut (median identity: 58.5%; red dashed line) [195].
Drug efflux is a widespread resistance mechanism common to many antibiotic classes, but individual transporters identified through computational annotation methods cannot be assigned antibiotic efflux properties without functional validation. Small changes in protein structure can change the drug specificity profile of an efflux pump and confer or eliminate resistance. Therefore, functional metagenomic selections are an attractive alternative for high-throughput characterization of transporters with tentative resistance annotations [208]. For instance, in their study of soil isolate metagenomes, Forsberg et al. identified a novel gene with only low identity to a drug/metabolite transporter. Were it not for functional selection of D-cycloserine resistance, this putative transporter would never have been identified as an AR gene [71]. McGarvey et al. identified five novel MFS transporters and two ABC transporters from an urban soil that conferred resistance to tetracycline, chloramphenicol, or trimethoprim [146]. Lang et al. successfully identified pexA, a novel amphenicol MFS transporter, despite its low identity (33%) to other drug resistance transporters, including any known chloramphenicol exporters [122].

1.3 Antibiotics and the developing infant gut microbiota and resistome

The microbial communities colonizing the human gut are tremendously diverse and highly personal. The composition and function of the microbiota play important roles in human health and disease, and considerable research has focused on understanding the ecological forces shaping these communities [137, 149]. While it is clear that factors such as diet, genotype of the host, and environment play important roles in adult gut community composition [159, 224, 120], recent work has emphasized the importance of early-life assembly dynamics in both the immediate and long-term personalized nature of the gut microbiota [22, 80, 174, 13, 45, 157]. While the mature adult gut microbiota is believed to
be relatively stable, the developing infant gut microbiota is highly dynamic and prone to disruption by external factors, including antibiotic exposure [211]. Studies have revealed both transient and persistent alterations to the adult gut microbiota community resulting from antibiotic treatment later in life [134, 104, 107, 56, 190]. As antibiotics are routinely prescribed at a greater rate in the first years of life [39, 35, 228], the impact of these interventions on the developing infant gut microbiota is emerging as a key research priority [81, 211]. In addition to understanding the impact of these disruptions on the infant gut microbial architecture and related host diseases, we need to understand the contribution of early life antibiotics to the selection of AR gene reservoirs in the microbiota, and their threat to successful treatment of infectious disease. Here I review the current understanding of the developmental progression of the infant gut microbiota and the impact of antibiotic therapies on its encoded reservoir of AR genes.

Antibiotics are the most prescribed medications in neonatal and pediatric populations in the US [39, 35, 228]. In neonatal intensive care units (NICUs), ampicillin and gentamicin are prescribed twice as frequently as the next most common medication [39]. In children age 0-18, antibiotics are prescribed to more than 50% of individuals [228] and account for approximately 25% of prescriptions, with amoxicillin, azithromycin, and amoxicillin/clavulanate being the most common [35]. Antibiotic perturbation of the actively developing infant gut microbiota likely has profound impacts on human health and disease throughout life, as alteration of the gut microbiota during this timeframe may disrupt metabolic and immune development [45]. Equally important is the potential enrichment of the reservoir of AR genes (‘resistome’) available for transfer to pathogens [195], compromising treatment of infections in vulnerable populations. The phylogenetic and resistome composition of the infant gut microbiota is likely connected, yet dynamic, with gut environment and antibiotic pressure increasing opportunities for HGT [114, 191, 199]. Until recently, the response of the infant gut microbiota and its resistome to antibiotic perturbation was largely characterized by culture- or PCR-based experiments [134, 104, 107, 149],
which underestimate novel resistance genes. This response can be influenced by many factors, including antibiotic spectrum, duration, and delivery route (oral vs. intravenous), as well as microbial community composition and antibiotic susceptibility. While it is clear that antibiotics disrupt the developing gut microbiota, eliminating taxa and enriching for AR genes, we are just beginning to understand the relative contribution of each of these factors to the community-wide taxonomic and functional response to antibiotics.

1.3.1 Normal infant gut microbiota and resistome development

The normal developmental progression of the infant gut microbiota is patterned, yet highly dynamic and individual specific, and is shaped by many factors, including host physiology, genetics, diet, and environment [159, 224, 120]. Upon birth, infants are exposed to a surge of microbes that colonize the epithelial surfaces, including the gastrointestinal system. The source and composition of this inoculating bacterial community is highly dependent on gestational age at time of delivery and, for term infants, mode of delivery [61, 139, 120]. Term infants born vaginally are initially colonized by microbial communities resembling maternal vaginal microbiota (enriched in *Lactobacillus* and *Prevotella* spp.), while those delivered by caesarean section harbor communities that more closely resemble the skin microbiota (enriched in *Staphylococcus* and *Propionibacterium* spp.) [61]. For preterm infants (gestational age <33 weeks) the early gut microbiota composition resembles bacterial communities colonizing hospital surfaces and feeding and intubation tubing and are enriched in *Staphylococcus epidermidis*, *Klebsiella pneumoniae*, and *Escherichia coli* [29]. Mode of delivery in preterm infants does not appear to significantly affect the initial colonizing community and is instead hypothesized to be highly influenced by environment [184, 29]. Following initial colonization, term and preterm infant gut microbiota alike begin to increase in diversity with continual dynamic turnover in bacterial composition driven primarily by chronological age; however, specific bacte-
rial succession patterns are unique to these two populations [184, 224, 120]. The most notable difference in succession patterns between infant populations includes an enrichment in Proteobacteria at <2 weeks in preterm infants. A detailed time series of a single term infant revealed the developing infant gut microbiota is initially dominated by Firmicutes, with low levels of Proteobacterial species introduced in the first week of life and persisting as minor components (<10% relative abundance on average) throughout the first 2.5 years of life [118]. In contrast, preterm infant gut microbiota are quickly dominated by Proteobacterial species within the first week of life and maintain high levels, comprising on average >75% relative abundance of the community, throughout the first month [153, 120]. In healthy term infants there is a dramatic increase in *Bifidobacterium* and *Bacteroides* spp. within the first 6 months of life. By the end of the first year of life the infant gut microbiota begins to resemble an adult-like microbiota, reaching full maturity by 2-3 years of age [159, 118, 224]. It is still unclear if preterm infants eventually follow a similar developmental pattern once ‘caught up’ to term infants in postmenstrual age (gestational age + chronological age) or if this population is set on a unique developmental trajectory.

The functional capacity encoded in the infant gut microbiota also changes dramatically in the first year of life. In term infants, a shift is observed from lactate metabolism when diet is comprised of human milk and formula, to carbohydrate utilization upon the introduction of solid foods [118]. While the gut-associated resistome comprises epidemiologically important functions, less is known about how this reservoir of genes develops in early life. Recent studies have shown that AR genes in the infant gut microbiota are likely established within the first week of life, even in the absence of antibiotic exposure [148, 227, 151]. Most investigations of the early resistome have employed culture- or PCR-based methods [134, 104, 107, 149]. Focusing on readily culturable bacteria and previously identified AR genes vastly underestimates the diversity and abundance of AR genes in the gut microbiome [195]. To overcome these challenges, a recent study used
culture-independent methods to characterize the gut resistome of 22 healthy infants and children aged 1 month to 19 years [151]. Employing high-throughput functional metagenomic selections [162], the authors demonstrate that the healthy pediatric gut resistome is established early in life and persists throughout childhood. Of the 18 antibiotics investigated, only gentamicin demonstrated age-discrimination independent of antibiotic exposure with children >12 months of age harboring significantly higher levels of gentamicin resistance compared to younger children [151].

### 1.3.2 Early-life antibiotics and the human microbiota

Preterm or very low birth weight infants are at highest risk for antibiotic associated perturbations, as they routinely receive empiric antibiotic therapy at birth [119, 80]. As with adults, short-term perturbations of the infant gut microbiota follow soon after antibiotic treatment, the broad characteristics of which are known through culture-based methods [22]. Recently, culture-independent methods for interrogating microbial communities have emerged, relying on DNA amplification and sequencing. When applied to the developing infant gut microbiota, some studies suggest both phylogenetic diversity and microbial load are depressed following antibiotic therapy. For example, 16S rRNA-based phylogenetic profiling of fecal microbiota from preterm infants receiving ampicillin and gentamicin during the first week of life had lower diversity compared to un-treated infants [80]. However, another study comparing the fecal microbiota composition of infants treated with oral cephalexin to infants receiving no treatment did not reveal significant differences during the month following therapy [202]. These differing findings may be due to different antibiotic regimens, routes of antibiotic administration, choice of statistical analytical methods, or other uncontrolled factors. The difficulties inherent to untangling these variables informs both the need for large cohort studies of specific antibiotic regimens and studies in controlled animal models. Bacterial load is another measure found to decrease in some
studies but not in others. Quantitative PCR of the 16S rRNA gene has been used to estimate bacterial load in the gut. In unrelated studies examining the infant gut microbiota following antibiotic therapy, bacterial load was found to be unaffected, slightly altered, profoundly decreased, or even increased following treatment [159, 13]. Again, the lack of a consensus may be due to uncontrollable variables inherent to infant cohorts.

Antibiotic treatment can also target specific phylogenetic subgroups of the infant gut microbiota. Treatment of preterm infants with a variety of antibiotics, including penicillin, ampicillin, cephalaxin, gentamicin, amikacin, erythromycin, vancomycin, clindamycin, and teicoplanin, have been found to increase the percentage of potentially pathogenic Enterobacteriaceae while lowering the relative percentage of microbial taxa linked to a healthy microbiota such as Bifidobacteriaceae, Bacilli, and Lactobacillales spp. [202, 13, 80]. In mice, reproducible effects on taxa have been noted. In mice exposed to sub-therapeutic antibiotics through drinking water, no overall change in microbial load was detected, but a significant decrease in the ratio of Bacteroides to Firmicutes was observed [38]. In another study examining the consistency of phylogenetic responses to antibiotic perturbation, mice were treated with amoxicillin, metronidazole, bismuth, cefoperazone, and in combination. Under these conditions Proteobacteria, and in particular Enterobacteriaceae, dominated the intestines of the treated animals immediately after cessation of therapy, accounting for 73% of sequences. After two weeks without perturbation the microbiota of these animals returned to a low percentage Proteobacteria state (5.77%), though still higher than in untreated mice (1.2%). Treatment with cefoperazone, a broad-spectrum antibiotic, was in particular associated with loss of microbial diversity without recovery even 6 weeks post therapy [12]. In another study in which mice were administered either vancomycin or streptomycin in their drinking water, only vancomycin treatment was associated with significant reductions in both bacterial load and diversity, including depletion of Bacteroidales and marked enrichment of Lactobacillus spp. [180]. An important variable in several studies is the antibiotic route of administration. In many mouse studies,
antibiotics are provided through the most facile means available, *e.g.* through the animal’s water supply or, in the case of infant mice, through the mother via milk [38, 45]. This is in stark contrast to antibiotic administration in the NICU, where the majority of antibiotics are provided through intravenous lines [120]. A recent study in mice found significant differences when tetracycline or ampicillin were administered orally versus intravenously, highlighting the importance of this variable in evaluating the translational significance of murine model systems [226].

### 1.3.3 Long-term effects of early-life antibiotic therapy

Infants exposed to antibiotics during microbiota development may experience long-term disruptions. For example, disruptions have been noted at 90 days following treatment with a variety of antibiotics and 3 months after treatment with oral amoxicillin [174, 13]. However, some studies have found no long-term microbial disruptions due to antibiotic use in human infants [159]. In mice, lack of recovery from antibiotic treatment at 6 weeks has been noted [12] and even sub-therapeutic antibiotics have been found to have long-term effects on taxa associated with healthy microbiota such as *Lactobacillus spp.*, *Bifidobacteriaceae* (decreased abundance) and *Enterobacteriaceae* (increased abundance) [202, 45].

Early-life antibiotic therapy has been linked to a variety of host outcomes and antibiotic-disrupted taxa have been linked causally to these as well. Broadly, antibiotic therapy can enrich for potentially pathogenic and antibiotic resistant *Enterobacteriaceae*, a bacterial family commonly resistant to beta-lactam antibiotics [13, 80]. Antibiotic therapy in infants has further been linked to increased risk of developing necrotizing enterocolitis (NEC), the leading cause of morbidity in NICU infants [132]. In one study of preterm infants, empiric antibiotic therapy lasting >5 days was associated with a significantly increased rate of sepsis, NEC, and death, with an attributable risk of 32 per 100 infants [119]. Another
retrospective study of extremely low birth weight infants found that courses of antibiotics >5 days in the first days of life were statistically linked to increased risk of developing NEC and higher mortality rates. It was found that each additional day of antibiotic treatment increased the odds of an infant developing NEC by 7% or developing NEC and dying by 4% [44]. In these studies causative taxa were not identified. However, other studies have demonstrated loss of Lactobacillus and Bifidobacterium spp. and increased Enterobacteriaceae as a result of antibiotic treatment [202, 45]. Taxa from the Lactobacillaceae and Bifidobacteriaceae families have been linked to the prevention of poor outcomes in infants and are known to be important components of a healthy developing infant gut microbiota and originate from the maternal microbiome [159, 139]. Probiotic treatment of very low birth weight infants with Lactobacillus acidolphilus and Bifidobacterium infantis has been shown to reduce morbidity in these cohorts, as well as increase daily weight gain and decrease hospital stays [2, 90]. One potential mechanism of this protection is through interactions between the gut microbiota and the host immune system. Specific taxa, such as Lactobacillus spp., have been shown in model organisms to promote a healthy gut immune response and healthy modulation of the intestinal epithelial layer [109]. Perturbation of the maternal and infant gut microbiota in a murine model was also found to modulate the levels of the IL-17 cytokine, leading to increased susceptibility to sepsis [55]. Outside of infancy, early life antibiotic use has also been linked to the development of other conditions later in life. Recent studies using a murine asthma model have found evidence implicating antibiotic-induced dysbiosis in increasing asthma rates in children [180]. Similarly, antibiotics have been found to play a role in the induction of hypersensitivity pneumonitis [181]. Antibiotic treatment has also been linked to obesity. Children exposed to antibiotics in the first 6 months of life were found to have a statistically significant increase in body mass. On the other hand, children treated with other medications or antibiotics after 6 months of life showed no such correlation [210]. In another study, antibiotic exposure during the first year of life was found to be associated
with being overweight at age 12, with the association particularly strong in males [16]. Similar effects have been seen in mice under controlled conditions. In a pair of studies in which sub-therapeutic antibiotics were administered to infant mice, treatment was found to induce metabolic changes in the host, including increased adiposity, modulation of liver mechanisms for cholesterol and lipid metabolism, and increased susceptibility to a high fat diet. Furthermore, these effects were directly linked to changes in the gut microbiota, including phylogenetic composition and metabolic function, and were found to transfer following administration of an altered microbiota to a healthy host [38, 45].

### 1.3.4 Enrichment of the infant antibiotic resistome

Significant alterations in the composition of the developing infant gut microbiota in response to antibiotic treatment can cause a similar transformation in functional capacity, the most clinically relevant example being AR. When exposed to constant antibiotic challenge \textit{in vitro}, microbial communities show evolution of multidrug resistance [207] as well as population-level resistance dynamics to antibiotic stress [125]. While the routes of evolution of AR and community-level dynamics are less well known \textit{in vivo}, antibiotic therapy has been shown to select for survival of resistant members of the microbial community or for members capable of acquiring AR genes [20]. The persistence of these populations after cessation of therapy poses a long-term threat to the host as these populations can include potential pathogens as well as act as reservoirs for AR genes for transfer to pathogens [189, 126]. For example, in a pair of studies in adults, treatment with clindamycin for 7 days resulted in rapid development of resistant \textit{Bacteroides spp.}, with resistant clones constituting 15% of the clones in the treated cohort compared to 0% in the control cohort. This condition persisted through the entire 2-year study. Similarly, the macrolide resistance gene \textit{ermF} was several logs higher in treated adults than in controls and persisted for at least 2 years [134, 106]. In another study, 1,000-fold enrichment
of the macrolide resistance gene \textit{ermB} was found following treatment with clarithromycin and metronidazole, and was observed up to 4 years later even in the absence of additional antibiotic therapy [104]. While comparable studies on the effect of antibiotics in early life are lacking, one similar culture-based study examined children treated with the antibiotic amoxicillin. Surprisingly, amoxicillin resistant bacteria were found both in children with and without drug treatment. In addition, of the amoxicillin resistant isolates, approximately 50% also showed resistance to penicillin, with others also demonstrating resistance to erythromycin and tetracycline [174]. Notably, the route of antibiotic administration can strongly impact the emergence of resistant populations in the gut. Mice provided with an oral inoculum of either tetracycline or ampicillin resistant bacteria were administered each antibiotic either by oral gavage or intravenously. The expansion and contraction of the known resistance genes in the resistant bacteria were monitored by quantitative PCR. Oral administration of ampicillin was found to result in an approximate 4-log increase in ampicillin resistant gene copy number over intravenous administration, while the increase seen for oral administration of tetracycline 2-log. The difference in effect was hypothesized to be a result of how the host clears each antibiotic, with ampicillin being cleared solely through the urine and not interacting with the gut microbiota [226].

1.3.5 Future Directions

Given the exceedingly personalized nature of the human gut microbiota, I anticipate that highly sampled, longitudinal infant cohort studies combined with controlled mouse models of therapeutic levels of antibiotic treatment will begin to deconvolute the forces shaping these developing microbiota and their encoded AR genes. As we begin to understand more about the extent to which AR spreads within and between microbial ecosystems, there has been a concurrent increase in emphasis on addressing the challenge of AR from an ecological perspective [149]. Importantly, this approach requires characteri-
zation of the overall abundance and diversity of AR genes in the environment and human-associated microbial communities [78, 155]. Using culture-independent metagenomic and functional metagenomic techniques, recent studies have shown the human gut microbiota to be an extensive reservoir of AR genes [195, 96], the abundance of which has been broadly correlated with antibiotic use practices by country [73]. While a number of studies described above have demonstrated a significant response of specific drug-resistant strains or specific AR genes to antibiotic therapies using culture or PCR-based methods, the effect of antibiotics on community-wide AR remain unclear. Functional metagenomic studies of AR genes harbored in the guts of healthy infants reveal high potential for mobilization and overall disconnection between AR gene and bacterial host [151], suggesting a much more complicated relationship between the community composition and functional response to AR in the developing gut. Integration of culture-independent methods for community-wide investigation of corresponding community composition and functions, such as metagenomic functional selections [162] combined with marker or shotgun DNA sequencing [72], will be essential in filling in the current gaps in our system-wide understanding of the effects of antibiotics on developing infant gut microbiota and resistomes.

1.4 Scope of thesis

Reservoirs of AR genes available for transfer to pathogenic bacteria pose a severe threat to treatment of infectious disease. AR in the clinic has been hypothesized to have environmental origins, however, metagenomic studies of how AR moves between ecologies as well as how reservoirs of AR develop are severely limited by the ability of current computational methods to accurately identify diverse and sequence divergent AR genes. In Chapter 2, I describe the development and application of Resfams, an AR-specific pHMM database, combined with functional metagenomic sequencing and whole genome sequencing to the study of diverse microbial communities in both the soil and the human
gut environments. I show that these two habitats harbor significantly distinct AR functions. Further, once identified, many AR genes with divergent functions contain similar or identical protein domains, requiring identification of high information content regions within the protein for quantification of specific AR families within short read metagenomic dataset. In Chapter 3, I describe the development and application of ShortBRED, a database of short high-information content protein markers, to studying AR in the human gut microbiota. We demonstrate that ShortBRED both increases specificity and speed of AR detection and quantification within microbial communities.

Antibiotic treatment has the potential to significantly disrupt microbial community structures and functions, including the ‘resistome’. This is especially true during human development, where the gut microbiota is relatively unstable and metabolic and immune systems are in a critical developmental timeframe. Infants born prematurely almost universally receive early and prolonged antibiotic therapy, however, it is currently unclear if and how these treatments disrupt microbial community composition and encoded reservoirs of AR. In Chapter 4, I apply high-throughput experimental methods (shotgun metagenomic sequencing and functional metagenomic selections for AR) combined with the described computational methods (Resfams and ShortBRED), to study the developing preterm infant gut microbiota in the first 2.5 months of life. Through this work, I showed that a subset of specific antibiotic therapies (meropenem, cefotaxime, and ticarcillin-clavulanate) result in significant reduction in species richness as well as significant enrichment or depletion of specific bacterial taxa and AR genes. In contrast, vancomycin and gentamicin, two of the antibiotics most commonly administered to preterm infants, result in a varied response in species richness. We show this response is predictable with 85% accuracy based on the relative abundance of only two bacterial species and two AR genes prior to treatment.

This body of work makes several contributions to the field of microbial ecology and antibiotic resistance. First, the development of computational methods for high-throughput analysis of antibiotic resistomes encoded by diverse microbial communities. Second, this
study demonstrates that while there are documented sources of transfer of AR genes from soil organisms to the clinic, bacteria in soil microbiota and human gut microbiota generally resist the same antibiotics through district mechanisms. Finally, we show the impact of specific antibiotic therapies on both the gut microbiota composition and function in a clinical cohort of preterm infants. Demonstrating a predictable response of the gut microbiota to antibiotic treatment based on the pre-antibiotic composition and function is the first step to precision medicine for the gut microbiota based on an individual’s pre-treatment gut microbiota state.

1.5 Acknowledgements

This chapter was synthesized from work done in collaboration with Mitch W. Pesesky, Terence S. Crofts, Erica C. Pehrsson, Kevin J. Forsberg, Sara Ahmadi, and Gautam Dan-tas. Sections of this chapter were published in Gibson, Pesesky et al. (J Mol Biol, 2014), Gibson, Crofts et al. (Curr Op Microbiol, In Press), and Pehrsson, Forsberg, Gibson et al. (Front. Microbiol, 2013).
Chapter 2

Improved Annotation of Antibiotic Resistance Determinants Reveals Microbial Resistomes Cluster by Ecology

2.1 Abstract & Introduction

Antibiotic resistance is a dire clinical problem with important ecological dimensions. While antibiotic resistance in human pathogens continues to rise at alarming rates, the impact of environmental resistance on human health is still unclear. To investigate the relationship between human-associated and environmental resistomes, we analyzed functional metagenomic selections for resistance against 18 clinically relevant antibiotics from soil and human gut microbiota as well as a set of multidrug resistant cultured soil isolates. These analyses were enabled by Resfams, a new curated database of protein families and associated highly precise and accurate pHMMs, confirmed for AR function and orga-
nized by ontology. We demonstrate that the AR functions that give rise to the resistance profiles observed in environmental and human-associated microbial communities significantly differ between ecologies. AR functions that most discriminate between ecologies provide resistance to beta-lactams and tetracyclines, two of the most widely used classes of antibiotics in the clinic and agriculture. We also analyzed the AR gene composition of over 6,000 sequenced microbial genomes, revealing significant enrichment of resistance functions by both ecology and phylogeny. Together, our results indicate that environmental and human-associated microbial communities harbor distinct resistance genes, suggesting that AR functions are largely constrained by ecology.

Multi-drug resistance in clinical pathogens continues to rise, while the pipeline for new antibiotic development and approval continues to dwindle [198]. Consequently, we face the prospect of returning to a pre-antibiotic era, where an increasing number of infections can no longer be treated effectively with our current arsenal of drugs. While it is clear that dissemination of AR extends beyond the clinic [71, 222] and includes many routes through agricultural and environmental microbial communities, the full impact of environmental AR on human health is still unknown [68]. A deep quantitative understanding of the ecological relationship between environmental and human-associated microbial resistomes is necessary to evaluate the relative importance of these diverse ecologies in AR gene acquisition by human pathogens.

Hindering our ability to study the ecology and transmission of AR genes between environmental and human-associated reservoirs is the difficulty in accurately identifying AR functions from sequence alone. This is emphasized by continual identification of sequence-novel AR genes in almost every microbial community, including soil [7, 71, 177, 208], activated sludge [152, 161], human gut and oral microbiomes [37, 57, 58, 195], and animal gut microbiomes [115]. This enforces the need for functional validation in studies of AR transmission [162] and for improved methods to identify sequence-divergent AR determinants in silico.
In order to quantitatively analyze the relationship between environmental and human-associated resistomes, we developed Resfams, a curated database of protein families and associated pHMMs, organized by ontology and confirmed for AR function. pHMMs have been widely adopted for improved functional annotation of general functions in microbial genomes and metagenomes [140, 147]. However, they have not yet been specifically applied to AR functions in microbial communities or genomes. Once developed and validated, Resfams pHMMs were applied to quantitatively understand the relationship between human-associated and environmental resistomes using both functional selections to 18 antibiotics from the soil and human microbiota as well as analysis of over 6,000 microbial genomes representing diverse phylogenies and habitats.

2.2 Results & Discussion

2.2.1 Optimization of antibiotic resistance profile hidden Markov models (Resfams)

Following extensive curation and functional validation, the full Resfams pHMM database contains 166 pHMMs representing all major AR gene classes, including AR genes against beta-lactams, aminoglycosides, fluoroquinolones, glycopeptides, macrolides, tetracyclines, as well as efflux pumps and transcription factors modulating AR. To improve Resfams prediction accuracy, we optimized profile specific gathering thresholds, which set an inclusion bit score cut-off for a protein sequence alignment on a profile-by-profile basis [169]. We achieved perfect precision and high recall (99 ± 0.02%) of all AR proteins used to train the Resfams pHMMs (Supplementary Figure 2-S1B). We tested the prediction accuracy of these optimized Resfams families on AR proteins from the Antibiotic Resistance Database (ARDB) [133] not used in training of the original pHMMs. We predicted the AR function
Figure 2-1: Annotation of functional metagenomic selections using Resfams. Unique full length ORFs (>90 a.a.) annotated as AR genes from functional metagenomic selections using Resfams pHMM database (blue) compared to hand curation (black) and BLAST to AR databases (red) from (a) MDR cultured soil bacteria and (b) human gut microbiota. The Resfams pHMM database identified no false positive annotations as measured by the hand-curated gold standard for cultured soil bacteria. For the amoxicillin selection, MetaGeneMark predicted one ORF in the hand-curated set as two independent ORFs and both were correctly identified by Resfams.

of 2,454 unique sequences, representing 54 Resfams protein families, and achieve perfect precision and high recall (98 ± 0.03%) for all protein families (Supplementary Figure 2-S1C). These recruited protein sequences were subsequently incorporated into the corresponding Resfams protein families, resulting in the final database of AR-specific pHMMs used for all further analysis in this study.

2.2.2 Resfams accurately predicts novel resistance functions from sequence alone

The prediction sensitivity and specificity of Resfams was evaluated using recent functional metagenomic studies investigating the resistomes of cultured soil microbiota and human gut microbiota [151, 71, 195]. We compared Resfams pHMMs to pairwise se-
sequence alignment (BLAST) against the ARDB [133] and the Comprehensive Antibiotic Resistance Database (CARD) [144] databases for their ability to predict AR function. Resfams demonstrated improved sensitivity: 64% of AR genes identified using Resfams in both the soil and the human gut microbiota were not identified by BLAST (Figure 2-1). Focusing on the hand-curated, gold standard set of AR genes identified in the set of 95 multidrug resistant (MDR) cultured soil isolates, we show that Resfams was able to predict over 95% of the full-length AR genes (>90 amino acids) (Figure 2-1A). In contrast, pairwise sequence alignment to AR specific databases using widely accepted identity thresholds for AR genes [133] annotated less than 34% of full-length AR genes. Importantly, Resfams did not identify any false positive AR proteins (i.e. not predicted by intensive hand-curation), ensuring that AR potential of microbial communities is not over-estimated. Generating the gold standard set of AR genes involved extensive hand-curation, including functional characterization, phylogenetic analysis, and primary literature validation to identify causative AR genes. This time-intensive process is prohibitive for large-scale studies of AR potential of microbial communities. In comparison, Resfams enables automated, rapid, accurate, high-resolution predictions of AR genes. As a demonstration of Resfams high-resolution and high-specificity AR functional predictions, we focused on beta-lactamases, one of the most widely disseminated and clinically relevant class of AR genes [50]. In the previously discussed soil and human gut functional metagenomic datasets, 113 unique, full-length beta-lactamase proteins were predicted by Resfams, which were not predicted as an AR gene by pairwise sequence alignment to AR databases (representing 45% of all identified beta-lactamases). beta-lactamases are commonly classified into four molecular classes (Ambler Classes A, B, C, and D) based on primary structure [10], with over 1,000 unique AR-related beta-lactamases identified to date [50]. We accurately predict the molecular class or subclass for over 60% of the novel beta-lactamases from the soil (Figure 2-2A) and over 80% of the novel beta-lactamases from the human gut (Figure 2-2B). Importantly, all of these predicted class or sub-class
Figure 2-2: Resfams improves the resolution of annotation of antibiotic resistance genes. The majority of beta-lactamases identified from (a) MDR cultured soil bacteria and (b) human gut microbiota in functional metagenomic selections that are highly sequence divergent from any gene in the ARDB or CARD databases (< 80% a.a. identity over 85% of target sequence) are annotated according to Ambler class (red) or subclass (cyan) level by Resfams. (c) All highly divergent beta-lactamases from (a) and (b) that are annotated at the Ambler class and subclass level (colored branches) by Resfams accurately cluster on a phylogenetic tree with previously verified beta-lactamases from all four Ambler classes (black branches). Pie charts in (c) depict the fraction of Resfams identified beta-lactamases and previously verified beta-lactamases represented on the phylogenetic tree in each Ambler class clade.

beta-lactamases clustered with previously reported beta-lactamases of the same molecular class on a phylogenetic tree (Figure 2-2C), confirming that Resfams accurately categorized these beta-lactamases by sequence class where conventional computational methods were unable to even predict general AR function.

2.2.3 Antibiotic resistomes cluster by ecology

Environmental, non-pathogenic, and commensal organisms have long been shown to harbor functional AR genes [23, 142, 177]. However, it was only recently established that exchange of multiple classes of AR genes has occurred between non-pathogenic environmental bacteria and human pathogens [71]. These findings indicate that environmental
bacteria serve as potential reservoirs of AR genes primed for exchange with pathogenic bacteria. This motivates a high-resolution characterization of the AR genes distinct to and shared between environmental and human-associated microbial communities. Analyses that employ Resfams protein families are capable of addressing this goal by reducing bias in comparisons across ecological barriers. As a demonstration, we re-annotated functional metagenomic selections against 18 antibiotics from uncultured human gut, soil microbiota, and MDR soil dwelling cultured-isolates using Resfams. To compare the diversity of underlying AR functions that give rise to observed AR profiles across samples and ecologies (Supplementary Figure 2-S2), a count matrix of unique gene sequences per Resfams AR family identified in each resistome was generated by summing across a subset of five antibiotic selections included in all three datasets (selections encompass four major classes of antibiotics: beta-lactams, tetracyclines, amphenicols, and glycopeptides). While previous studies have predicted an enrichment of AR genes in the human gut microbiota [96], we found no significant differences in the number of distinct AR genes between the human gut and soil microbiota when screened for functional AR. Resistomes cluster by ecology using both Bray-Curtis and Jaccard distance metrics (p<0.001, ANI-SOM; Figure 2-3, Supplementary Figure 2-S3), suggesting that the soil and human gut microbiota consist of functionally distinct AR gene architectures.

In order to determine the Resfams families that most discriminate antibiotic resistomes between ecologies, we used the supervised learning technique Random Forests [117]. This analysis revealed that the separation of soil and gut resistomes is driven mainly by beta-lactamase and tetracycline resistance functions (Figure 2-3). Subclass B3 beta-lactamases (Class B) are mainly associated with soil resistomes, while CbIA (Class A) and CMY/LAT/MOX (Class C) beta-lactamases are found mainly associated with the human gut. CbIA beta-lactamases are species-specific cephalosporinases that have been primarily identified in Bacteroides uniformis, a common resident of the human gut microbiota [192]. While this particular Class A beta-lactamase uniquely discriminates between
Figure 2-3: Resistomes differ by ecology. (a) PCoA plot depicting Bray-Curtis distances between resistomes of the soil (green), human gut (magenta), and MDR soil isolates (blue), calculated using unique ARG counts. Resistomes of different ecologies cluster separately (p<0.001, ANOSIM). Function biplot coordinates (squares) represent the weighted average of the top six most discriminating ARG functions between ecologies across all samples. Separation of resistomes is heavily influenced by beta-lactamase (orange squares) and tetracycline resistance functions (yellow squares). (b) Bipartite network diagram of normalized ARG counts across all resistomes. Edges connect sample nodes (squares) to ARG function (circles). Edges and sample nodes are colored by sample ecology (green, soil; magenta, human gut; blue, MDR soil isolates) and ARG functions are colored by extent of sharing across ecologies (white, unique to ecology; gray, shared between two ecologies; black, shared across all three ecologies). Inset pie chart represents the percentage of AR Resfams families that belong to each group.
soil and human gut resistomes, Class A beta-lactamases in general are prevalent across all environments and samples (94% of soil resistomes and 89% of human gut resistomes). The low ratio of Class C beta-lactamases to Class B beta-lactamases in the soil supports previous findings in studies of the soil resistome [7]; however, our results indicate that this trend is opposite in the human gut resistome (Class C/Class B ratio: 0.42, soil; 4.3, gut). While Class A and Class C beta-lactamases have long been considered the most clinically important classes of beta-lactamases, Class B beta-lactamases are a growing concern in the effective treatment of infectious disease [175], exemplified by the NDM-1 carbapenemase disseminating in MDR pathogens and environmental habitats [215]. Therefore, even though resistomes appear to be largely constrained by ecology, our results continue to emphasize the importance of environmental reservoirs of AR in the emergence of novel clinical resistance, particularly in this important class of AR genes. In addition to Resfams families that discriminate resistance reservoirs, we sought to determine if there was a core resistome across all samples and across all habitats. Highlighting the extreme diversity of AR genes, there was no single Resfams family that was shared across all samples, only two Resfams families were shared across >50% of metagenomic samples (Class A beta-lactamases and MFS antibiotic efflux), and only six Resfams families were found in at least one sample from every habitat investigated (beta-lactamase, chloramphenicol acetyltransferase, Class A beta-lactamases, Class C beta-lactamase, MFS antibiotic efflux, and TetA).

2.2.4 Antibiotic resistance potential encoded in microbial genomes

To understand the relative impacts of phylogenetic origin and ecological factors in shaping AR reservoirs, a greater appreciation for the relationship between AR function and bacterial community composition is needed. Accordingly, we used Resfams pPHMMs to predict the AR potential encoded in 6,179 microbial genomes from the Integrated Mi-
crobial Genomes (IMG) database [140], representing diverse phylogenies and habitats. We summed all Resfams counts in a genome by the AR mechanisms and calculated enrichment of a particular mechanism in (1) phyla, (2) habitat, as well as (3) phyla by habitat (Figure 2-4). Importantly, these results provide confirmation that Resfams accurately annotates AR function. For example, as was observed with functional selections, we found no significant difference in the percentage of total AR functions encoded in microbial genomes between habitats (Supplementary Figure 2-S4), which contradicts previous predictions using pairwise sequence alignment annotation methods [96]. Further confirmation can be seen by examining resistance to glycopeptides, a class of antibiotics only active against Gram-positive bacteria due to their large molecular size which prevents their transport across outer membrane porins of Gram-negative bacteria [168]. Glycopeptide resistance mechanism distributions across phyla reflect this AR pressure as they were predicted exclusively in Gram-positive organism genomes.

While AR is present in nearly all microbial genomes (84% of microbial genomes investigated encode at least one AR gene), there is significant AR mechanism enrichment by bacterial phyla and habitat (Figure 2-4). For example, resistance to beta-lactams, one of the most clinically important classes of antibiotics, is significantly enriched in Actinobacteria relative to other phyla (P<0.01, Fisher’s exact, Figure 2-4A), when summed across habitats. This is consistent with the Actinomycete class of bacteria being responsible for synthesis of the vast majority of natural beta-lactam antibiotics, therefore requiring self-resistance. In addition, beta-lactamases are enriched in soil bacteria versus other habitats (P<0.01, Fisher’s exact, Figure 2-4B), consistent with the vast majority of beta-lactam producing bacteria originating in the soil. However, our results show that beta-lactamase resistance genes in the soil are distributed across all phyla with no significant enrichment in Actinobacteria. This suggests that many soil bacteria, regardless of phylogeny, have evolved to confer resistance to beta-lactams, revealing a strong habitat by phylogeny relationship. As resistance to tetracyclines represented some of the most discriminating
Figure 2-4: Resistomes annotated by Restams across phylogeny and habitats of 6,179 sequenced bacterial isolate genomes. Binary heatmaps of genomes organized by (a) phylogeny, (b) habitat, and (c) phylogeny within habitat. Sections of the heatmaps are colored if a particular AR mechanism is significantly enriched within a particular phyla or habitat (p < 0.01, Fisher’s exact). Enrichment of beta-lactamase Ambler class and tetracycline resistance functions is depicted across (d) phyla and (e) habitat (*p<0.01, Fisher’s exact).
Resfams functions between the soil and the human gut microbiota from our functional metagenomic analyses (Figure 2-3), we were prompted to further investigate tetracycline resistance mechanisms encoded by microbial genomes (Figure 2-4D,E). There are three known mechanisms of tetracycline resistance: MFS efflux, ribosomal protection, and drug inactivation. Our results suggest that the mechanism by which bacteria resist tetracycline antibiotics is heavily biased by habitat. Soil bacteria are significantly enriched for tetracycline MFS efflux pumps, while human-associated bacteria are significantly enriched for tetracycline ribosomal protection genes (P<0.01, Fisher’s exact). These results are consistent with our findings from functional metagenomic selections (Supplementary Figure 2-S5). Conversely, pairwise sequence alignment to AR specific databases incorrectly predicts enrichment of all tetracycline resistance mechanisms in the human gut versus soil (Figure 2-5B). Finally, we observed that pathogenic organisms are enriched for all surveyed AR mechanisms, excepting antibiotic inactivation and beta-lactamases (Supplementary Figure 2-S6). Beta-lactamases have long been recognized as one of the most widely distributed mechanism of AR and were found to be encoded in over 60% of bacterial genomes. Our results, however, emphasize that beta-lactamases are commonly found in non-clinical environments and commensal organisms, which has implications for understanding the further spread of beta-lactamases to pathogens via non-clinical AR reservoirs [99, 215].

The described resistome analysis would yield significantly different conclusions if microbial genomes were analyzed using pairwise sequence alignment to AR specific databases. Not only would the total resistance potential of microbial genomes be significantly lower (p<0.001, Wilcoxon Rank Sum Test), but also the phylogenetic and ecological distribution of AR mechanisms would be significantly biased towards more heavily studied human-associated environments (Figure 2-5). Using the Resfams pHMM database, we predict that a number of AR functions are enriched in the soil, which is consistent with evidence that antibiotic producers are primarily soil dwelling and that AR is ancient, having evolved
Figure 2-5: Annotation of antibiotic resistance across habitats using Resfams compared to pairwise sequence alignment. Predicted enrichment of beta-lactamase Ambler classes and tetracycline resistance mechanisms (*p<0.01, Fisher’s exact) in sequenced genomes across habitats using (A, B) Resfams family pHMMs compared to (c, d) BLAST to the ARDB and CARD databases.
in soil for over the past 30,000 years [52]. Again, this result is not recapitulated using pairwise sequence alignment to AR specific databases (Figure 2-5).

2.2.5 Discussion

Using Resfams, we show that antibiotic resistomes cluster by ecology, with no core resistome shared between all samples. For example, while all communities display resistance to tetracycline, soil bacteria mainly resist tetracycline through MFS antibiotic efflux, while bacteria in the human gut microbiota typically resist tetracycline via ribosomal protection mechanisms. Importantly, these results are consistent between both the functional metagenomic datasets as well as the sequenced microbial genomes. Our ability to accurately identify and annotate AR functions in microbial genomes and communities has important implications for our ability to fight infectious disease. It improves our understanding of the evolution, ecology, and transmission of AR in pathogens, as well as has a direct impact on clinical diagnostics. Currently, the most common method used to characterize resistome composition from sequencing data is pairwise sequence alignment to AR specific databases [73, 96]. This approach biases towards human-associated organisms, vastly underestimating the potential impact of environmental resistance reservoirs on AR in pathogens. In order to address this problem, we developed and benchmarked a set of AR specific gene families (Resfams) and associated pHMMs and applied them to functional metagenomic datasets from the soil and human gut microbiota as well as to over 6,000 sequenced microbial isolate genomes representing diverse phylogenies and habitats. By using a consensus model approach, we are able to significantly increase our ability to characterize highly diverse and under studied reservoirs of resistance while minimizing bias. In order for the full potential of Resfams AR protein families to be realized, in depth functional validation of genotype to phenotype predictions are necessary. In addition, Resfams AR protein families need to be continually updated and maintained in
order to keep up with rapidly evolving bacterial AR. These challenges emphasize the im-
portance of continued functional investigation of environmental and clinical AR reservoirs,
and for these investigations to be intimately connected to the improvement of methods for
annotation of AR phenotype from genotype.

2.3 Materials & Methods

2.3.1 Training and Benchmarking of Resfams pHMMs

Resfams AR family pHMMs were built by (1) generating a multiple sequence alignment
for each AR family using MUSCLE [65] v3.8.31 with default parameters and (2) training
pHMMs using the hmmbuild function of the HMMER3 [70] software package using default
parameters. Gathering thresholds (GA) were added to the pHMMs by using a test set
of known AR genes and optimizing precision and recall metrics. Resfams pHMMs were
first trained using 2,097 unique AR protein sequences from the CARD database [144],
the Lactamase Engineering Database (LacED) [204], and Jacoby and Bush’s collection
of curated beta-lactamase proteins (http://www.lahey.org/Studies/).

This core database of 119 pHMMs was supplemented with an additional 47 pHMMs
from the Pfam [21] and TIGRFam [86] databases to generate the full Resfams pHMM
database, resulting in a total of 166 AR specific pHMMs. The 47 additional pHMMs in-
cluded in this version of Resfams have been curated and demonstrated to identify pro-
tein families that commonly contribute to AR, such as acetyltransferases, AraC transcrip-
tional regulators, Major Facilitor Superfamily (MFS) transporters, ATP-binding cassette
(ABC) efflux pumps, etc. These supplementary pHMMs were verified using functional
assays, including functional metagenomic selections of the soil microbiota and the hu-
man gut microbiota. The full version of the Resfams database is only utilized when there
is previous functional evidence of AR activity, such as functional metagenomic selections. All versions of the Resfams database and supporting datafiles are available at http://dantaslab.wustl.edu/resfams.

A curated list of AR proteins used to generate Resfams pHMMs, along with associated reference gene sequences, is available at http://dantaslab.wustl.edu/resfams. These proteins were compiled using CARD [144], ARDB [133], the Lactamase Engineering Database (LacED) [204] and Jacoby and Bush’s collection of curated beta-lactamase proteins (http://www.lahey.org/Studies/). All proteins were hand-curated to ensure functional homology using phylogenetic analysis and literature searches, eliminating all incorrectly annotated and truncated protein sequences.

Resfams pHMMs were trained using CARD [144], LacED [204], and Jacoby & Bush’s collection of beta-lactamases while retaining ARDB [133] as an independent test set. Gathering thresholds were optimized such that precision and recall metrics on this independent set of known genes were as close to 1.0 as possible. Precision was calculated as (the total number of correct annotations)/(total number of annotations). A precision of 1.0 for a pHMM indicates that all annotations of proteins recruited by that pHMM in the ARDB test set matched the annotation of the pHMM. Recall was calculated as (the total number of correct annotations)/(the total number of proteins in ARDB with annotations matching the annotation of the pHMM). A recall of 1.0 for a pHMM indicates that all proteins contained in the ARDB test set that matched the annotation of the pHMM were correctly recruited by the pHMM.

### 2.3.2 Resfams pHMM Precision and Recall Optimization Overview

The full set of precision and recall metrics of Resfams pHMMs can be found in Supplementary Figure 2-S1. We first evaluated the ability of each pHMM to recruit the sequences used to initially train the pHMMs. Protein sequences were recruited into an AR family or
sub-family if they covered greater than 80% of the pHMM and had an e-value score of less than $1 \times 10^{-50}$. These global coverage and e-value scores were optimized to achieve maximum precision and recall. The distribution of precision and recall across all Resfams using these universal significance thresholds shows low precision for individual protein families (Supplementary Figure 2-S1A). These results indicate that a large number of antibiotic resistance families are structurally similar, highlighting an important challenge in predicting antibiotic resistance functions from sequence.

To improve Resfams prediction accuracy, we optimized profile specific gathering thresholds, which set an inclusion bit score cut-off for a protein sequence alignment on a profile-by-profile basis (Supplementary Figure 2-S1B). Finally, we tested the prediction accuracy of these optimized Resfams families on AR proteins from ARDB [133] not used in training of the original pHMMs (Supplementary Figure 2-S1C). These recruited protein sequences were subsequently incorporated into the corresponding Resfams protein families, resulting in the final database of AR profile hidden Markov models used for all further analysis in this study.

2.3.3 Protein annotation using Resfams pHMMs or BLAST to AR specific databases

Proteins were aligned to the core Resfams database of pHMMs (Resfams.hmm) for microbial genome annotation or the full Resfams database of pHMMs (Resfams-full.hmm) for functional metagenomic selections using the hmmscan function of the HMMER3 [70] software package using the following parameters: --cut_ga, --tblout. Antibiotic, mechanism, and beta-lactamase class classification used in this analysis can be found in supplementary data at:

http://dantaslab.wustl.edu/resfams.
The ARDB [133] and CARD [144] were used for comparison of Resfams PHMMs to pairwise sequence alignment to known AR genes. A protein was called an AR gene if it had an amino acid identity greater than or equal to the class specific identity threshold defined by ARDB's documentation over greater than 85% of the length of the target sequence. If no class specific identity threshold was defined or the top hit was to a protein from the CARD [144] database, a protein was called an AR gene if it had greater or equal to 80% amino acid identity over greater than 85% of the length of the target sequence.

2.3.4 Resistome analysis using functional metagenomic selections

All functional metagenomic selections in this analysis were selected, sequenced and assembled into contigs as previously described [72, 71, 151]. Briefly, metagenomic plasmid libraries prepared in *E. coli* DH10B host were selected for resistant inserts on Luria-Bertani or Mueller-Hinton agar plates containing kanamycin (50 μg/mL; plasmid resistance marker) plus the antibiotic of interest at a concentration toxic to wild-type *E. coli* host. Resistant inserts were then amplified and sequenced using the Illumina Hi-Seq Pair-End (PE) 76 or 101 bp sequencing protocol. Sequencing reads were then assembled into contigs using the PARFuMS (Parallel Annotation and Re-assembly of Functional Metagenomic Selections) pipeline [71]. Open reading frames (ORFs) were predicted in assembled contigs using the gene-finding algorithm MetaGeneMark [231] using default parameters. Functional metagenomic selections used for this study were prepared from (1) MDR cultures soil isolates [195], (2) pediatric human fecal samples [151], and (3) Cedar Creek (CC) and Kellogg Biological Station (KBS) soils [72]. Resistome comparisons were performed using a combination of R, Cytoscape, and the QIIME software package [31, 187].

Functional selections using antibiotics common to all three functional metagenomic selections described were used for comparative resistome analysis (Penicillin, Piperacillin,
Tetracycline, Chloramphenicol, and Gentamicin). Contigs were assembled using PAR-FuMS [71] and open reading frames were predicted using MetaGeneMark [231] as described above. Within each sample, all genes 100% identical over the length of the shorter sequence were collapsed into a single sequence using CD-HIT with the following parameters: -c 1.0 -aS 1.0 -g 1-d 0. All proteins over 350bp and unique within sample were then used for downstream analysis. All proteins were then annotated using Resfams pH-MMs as described above, resulting in a total of 3,099 antibiotic resistance proteins used for comparative resistome analysis (64, MDR soil isolates; 1,082, pediatric gut resistome, 1,953; soil resistome). A count matrix of unique gene sequences per Resfams family annotation for each resistome sample was generated by summing unique annotation counts across all antibiotic selections for a sample and normalizing them by metagenomic library size. The normalized antibiotic resistance gene count table was then used to generate bray-curtis and binary jaccard distance matrices and perform principal coordinate analysis (PCoA) using the beta_diversity.py and principal_coordinates.py scripts (QIIME [31]). The significance of clusters by was determined using ANOSIM and performed using the compare_categories.py script (QIIME [31]). Random forests analysis was performed using the supervised_learning.py script (QIIME [31] and randomForest R package [131]) to determine the Resfams families that most discriminate resistomes between habitats. PCoA plots were plotted along with the six most discriminating Resfams families as determined by random forests analysis as biplots for bray-curtis distance matrix. Biplot positions were calculated as the weighted average of the coordinate positions of all samples along the first two PCoA axes, where the weights are the relative abundances of the Resfams family. The size of the biplot points represents the aggregate abundance of the Resfams family across all samples. A bipartite network (Figure 2-3) was generated from the normalize antibiotic resistance gene count table using the make_bipartite_network.py script (QIIME [31]) and then visualized using Cytoscape [187] version 3.0.2 using the edge weighted spring embedded format.
2.3.5 Summary of microbial genomes used in analysis

A total of 6,179 microbial proteomes were downloaded from the IMG database on August 18, 2013. ORFs were called using the IMG pipelines as previously described [140] and protein sequences downloaded from IMG were used in all downstream annotation and analysis. Habitat and potential pathogen status was curated using the metadata available from the IMG database, using the 'Habitat' and 'Diseases' fields.

2.3.6 Resfams pHMM Annotation from Microbial Sequence Alone

To test the ability of Resfams to accurately distinguish between AR genes and all other genomic functions, we used the well-curated UniProtKB/SwissProt database (Supplementary Figure 2-S7). The UniProtKB/Swiss-Prot protein database was downloaded on September 15, 2013, containing 540,732 reviewed proteins. The full set of reviewed proteins was aligned to the core Resfams database of pHMMs (Resfams.hmm) using the hmmscan function of the HMMER3 [70] software package using the following parameters: --cut_ga, --tblout. All proteins in the UniProtKB/Swiss-Prot database that were recruited to at least one Resfams AR gene family are represented on Supplementary Figure 2-S7. Hits were designated as ‘true positive’ if the annotation in the UniProtKB/Swiss-Prot database matched the Resfams AR family annotation for the top hit. In addition, for all efflux/transporter and quinolone resistance AR mechanisms, ‘true positive’ hits required the protein to be designated as an ‘Antibiotic Resistance’ protein in the UniProtKB/Swiss-Prot database as these proteins are often also associated with other functions beyond resistance. All other hits were designated as ‘false positive’ hits.

Resfams AR protein families had less than 5% false discovery rate with the exception of ABC transporters, a class of transmembrane proteins with extremely diverse functions that are categorized together due to a common ATP binding domain. Because ABC trans-
Porters are difficult to predict \textit{a priori} as resistance genes, these Resfams annotations were excluded from analysis in the absence of functional confirmation. Excluding ABC transporters, Resfams very accurately predicted AR function from microbial genomes. Combined with its ability to annotate sequence-divergent AR genes, this indicates that Resfams is adept at predicting AR without the need for a functional verification assay (\textit{e.g.} functional metagenomics).

The accuracy of Resfams pHMMs for predicting AR function from microbial sequence alone is supported by the results obtained from functional metagenomic selections of the human gut and soil microbiotas. For example, we found that antibiotic resistance to tetracycline is mediated almost exclusively by tetracycline MFS efflux pumps in the soil microbiota and by ribosomal protection (TetM/TetO/TetW/TetS) in the human gut microbiota (Figure 2-3). Resfams pHMMs predict the same profile of tetracycline resistance mechanisms across habitats obtained from functional selections in bacterial genomes. Conversely, pairwise sequence alignment to AR specific databases incorrectly predicts enrichment of all tetracycline resistance mechanisms in the human gut versus soil (Figure 2-5b).

### 2.3.7 Comparison of Resfams pHMMs to BLAST to AR gene databases

For comparison of Resfams pHMMs to BLAST to AR gene databases in functional metagenomic selections (Figure 2-1), we used assembled contigs from functional metagenomics studies of MDR soil isolates [195] and the pediatric gut resistome [151]. A total of 161 assembled contigs from the MDR soil isolate resistome study and 3,692 assembled contigs from the pediatric gut resistome study were used for method comparison. ORFs were predicted in the assembled contigs using the stand-alone version of MetaGeneMark [231] with default parameters. Within each functional selection investigated, all proteins 100% identical over the length of the shorter sequence were collapsed into a single se-
sequence using CD-HIT with the following parameters: -c 1.0 -aS 1.0 -g 1-d 0. The longest protein in the identified cluster was retained for downstream analysis. A total of 281 and 9,795 proteins were identified in the MDR soil isolates and pediatric gut resistome study, respectively. The same predicted protein sequences for each study were then annotated using either the full Resfams database of pHMMs (Resfams-full.hmm) or BLAST to antibiotic resistance specific databases as described above. The hand curated annotations for the MDR soil isolates study previously reported in Forsberg et al. 2012 were used as a gold standard for that study for comparison of BLAST and Resfams pHMMs.

A recent report using pairwise sequence alignment to the ARDB [133] concluded that antibiotic resistance is highly enriched in the human gut as compared to natural environments, such as the soil [96]. In contrast, we find no statistical difference between the total AR genes using either functional selection data from metagenomes or sequenced isolate genomes. This emphasizes that studies of AR in microbial genomes and communities and comparisons across habitats requires functional or consensus-based annotation methods in order to provide a complete, unbiased representation of antibiotic resistance reservoirs.

2.3.8 Generation of binary heatmaps for genome comparisons

Binary heatmaps corresponding to genome annotations (e.g., Figure 2-4) were created using the Heatplus package in R. Sections of the heatmap were colored and outlined if there was a significant enrichment of that antibiotic resistance mechanism as determined by Fisher's Exact Test (p<0.01). If there was not a significant enrichment, the heatmap was colored black.
2.4 Acknowledgments

This work was done in collaboration with Kevin J. Forsberg and Gautam Dantas. This chapter is a manuscript published in 2015 in *The ISME Journal*. M.K.G. built and benchmarked Resfams pHMM database, annotated functional selections and microbial genomes, performed ecological analysis and wrote the manuscript, with assistance from K.J.F. and G.D.

Research reported in this publication was supported in part by the NIH Director’s New Innovator Award (http://commonfund.nih.gov/newinnovator/), the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK: http://www.niddk.nih.gov/), and the National Institute of General Medical Sciences (NIGMS: http://www.nigms.nih.gov/), of the National Institutes of Health under award numbers DP2DK098089 and R01GM099538 to G.D. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. M.K.G. is a NSF graduate research fellow (award number DGE-11143954) and a Mr. and Mrs. Spencer T. Olin Fellow. K.J.F. received support from the NIGMS Cell and Molecular Biology Training Grant (GM 007067) and the NHGRI Genome Analysis Training Program (T32 HG000045). M.K.G. is supported by a Mr and Mrs Spencer T. Olin Fellowship at Washington University. M.K.G. and K.J.F. are NSF graduate research fellows (award number DGE-11143954).
2.5 Supplementary Figures

Figure 2-S1: Resfams family precision and recall distributions. The x-axis represents the precision or recall values while the y-axis represents the number of Resfams families of a given value. (a) Precision and recall distributions of Resfams families on the proteins used to train the original profile using universal thresholds. (b) Precision and recall distributions of Resfams families on the proteins used to train the original pHMMs using defined gathering thresholds. (c) Precision and recall distributions of Resfams families on an independent validation set of antibiotic resistance proteins not used to train the Resfams pHMMs using defined gathering thresholds.
Figure 2-S2: Phenotypic resistance profiles across ecologies. Binary heatmap (gray, resistance observed; white, no resistance) of phenotypic profiles of 16 soil microbiota (green) and 18 human gut microbiota (magenta) to 18 antibiotics. Samples are clustered using the Jaccard Index and resistance profiles are herarcically clustered.
Figure 2-S3: Principal coordinates (PCoA) plots of functional metagenomic selections. PCoA plot depicting binary Jaccard distances between resistomes of soil microbiota (green), human gut microbiota (magenta), and MDR soil isolates (blue), calculated using unique ARG counts generated by Resfams annotations. Resistomes of different ecologies cluster separately (p<0.001, ANOSIM).
Figure 2-S4: Distribution of total AR genome composition across habitat and bacteria phyla. The total AR genome composition was calculated by taking the total number of genes annotated by Resfams pHMMs as AR genes in the genome divided by the total number of genes in the genome. The total AR genome composition was averaged across (a) habitats and (c) bacterial phyla. Error bars represent one standard deviation. The raw number of total AR genes per genome was averaged across all genomes by (b) habitat. Error bars represent one standard deviation.
Figure 2-S5: Enrichment of AR mechanisms in functional metagenomic selections. Distribution of Resfams family AR genes identified in functional metagenomic selections of the soil (green) and human gut (red), including (a) all mechanisms, (b) beta-lactamase Ambler classes, and (c) tetracycline resistance mechanisms. The normalized count of unique antibiotic resistance genes per metagenome investigated in each mechanistic category along the x-axis is depicted along the y-axis (*p<0.05, **p<0.001; Student’s t-test).
Figure 2-S6: Enrichment of antibiotic resistance functions in pathogens. (A) Binary heatmap of resistomes organized by pathogen status of 2,966 genome sequenced bacterial isolates. The heatmap is colored by enrichment of a particular AR mechanism within pathogenic or non-pathogenic organisms (p < 0.01, Fisher’s exact). Enrichment of (B) beta-lactamase ambler class and (C) tetracycline resistance functions within pathogenic or non-pathogenic organisms (*p<0.01, Fisher’s exact).
Figure 2-S7: Resfams pHMM annotation of 540,732 proteins in the UniProtKB/Swiss-Prot curated database. The y-axis represents the total number of proteins recruited to Resfams families in each of the antibiotic resistance mechanism categories represented along the x-axis. True positive hits (blue) both match the annotation category and are flagged as antibiotic resistant protein forms in the UniProtKB/Swiss-Prot Database. (a) Recruitment distribution of all Resfams families. (b) ABC Transporter mechanistic class broken down into the individual Resfams families, showing that the majority of false positive arise from the general ABC Transporter pHMM. (c) Recruitment distribution of all Resfams families used in annotation of genomes in absence of functional assay.
Chapter 3

Fast and accurate metagenomic characterization of antibiotic resistance reservoirs

3.1 Abstract & Introduction

Profiling microbial community function from metagenomic sequencing data remains a computationally challenging problem. Mapping millions of DNA reads from such samples to reference protein databases requires long run-times, and short read lengths can result in spurious hits to unrelated proteins (loss of specificity). We developed ShortBRED (Short, Better Read Extract Dataset) to address these challenges, facilitating fast, accurate functional profiling of metagenomic samples. ShortBRED consists of two components: (i) a method that reduces reference proteins of interest to short, highly representative amino acid sequences ("markers") and (ii) a search step that maps reads to these markers to quantify the relative abundance of their associated proteins. After evaluating ShortBRED on synthetic data, we applied it to profile AR protein families in the gut micro-
biomes of individuals from the United States, China, Malawi, and Venezuela. Our results support AR as a core function in the human gut microbiome, with tetracycline-resistant ribosomal protection proteins and Class A beta-lactamases being the most widely distributed resistance mechanisms worldwide. ShortBRED markers are applicable to other homology-based search tasks, which we demonstrate here by identifying phylogenetic signatures of AR across approximately 3,000 bacterial genomes. ShortBRED can be applied to profile a wide variety of protein families of interest; the software, source code, and documentation are available for download at http://huttenhower.sph.harvard.edu/shortbred.

Quantifying proteins of interest from metagenomic sequencing data in a fast and accurate manner is a central challenge in microbial community analysis. Whole metagenome shotgun (WMS) sequencing provides millions of short nucleotide sequencing reads (often 100-250 bases long) from the mixture of bacterial genomes in a microbial community. A common approach to profiling protein families from these data involves (i) mapping reads to a database of reference protein sequences followed by (ii) interpreting the mapping results to estimate protein family relative abundance. This process is complicated by regions of local similarity in otherwise unrelated protein families: reads drawn from such regions will map non-specifically, which can result in false positive identifications (reduced specificity). Reducing the time spent on mapping reads is also an important task, as typical metagenomic sequencing depths and reference database sizes continue to grow rapidly.

Protein families are typically profiled in metagenomic sequencing data by one of three approaches: (i) mapping DNA reads to a database of nucleotide sequences [191], (ii) mapping translated DNA reads to a database of protein sequences [3], or (iii) assembling full-length genes from DNA reads de novo and then annotating them via comparison with reference databases [170]. Approaches (i) and (ii) rely on homology-based searches, as enabled by programs such as BLAST [8], USEARCH [66], and RAPSearch2 [230]. Methods such as MEGAN [101] and HUMANn [3] can achieve very high sensitivity by mapping reads to large nucleotide and protein reference databases. However, this approach is
vulnerable to false positives, as a read derived from a given protein-coding sequence may spuriously align to other genes as a result of local sequence homology, as mentioned above [33]. Moreover, searching nucleotide sequences against large, full-length protein reference databases comes at great computational cost, as search time is roughly proportional to database size and translated search is more computationally demanding than searching in nucleotide space. Assembly-based methods, while advantageous for identifying new genes, tend to underrepresent known, low-abundance genes, as more reads are required to assemble a gene than to identify it by homology-based search. Like search-based methods, assembly is also challenged by regions of local homology, which may lead to the construction of chimeric contigs.

False positive hits to regions of local homology can be mitigated by identifying and mapping against only unique substrings of protein sequences. This has the added benefit of reducing search time, as the unique portions of each database sequence constitute a smaller search space. One approach to this method involves identifying unique k-mers within bacterial protein sequences relative to a larger reference database [25]. While this approach makes progress toward increasing specificity and decreasing runtime, exact k-mer matching is not always biologically satisfying as it can fail to model common patterns in protein sequence evolution. For example, two peptide k-mers that differ by a single substitution between glutamic and aspartic acid (biochemically similar amino acids) are biologically similar, but would be scored as completely distinct by exact k-mer matching. Moreover, while k-mer approaches focus on matches to unique substrings of specific protein sequences, many metagenomics applications, particularly those involving poorly characterized microbial communities, benefit from alignment to more sequence-diverse protein families.

Here we present ShortBRED (Short Better Read Extract Dataset): a method for profiling protein family abundance in metagenomic data by first identifying short peptide markers that (i) are conserved within protein families and (ii) uniquely distinguish fami-
lies from one another. ShortBRED achieves equivalent sensitivity, enhanced specificity, and enhanced speed relative to profiling strategies that map reads against full-length protein sequences. Unlike k-mer based profiling, ShortBRED relies on standard sequence homology-based methods to map reads to peptide markers, thus making it robust to common patterns in protein sequence evolution. By enabling faster, more accurate profiling of protein families in large metagenomes, ShortBRED allows researchers to better measure the prevalence and abundance of proteins families of interest, and can lead to better understanding of biological phenomena. As proof-of-principle, we applied ShortBRED to profile AR families in both human microbiomes and bacterial isolate genomes, revealing new, population-specific, and phylogenetic trends in the distribution of this important class of proteins.

3.2 Results & Discussion

We developed ShortBRED as a method to quickly and accurately quantify the relative abundance of protein families in WMS sequencing data. ShortBRED profiles protein family abundance in metagenomes by a two-step process: (i) ShortBRED-Identify isolates representative peptide sequences (markers) for the protein families, and (ii) ShortBRED-Quantify maps metagenomic reads against these markers to determine the relative abundance of their corresponding families (Figure 3-1). To evaluate ShortBRED, we measured its speed and accuracy in profiling synthetic metagenomes, and then tested its specificity by searching for yeast proteins in a synthetic bacterial metagenome. We next applied ShortBRED to profile AR genes in the gut microbiomes of healthy American [98], Chinese [171], and Venezuelan and Malawian [224] populations, as well as approximately 3,000 microbial isolate genomes.
Figure 3-1: The ShortBRED algorithm. ShortBRED-Identify creates distinctive markers for protein families of interest. ShortBRED-Quantify maps nucleotides reads to markers and normalizes abundance.
3.2.1 Profiling protein families in metagenomes using representative marker sequences

Current approaches to functional profiling of metagenomic samples often involve mapping reads to full-length protein sequences (e.g. centroid sequences of previously defined protein families). ShortBRED obtains higher speed and specificity relative to these approaches by reducing protein families to short, highly representative peptide sequences (markers), and then mapping reads against only those markers. To create the markers, ShortBRED-Identify uses two inputs: (i) a FASTA file of proteins-of-interest and (ii) a comprehensive reference database of additional protein sequences (provided as a FASTA file or preformatted BLAST database; Figure 3-1). ShortBRED-Identify first clusters the protein sequences of interest to identify protein families by global sequence homology, with each collapsed to form a single consensus sequence. Regions of a family’s consensus sequence that share strong, local sequence homology (‘overlaps’) with proteins outside of the family are then penalized. Based on these overlaps, ShortBRED then isolates short peptide markers from the consensus that best represent the protein family. We classify these markers into three groups: True Markers, which do not overlap with the other protein families, Junction Markers, which overlap partially with the other protein families, and Quasi Markers, which are completely overlapped by another protein family.

The marker creation process is run once for a given set of proteins, resulting in a reusable and distributable marker database. ShortBRED-Quantify then (i) maps WMS sequencing reads against a given protein marker database using a translated search, (ii) counts high-quality hits, (iii) normalizes the counts based on marker length and sequencing depth, and (iv) produces a relative abundance profile of the protein families of interest represented by the marker database (Figure 3-1). Creating a highly specific marker sequence database has two major advantages: (i) searches against this database are more accurate, as the exclusion of non-specific (overlap) regions reduces false positive hits,
and (ii) searches proceed more quickly, as the search space is considerably reduced relative to the full database.

### 3.2.2 ShortBRED achieves better specificity than mapping reads to centroids

We constructed synthetic datasets to train ShortBRED’s default parameter settings and validate its performance. For one set of AR protein families (ARDB [133]) and one set of virulence factor protein families (VFDB [36]), we created three synthetic bacterial metagenomes spiked with the proteins of interest at 5%, 10%, and 25% relative abundance. We first tuned ShortBRED’s ability to correctly call the presence and absence of protein families in the 5%-spiked metagenomes by varying the initial protein clustering identity (80%, 85%, 90%, 95%, and 100%) and minimum allowed marker length (8, 10, 12, 15, 18, 20, 22, 25, and 30 amino acids; Figure 3-S1). 85% initial clustering identity and a minimum allowed marker length of 8 amino acids provided an optimal balance of specificity and sensitivity for the ARDB- and VFDB-spiked datasets. These parameter settings are the current defaults in ShortBRED and were used for the remaining analyses in this work; they can be easily tuned with command-line arguments for other applications.

To further validate our parameter settings and ShortBRED’s performance, we generated markers for the ARDB and VFDB protein families based on the optimal settings described above (Table 3-1). We then used these markers to profile all six synthetic metagenomes, including the 10%- and 25%-spiked metagenomes that were not used in the training process. We compared ShortBRED to an alternative profiling strategy in which reads were mapped directly to the centroid sequences of protein families. Centroids were obtained by clustering the proteins of interest at 85% identity; during the quantification stage, any matches to centroids with length \( \geq 30 \) amino acids and \( \geq 95\% \) identity were considered valid hits.
Table 3-1: Characteristics of ShortBRED Markers Used to Profile Synthetic Metagenomes

<table>
<thead>
<tr>
<th></th>
<th>ARDB</th>
<th>VF</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Families after Initial Clustering</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Families with True Markers</td>
<td>618</td>
<td>2,089</td>
</tr>
<tr>
<td>Families without True Markers</td>
<td>594</td>
<td>2,083</td>
</tr>
<tr>
<td><strong>Total Markers</strong></td>
<td>2,886</td>
<td>3,955</td>
</tr>
<tr>
<td>True Markers</td>
<td>2,845</td>
<td>3,940</td>
</tr>
<tr>
<td>Junction Markers</td>
<td>37</td>
<td>15</td>
</tr>
<tr>
<td>Quasi Markers</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

An ideal search methodology will correctly identify all protein families present in a metagenome (true positive rate, TPR, equal to 1) and will not erroneously identify any protein families absent from the metagenome (false positive rate, FPR, equal to 0). As we intensify the criteria for calling a family as present (e.g. requiring a higher normalized count for the family), TPR and FPR will both decrease: a tradeoff we quantify using a receiver operating characteristic (ROC) curve (Figure 3-2A and B; Table 3-S1). Notably, even treating a single hit to a ShortBRED marker as evidence of the corresponding protein family’s presence resulted in exceptional sensitivity with very low false positives (<5%). As we increased the number of protein families present and the share they comprised of the metagenome, ShortBRED achieved TPR and FPR values comparable to or exceeding those of the centroids method (Table 3-S1). The centroids method performed well for the 5%-spiked metagenomes, but experienced a substantial drop in specificity at similar levels of sensitivity to
Figure 3-2: Accuracy of ShortBRED and centroid-based profiling within synthetic metagenomes. (a, b) ROC curves report the sensitivity and specificity of the two methods for correctly identifying the presence and absence of protein families of interest in six synthetic metagenomes, spiked with 5%, 10%, and 25% of their material from the ARDB (panel a) and VFDB (panel b). (c, d) Scatterplots of protein family ‘predicted from mapping’, the abundance values calculated by ShortBRED and the centroids, versus ‘expected from gold standard’, the abundance values of the protein families in the 10% synthetic metagenome.
3.2.3 ShortBRED in analyses of the 10%- and 25%-spiked metagenomes

Beyond correctly calling protein family presence and absence, an ideal search strategy will be able to accurately quantify the relative abundances of these families in a metagenome, which may vary over several orders of magnitude. Using Spearman's correlation to compare known and predicted relative abundances, ShortBRED out-performed the centroid-based method in five out of six cases (median $r=0.948$ versus 0.882; Table 3-S1). The weaker performance of the centroid-based method was due in part to a larger fraction of false positive detection events (defined to have 0 expected abundance; Fig 2C and D). Thus, ShortBRED's increased specificity not only provides a more accurate qualitative profile of protein family presence and absence, but also contributes to more accurate quantitative profiling.

As an additional evaluation of specificity, we applied ShortBRED and the centroid-based profiling method to search for yeast proteins in a synthetic bacterial metagenome. Given that yeast and bacteria are extremely distantly diverged [63], homology between a short, bacterial DNA sequence and a yeast protein is likely to have resulted from chance. ShortBRED did not identify any false positive hits to yeast proteins among the bacterial DNA reads, while the centroid-based method produced fifteen high-identity and long-length hits. The centroid-based profiling method offered some advantages over ShortBRED when working with shallow sequencing data (e.g. as derived from older 454 sequencing experiments), wherein reads are less likely to have been sampled from marker regions (Figure 3-S2). However, this limitation vanishes when working with typical modern sequencing depths, while the drawbacks of the centroid-based approach will only grow as typical depths continue to increase.
3.2.4 ShortBRED is faster than centroid-based profiling

In addition to increasing the accuracy of metagenomic search, ShortBRED’s focus on a reduced sequence database (the markers) results in considerably shorter run-times relative to searching against full-length centroids (Figure 3-3). Focusing on metagenomes spiked with proteins from ARDB, we were able to process 8,500 reads/sec (on average) by mapping against full-length ARDB centroid sequences, while ShortBRED processed 16,000 reads/sec using the previously generated ARDB marker sequences (a 1.9x speedup). For the VFDB-spiked metagenomes, we were able to map 2,500 reads/sec against centroid sequences, while ShortBRED processed 5,700 reads/sec (a 2.2x speedup). All mapping experiments were carried out on the same computer hardware using 10 CPU cores and the same underlying mapping program (USEARCH); hence, ShortBRED’s increased speed can be attributed to the reduced size of the marker database. For a modern metagenomics study producing 100s of millions of reads for 100s of samples, this speedup corresponds to savings of 100s of CPU-hours of compute time.
3.2.5 Antibiotic resistance in the human gut microbiome worldwide

We leveraged the improved specificity of ShortBRED to accurately quantify AR worldwide in the human gut microbiomes of 552 individuals from the United States [98, 224], China [171], Malawi, and Venezuela [224] (Figure 3-4). We identified centroid sequences and ShortBRED markers for 849 AR protein families derived from the ARDB and independent curation [78]. These families were further grouped into broader classes such as such as ‘Class A beta-lactamase’ and ‘quinolone resistance’. 107 microbiome samples based on older 454 sequencing methods were mapped to centroid sequences to avoid loss of sensitivity from low sequencing depths; all other samples were profiled with ShortBRED (see Methods).

Our results support AR as a core function in the human gut microbiome, with every individual gut microbiome containing at least one AR determinant (Figure 3-4). As previously observed [73, 96], resistance to the tetracyclines was the most widespread AR function in the human gut microbiome, with at least one tetracycline resistance mechanism being identified in 99% of individuals globally (97% ribosomal protection; 87% efflux; 3% inactivation). In addition, Class A beta-lactamases were identified in 90% of individuals and were widespread throughout all four global populations, with CfxA and CblA the most common variants (as represented by families P30898 at 68.5% prevalence and P30899 at 60.1% prevalence). Consistent with previous findings [72], this global distribution of AR determinants in the human gut microbiome appears to be driven by the underlying bacterial phylogenetic profile. For example, while Class A beta-lactamases are known to be the most diverse and widely disseminated class of beta-lactamase genes [91], the most abundant variants (CfxA and CblA) have been previously shown to be specific to species Bacteroides [192, 160]. Hence, enrichment for these families may be a direct marker for the presence of specific bacterial phyla in the gut microbiota rather than a response to selective pressures of individual-specific antibiotic use. The relationship be-
**Figure 3-4: Antibiotic resistance in the human gut microbiome.** RPKM values produced by ShortBRED for antibiotic resistance protein families, summed by class of resistance. Samples in the USA-Global, Venezuela, and Malawi cohorts were profiled by mapping reads to centroids due to their lower sequencing depth. Marker information is listed in Table 3-2. Samples (columns) were clustered according to Canberra distance and antibiotic resistance families (rows) were clustered according to Euclidean distance.

Between microbiome-specific AR and phylogenetic profiles are addressed in greater detail in subsequent sections.

In addition to the universal AR trends described above, ShortBRED revealed several consistent differences in AR profiles between global populations. For example, Chinese individuals were particularly enriched for a number of AR factors: quinolone resistance, aminoglycoside acetyltransferases, and genes modulating antibiotic efflux. Among these individuals, the two most prevalent quinolone resistance families (variants of fluoroquinolone-resistant DNA topoisomerases) were found in 78% and 28% of individuals, the most prevalent aminoglycoside acetyltransferase (YP_002559372) was found in 99.7% of individuals, and the next-most-prevalent (P13246) followed at 19.2% of individuals. Four gene-modulating antibiotic efflux families (phoQ_1, soxR_5, marA_1, and baeR_2) had...
individual prevalence values exceeding 58%. In addition, while many AR genes were discretely strongly present or absent within the Chinese cohort (Figure 3-4), their gut resistomes were differentiated into four clear clusters based largely on the abundance of antibiotic efflux pumps, including major facilitator superfamily (MFS) antibiotic efflux, resistance/nodulation/cell division (RND) antibiotic efflux, and small multidrug resistance (SMR) antibiotic efflux pumps. Many multidrug antibiotic efflux pumps are chromosomally encoded and highly conserved across all members of a given bacterial species [9], further suggesting that observed AR distribution patterns are driven by underlying community membership and phylogeny. In comparison with the Chinese cohort, gut microbiome samples from the American (HMP) cohort were much more homogeneous. This difference is likely influenced by the greater diversity in membership among the Chinese cohort, which contained individuals with and without type II diabetes and a wide range of ages (13-86). In comparison, the HMP cohort consisted solely of young (ages 18-40), healthy individuals. Differences between the cohorts may also reflect variation in the sampling and sequencing protocols used by their corresponding studies (in addition to real biological variation). American individuals were characterized by increased abundance of four protein families within the Class A beta-lactamases (CfxA_11, AAA22905, P30898, and P30899). Conversely, these individuals were depleted for aminoglycosides and acetyltranferases. These observed differences between the American and Chinese cohorts stress that, while AR is (at a high level) core to the global human gut microbiome, variation emerges in specific resistances present in individual populations.

3.2.6 Connecting antibiotic resistance to phylogeny

In order to understand and control the spread of AR, it is necessary to characterize the connections between AR determinants, source genomes and their phylogeny, and the relative propensity of HGT. In addition to their usefulness in metagenomics profiling,
ShortBRED markers can aid in this goal by providing highly specific signatures of AR protein families for microbial genome annotation. Along these lines, we used ShortBRED to profile the 849 AR protein families introduced above across 3,305 phylogenetically diverse microbial isolate genomes [140] (see Methods).

Over 40% of microbial isolate genomes surveyed encoded at least one AR determinant, with significant enrichments among particular genera (Figure 3-5). For example, *Escherichia* and *Staphylococcus* are closely related bacterial genera that contain many human pathogens [113, 41]; both were highly enriched for AR determinants. Specifically, all *Escherichia* and *Salmonella* encoded at least one AR class, with an average of 20.3 AR genes for *Escherichia* and 11.2 AR genes for *Salmonella*. In addition, while these two genera shared many similar AR determinants, they appear to resist beta-lactam antibiotics using largely orthogonal mechanisms: 94.6% of *Escherichia* genomes were enriched for Class C beta-lactamases and were completely depleted of Class B beta-lactamases, while *Salmonella* showed the opposite trend (6.5% of genomes encoded Class C resistance, while all genomes encoded Class B resistance, Figure 3-S2). While these examples illustrate cases of strong coupling between AR determinants and particular genera, this was not always the case. For example, glycopeptide resistance was highly variable within the genus *Enterococcus*, with approximately $\frac{1}{3}$ of isolate genomes possessing the function while the remaining $\frac{2}{3}$ lacked it.

Our observations further suggested that AR functions could be subdivided into two categories of phylogenetic distribution: (i) functions that are clade-specific, *i.e.* highly conserved across all members of a bacterial clade, and (ii) functions that are broadly distributed across the phylogenetic tree. Both distribution patterns were observed among abundant AR classes in the human gut microbiome (Figure 3-5 and 3-S3). For example, multi-drug antibiotic efflux pumps and rRNA methyltransferases showed strong signatures of clade-specific enrichment among the *Staphylococcus*., *Escherichia*, *Salmonella*, and *Yersinia* genera. Functions that are tightly linked to particular clades are notable
Figure 3-5: Prevalence of antibiotic resistance across bacterial isolate genomes. Phylogenetic tree of bacterial genomes from IMG [140] overlaid with presence/absence of ShortBRED antibiotic resistance protein families. The outermost ring indicates the share of genes in each species’ genome that mapped to any of the AR protein families.
in that their presence and abundance can be inferred from profiles of community com-
position alone, including profiles based on lower-resolution amplicon sequencing [123].
Conversely, tetracycline ribosomal protection determinants were widely dispersed across
the phylogenetic tree, a pattern more consistent with recent spread by mechanisms such
as HGT [191]. The presence and abundance of functions in this category would be diffi-
cult to infer from community profiling and are best quantified directly from a metagenome,
a process facilitated by ShortBRED.

3.2.7  Predicting antibiotic resistance profiles from community com-
position

The previous section stressed that, while some AR functions can be accurately quan-
tified based on microbial community composition, broadly distributed functions pose a
greater challenge. To further explore this idea, we compared observed and predicted AR
profiles for 82 gut metagenomes from HMP individuals. We predicted the AR profile for a
given gut metagenome by first quantifying the sample’s microbial community composition
with MetaPhlAn [186]. This step resulted in a vector of relative abundance measurements
for species present in the sample (in RPKM units). Then, using the ShortBRED based-
annotations of AR functions in bacterial genomes described above, we computed the
abundance of each AR function in the sample by multiplication. For example, if species A
had a relative abundance of 5 RPKM and contained 1 copy of AR protein X, while species
B had a relative abundance of 10 RPKM and contained 2 copies of AR protein X, then the
total abundance of AR protein X in the metagenome was predicted to be: 1(5 RPKM) +
2(10 RPKM) = 25 RPKM. This procedure was repeated for all samples and AR functions.

At the level of individual AR gene families, ShortBRED and the predicted profiles co-
detected 63 families, 57 were detected by ShortBRED but never observed in the predicted
profiles, and 14 were predicted to be present but never confirmed by ShortBRED. Among
the co-detected families, the average quantitative agreement between the ShortBRED and predicted profiles (as measured by Spearman’s correlation) was 0.43. When gene families were grouped into broader AR classes, 17 were co-detected, 5 were found only by ShortBRED, and 1 was predicted to occur but not confirmed by ShortBRED. Average quantitative agreement for the 17 co-detected classes was 0.33 (Spearman’s correlation). The AR classes most under-represented by the community composition-based predictions were tetracycline ribosomal protection, Class A beta-lactamase, rRNA methyltransferase, MFS antibiotic efflux, and RND antibiotic efflux. Notably, tetracycline resistance was also among the most broadly-distributed AR classes.

In addition, individual-specific ShortBRED-based versus predicted AR profiles showed poor quantitative agreement (average Spearman correlation = 0.53). There are a number of reasons why the two profiles would agree poorly on an individual basis or for particular AR families. While ShortBRED is able to profile AR gene abundance in cases where the genes are present in uncharacterized genomes, the taxonomic profile method is limited to species with known isolate genomes. Hence, predicting AR content from taxonomic composition will tend to underestimate AR content, and explains why ShortBRED detects several families that the predictive method does not. In instances where multiple isolate genomes were available for a species detected in a sample, the species’ contributions were based on the median gene copy number for each AR family across its isolate genomes. If the sample isolate contained fewer copies of an AR gene than the median estimate, then we would tend to overestimate its abundance; conversely, if the sample isolate contained more copies of an AR gene than the median estimate, then we would tend to underestimate its abundance (both serving to weaken signal-to-noise ratio among the predictions). For these reasons, directly profiling AR content in a metagenome with ShortBRED is preferable to predicting functional content from community composition.
In this work, we have presented and validated ShortBRED: a tool to build short peptide markers for protein families and then apply them to profile protein family content in a metagenomic sequencing sample. We demonstrated that ShortBRED is both faster and more accurate than the common approach of mapping reads to full-length protein family centroid sequences. ShortBRED is extensible to a diverse collection of functional profiling tasks. The most straightforward of these was demonstrated in our profiling of AR in human gut metagenomes, which we discuss further below. Although this example was based on DNA sequence data, ShortBRED’s markers are also applicable for profiling microbial community RNA-Seq data (metatranscriptomics), which reveals the relative functional activity of protein families in a community. In addition to profiling metagenomic sequencing data, ShortBRED’s markers have proven useful for identifying protein families of interest in microbial isolate genomes, as the markers’ small sizes and highly representative sequences facilitate rapid, unambiguous gene annotation. The functional profiles produced in these applications are amenable to a variety of downstream analysis methods, including comparing functional composition in case versus control samples or monitoring temporal variation in functional composition or activity from longitudinal samples. Mapping metagenomic reads to protein families of interest is an example of a search problem in which new queries (samples) arise more frequently than changes to the database (proteins of interest). In such cases, it is desirable to pre-process the database in order to accelerate downstream search. ShortBRED accomplishes this by reducing large numbers of protein sequences first to clusters of related proteins (families) and then to representative peptide markers. Searching a new metagenomic sample against these marker sequences represents a considerable savings in computation relative to searching against the full database. ShortBRED’s pre-processing steps, while not computationally trivial, can reduce a collection of 1,000 protein families to identifying markers in a
matter of hours on typical desktop or server hardware (i.e. taking advantage of multiple CPU cores for parallelization, but not requiring special high-memory or accelerated file I/O configurations). The bottleneck in this process is the BLAST-based search of the proteins of interest against the universal protein reference database. In the future, it may be possible to further accelerate pre-processing steps by incorporating an alternative program for protein homology search, provided that it meets or exceeds BLAST’s sensitivity. In the same vein, downstream performance mapping reads to markers depends largely on the speed of ShortBRED’s chosen translated search tool (currently USEARCH), which could also be replaced or supplemented by future alternatives. In our evaluations, the vast majority of protein families could be identified by one or more unique amino acid subsequences (True Markers). Although these sequences are used here for protein family identification and quantification, they are themselves interesting targets for investigation. For example, the conservation of these sequences within a family may indicate the presence of a functionally relevant domain, peptide recognition motif, or enzyme active site. The small minority of protein families that lacked unique identifying subsequences are also worthy of consideration (Tables 3-1 and 3-2). In such cases, ShortBRED constructs a Quasi Marker to represent the family: i.e. the amino acid sub-sequence which, while not unique to the family, is found in a minimal number of other families. Users may wish to exclude Quasi Markers (and their associated families) in their analyses to increase specificity. Quasi Markers were always included in the analyses reported here and were found to compromise specificity only slightly (far less than the centroid-based approach; Figure 3-2 and Table 3-2). In the future, an expectation maximization (EM) step could be incorporated in ShortBRED-Quantify to improve the accuracy of protein family quantification when mapping reads to ambiguous Quasi-Markers.

We demonstrated ShortBRED’s utility by generating and applying AR gene markers to profile AR gene content in 552 human gut metagenomes and 3,305 bacterial isolate genomes. AR determinants in pathogens are increasingly compromising infec-
Table 3-2: Characteristics of ShortBRED Markers Used to Profile Metagenomes and Bacterial Genomes. ** Centroids were used to profile the Yatsunenko dataset.

<table>
<thead>
<tr>
<th>Parameter Settings</th>
<th>HMP</th>
<th>T2D</th>
<th>T2D_Short</th>
<th>Yatsunenko</th>
<th>Bacterial Genomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustering ID</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Minimum Marker Length</td>
<td>85%</td>
<td>85%</td>
<td>85%</td>
<td>85%</td>
<td>85%</td>
</tr>
<tr>
<td>Average Read BP</td>
<td>101</td>
<td>90</td>
<td>75</td>
<td>450</td>
<td>100</td>
</tr>
<tr>
<td>Min Trusted BP</td>
<td>90</td>
<td>81</td>
<td>68</td>
<td>200</td>
<td>30</td>
</tr>
<tr>
<td>QMLength</td>
<td>30</td>
<td>27</td>
<td>22</td>
<td>66</td>
<td>33</td>
</tr>
<tr>
<td>Statistics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Families after Initial Clustering</td>
<td>849</td>
<td>849</td>
<td>849</td>
<td>849</td>
<td>849</td>
</tr>
<tr>
<td>Families with True Markers</td>
<td>820</td>
<td>820</td>
<td>820</td>
<td>**</td>
<td>820</td>
</tr>
<tr>
<td>Families without True Markers</td>
<td>29</td>
<td>29</td>
<td>29</td>
<td>**</td>
<td>29</td>
</tr>
<tr>
<td>Total Markers</td>
<td>4132</td>
<td>4135</td>
<td>4142</td>
<td>**</td>
<td>4132</td>
</tr>
<tr>
<td>True Markers</td>
<td>4078</td>
<td>4078</td>
<td>4078</td>
<td>**</td>
<td>4078</td>
</tr>
<tr>
<td>Junction Markers</td>
<td>48</td>
<td>50</td>
<td>61</td>
<td>**</td>
<td>48</td>
</tr>
<tr>
<td>Quasi Markers</td>
<td>6</td>
<td>7</td>
<td>3</td>
<td>**</td>
<td>6</td>
</tr>
</tbody>
</table>
tious disease treatment due to their acquisition from commensal or environmental bacteria [71, 189]. The human gut microbiome serves as a transferable reservoir of AR readily available to human pathogens [189], leading to an increased focus on characterization and quantification of AR genes in large metagenomic studies [73, 96, 155]. However, accurate quantification of AR genes using short reads is challenging: AR determinants are often originally genes with diverse native functions repurposed through mutation or expression modulation to provide AR, therefore sharing large sequence similarity to genes with no AR function. For example, when particular genes in the RND efflux pump superfamily (such as CmeABC, AcrB, and Mex) are highly expressed, they are capable of exporting multiple antibiotics [218]. However, members of the RND efflux pump superfamily also serve important functions as transporters of proteins required for nodulation and cell division and share high sequence similarity with proteins shown to serve resistance functions. As a result, previous attempts to profile AR in human gut metagenomes by mapping short reads to full-length protein sequences may have been compromised by spurious mapping events. Our ShortBRED-based profiles avoided this complication by using only the most information-rich portions of AR genes for identification and quantification of AR in microbial communities and isolate genomes. Notably, our results agree with several of the major findings from previous profiling attempts, specifically (i) high relative abundance of AR genes in the gut microbiota of Chinese individuals compared to individuals from other countries as well as (ii) ubiquity of tetracycline resistance worldwide [73, 96]. Hence, we can be confident that these results are not the result of spurious mapping to full-length protein sequences. ShortBRED demonstrated increased sensitivity for identification of additional classes of AR genes, including resistance to the quinolone class of antibiotics. The application of ShortBRED to the identification and quantification of AR genes in microbial communities addresses a significant challenge in the computational investigation of AR using high-throughput sequencing technology. In addition, just as ShortBRED markers enabled confident differentiation
between closely-related AR and non-AR proteins in metagenomes, the same advantage applied to annotating full-length protein sequences in bacterial isolate genomes. Indeed, we used this method to dissect phylogenetic properties of the AR families under study, revealing distinct patterns of clade-specific versus broad distribution. In the future, the same technique could be applied to quickly and accurately determine AR gene content in a newly-sequenced bacterial strain, an application with relevance to infectious disease management. To facilitate such applications, the AR markers produced here are available for download, along with the ShortBRED software and documentation, at the ShortBRED website: http://huttenhower.sph.harvard.edu/shortbred. Although the preceding examples have focused on applying ShortBRED to profile AR in genomes and metagenomes, this is only one possible application. Indeed, the same analyses described above can be applied to a wide range of protein families of interest, limited largely by the imagination of the user. To that end, users who produce marker sets with ShortBRED and who would also like to share them with the scientific community are encouraged to submit the markers (along with a relevant citation) for posting on the ShortBRED website.

3.3 Materials & Methods

3.3.1 Creating protein family-specific marker sequences with ShortBRED-Identify

ShortBRED-Identify takes two inputs: (i) a FASTA file of proteins of interest and (ii) a comprehensive catalog of reference protein sequences (as a FASTA file or preformatted BLAST database). The reference database used here was based on version 3.5 of the Integrated Microbial Genomes database [140]. The full version of this database contained 12,607,998 protein coding sequences, which we previously reduced to 4,981,629
representative protein coding sequences proteins by clustering at 80% nucleotide identity [185]. ShortBRED-Identify first clusters the proteins of interest at 85% identity using CD-HIT [74, 128] to group them into highly conserved protein families. For each clustered protein family, ShortBRED-Identify first calls MUSCLE [65] to generate a multiple sequence alignment (MSA) for the family, then uses Biopython [42] to generate a consensus sequence for the MSA. If the most common amino acid for a given MSA column occurred in less than 95% of sequences, the corresponding position in the consensus sequence is marked as ambiguous (‘X’). ShortBRED-Identify then uses BLAST [8] to query consensus sequences (i) against one another and (ii) against the reference protein database. The results of these searches are used to identify short segments of each consensus sequence that align with high sequence identity (≥90%) to unrelated proteins in the reference database, or share high identity with a length greater than 80% of minimum marker length with other consensus sequences. (A short sequence is defined as having a length between 80% of the minimum marker length and 15% of a target sequence in the reference database.) Metagenomic reads derived from such segments will be prone to false positive matches across protein families. ShortBRED-Identify thus interprets the BLAST results to find segments of a consensus sequence that participate in a minimal number of such alignments (markers) and then uses these sequences as a basis for more accurate functional profiling. Consensus sequences from different families can share long regions of similarity even after initial clustering at high sequence identity. Because of this, ShortBRED-Identify penalizes high-identity alignments of any length greater than 80% of marker length between pairs of consensus sequences in order to minimize inter-family false positives. ShortBRED does not penalize high-coverage, high-identity alignments between a consensus sequence and a protein from the reference database, as such proteins are likely members of the protein family represented by the consensus. ShortBRED counts the number of times each amino acid of each consensus sequence appeared in a valid alignment with another protein. These ‘overlap counts’ are then used
to identify the most representative segments (markers) for the consensus. For a given consensus sequence, ShortBRED will first try to build as many ‘True Markers’ as possible. A True Marker is a contiguous sequence of amino acids with zero overlap count; i.e. the corresponding peptide was unique among the consensus sequences and non-member reference sequences. If no True Markers are found above a minimum length (with a default of 8 amino acids), ShortBRED next tries to make up to three ‘Junction Markers’ for the consensus sequence. A Junction Marker is a sequence of amino acids that partially overlaps with other consensus sequences or reference sequences, but is not completely overlapped by any single consensus or reference sequence. Note that when mapping reads to marker sequences, ShortBRED-Quantify requires high-identity ($\geq 95\%$) and high-length (minimum of the marker’s length or 95% of a read length), and hence these partial overlaps will not lead to false positive matches. If it is not possible to build a True Marker or a Junction Marker for a consensus sequence, ShortBRED-Identify will create a single ‘Quasi Marker’ for the consensus, which is a sequence of amino acids above a given minimum length (with a default of 33 amino acids) that has the lowest total adjusted overlap count. The adjusted overlap count is the fourth root of the raw overlap count, and helps to down-weight very short outlier regions with extremely high overlap counts. Protein families with similar Quasi Markers and Junction Markers ($\geq 95\%$ identity) are merged, and then all marker sequences are output as a FASTA file for use by ShortBRED-Quantify. For Junction Markers and Quasi Markers, ShortBRED also lists the percentage of each marker that overlaps with each other consensus sequence. Each sequence is given a weight, which is defined as its total length in amino acids divided by the sum of that value, and all overlapping amino acids from other reference or consensus sequences. The weight is printed in the FASTA header, along with other highly overlapping consensus sequences from the input database. An additional text file lists the overlapping regions from the consensus sequences and the reference database. This ShortBRED-Identify process requires 100 CPU-hours to complete given a set of proteins of interest.
which cluster to 1000 protein families. The bottleneck in this process is the BLAST-based search of the protein family consensus sequences against the comprehensive reference database. Notably, this process is highly parallelizable, as each consensus sequence can be searched independently of the others. By allowing ShortBRED-Identify to use multiple cores during the search process, the actual run-time can be reduced considerably. Once the initial BLAST results have been generated, new markers can be generated in a few minutes provided that the initial clustering identity and consensus thresholds are not changed. Precomputed markers for the AR proteins (ARDB) [133] and virulence factors (VFDB) [36] are available for download at http://huttenhower.sph.harvard.edu/shortbred. Notably, the ShortBRED-Identify process needs to be applied only once to produce a set of markers, which can then be used repeatedly to profile metagenomic datasets using ShortBRED-Quantify.

### 3.3.2 Profiling protein family metagenomic abundance with ShortBRED-Quantify

After markers have been created for each protein family, the user can call ShortBRED-Quantify to profile the relative abundance of these families in a WMS sequencing sample. ShortBRED-Quantify calls USEARCH [66] to find the best matching marker for each nucleotide read. USEARCH specializes in fast search for high-identity matches, which fits with ShortBRED’s objective of profiling metagenomic samples quickly with high specificity. By default, ShortBRED-Quantify will record a hit to a marker if the resulting alignment has at least 95% identity, and is at least as long as the minimum of (i) the marker length or (ii) 95% of the read length. For each marker, ShortBRED-Quantify computes an adjusted marker length, which takes into account how much of the marker is available to participate in a hit meeting our length and percent identity requirements. When a marker of length $L$ is longer than the average read length ($R$), a read from the corresponding gene
anywhere in the region from 5% downstream of the marker to 5% upstream of the marker can successfully align to the marker. Therefore, the adjusted marker length ($L'$) is:

$$L' = L - 0.9R + 1$$

When the marker is shorter than the expected read length ($L < R$), the entire marker must align to the read. Thus, the adjusted marker length is:

$$L' = R - L - 1$$

ShortBRED then normalizes the number of raw USEARCH hits to a marker ($H$) to produce a normalized count ($C$), adjusting for average read length, marker length, and sequencing depth (number of reads in the sample, $N$):

$$C = \left( \frac{H}{L'} \right) \left( \frac{N}{10^6} \right) = \left( \frac{H}{L'N} \right) 10^9$$

The normalized count is in units of RPKMs (reads per kilobase of reference sequence per million sample reads). For protein families characterized by multiple markers, a normalized count is first computed for each marker separately and then the median of these values is taken to represent the protein family; this procedure adds robustness to variation in sequencing depth across the markers. Finally, ShortBRED-Quantify outputs these normalized counts as a relative abundance table for the protein families of interest.

### 3.3.3 Creation of synthetic spiked metagenomes

We used GemSim [145] to create synthetic metagenomes containing five million 100 nucleotide-long reads, designed to mimic a typical WMS-sequencing run on an Illumina HiSeq instrument (Illumina, San Diego, CA). Reads were drawn from twenty bacterial
genomes obtained from the KEGG database \cite{112, 111}. We used USEARCH \cite{66} to identify and exclude from these genomes any naturally-occurring AR genes and virulence factors (defined as a sequence matching a gene from the ARDB or VFDB with >90\% identity). This ensured that the only ARDB and VFDB sequences in our synthetic metagenomes were those that had been artificially spiked in for the purposes of evaluating ShortBRED. Each bacterial genome was assigned an abundance value drawn from a log-normal distribution with unit mean and standard deviation. We created six Illumina-like synthetic metagenomes based on with ARDB and VFDB sequence datasets. Three metagenomes were made for each dataset, with 150, 500, and 1,000 genes from the corresponding protein dataset spiked among the genomic reads at 5\%, 10\%, and 25\% relative abundance. An additional set of six metagenomes were created using the same procedure but based on a simulated 454 sequencing instrument (454 Life Sciences, Branford, CT); these samples contained only 155,890 reads each, consistent with the lower sequencing depth of the 454 platform. We used 164 nucleotide sequences corresponding to ARDB protein sequences as a base for the ARDB metagenomes and 2,296 VFDB nucleotide sequences as a base for VFDB metagenomes. Nucleotide sequences were not always provided for ARDB proteins; in these cases, we used the EMBOSS program backtranseq \cite{175} to create nucleotide sequences that were compatible with the available amino acid sequences.

### 3.3.4 Application of ShortBRED to human gut metagenomes

We applied ShortBRED to profile AR in the human gut microbiome. We first produced a set of new AR marker sequences by applying ShortBRED-Identify to a combination of (i) a curated version of the ARDB which we obtained by deleting sequences no longer stored at NCBI and (ii) a set of known AR genes obtained from resistant bacterial libraries. We then used ShortBRED-Quantify to profile the relative abundance of corresponding
AR protein families across 552 gut metagenomes: 82 from U.S. adults sampled during the Human Microbiome Project (HMP) [98], 363 from Chinese adults with and without diabetes [171], and 107 individuals from Malawi, Venezuela, and the U.S. [224]. We used the first-visit samples from multi-visit HMP subjects to avoid redundancy. For 454-based samples characterized by sub-optimal sequencing depth, we mapped reads to full-length centroid sequences to avoid compromising sensitivity.

### 3.3.5 Application of ShortBRED to bacterial isolate genomes

ShortBRED can be applied to identify protein families in a bacterial isolate genome given a corresponding set of ShortBRED markers for those families. To do so, ShortBRED first creates a USEARCH database for the genome and then searches the markers against that database (allowing for multiple hits per marker query). For protein families characterized by more than one marker sequence, ShortBRED requires that a critical fraction of the markers map to a gene in the genome before assigning it to that protein family. The default value for this cutoff is 10% (i.e. 1 in 10 markers), but it can be tuned to be more conservative.

### 3.4 Acknowledgements

The work presented in this chapter was done in collaboration with Jim Kaminski, Eric A. Franzosa, Nicola Segata, Gautam Dantas, and Curtis Huttenhower and is currently under review for publication. J.K., M.K.G., E.A.F, N.S., G.D., and C.H. conceived and designed the experiments, J.K. and M.K.G. performed the experiments, J.K. wrote the software, J.K., M.K.G., E.A.F., G.D. and C.H. analyzed the data, J.K., M.K.G., E.A.F., G.D., and C.H. wrote the manuscript.
Levi Waldron, Timothy L. Tickle, Joseph Moon, Xochitl C. Morgan, and Daniela Börnigen provided helpful feedback throughout ShortBRED's development and initial application. Research reported in this publication was supported in part by NIH grants of a Director's New Innovator Award, DP2DK098089, and R01GM099538, and the March of Dimes Foundation 6-FY12-394 to G.D; and by NIH grants R01HG005969, NSF grant DBI-1053486, ARO grant W911NF-11-1-0473, and Danone Research grant PLF-5972-GD (PI Wendy Garrett) to C.H. MKG is supported by a Mr. and Mrs. Spencer T. Olin Fellowship at Washington University and is a NSF Graduate Research Fellow (DGE-11143954). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.
3.5 Supplementary Tables & Figures

**Figure 3-S1**: Accuracy of ShortBRED and centroid-based mapping in applications involving 454 sequencing data. Values reflect area under the ROC curve (AUC) as minimum marker length and initial clustering ID are varied. This analysis was based on the 5%-spiked synthetic metagenomes containing ARDB (a) and VFDB (b) sequences.

**Table 3-S1**: Performance of ShortBRED and Centroid Method across Six Synthetic Metagenomes. Results from using ShortBRED with minimal marker length of 8, initial clustering ID of 85% to profile six synthetic metagenomes. Centroids were also obtained by clustering at 85% identity. Any result mapping to centroid with length ≥30 and ID ≥95% was considered a “match”.

<table>
<thead>
<tr>
<th></th>
<th>ShortBRED</th>
<th>Centroids</th>
<th>ShortBRED</th>
<th>Centroids</th>
<th>ShortBRED</th>
<th>Centroids</th>
<th>ShortBRED</th>
<th>Centroids</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC</td>
<td>AUC</td>
<td>Sens</td>
<td>Sens</td>
<td>Spec</td>
<td>Spec</td>
<td>Corr</td>
<td>Corr</td>
</tr>
<tr>
<td>A05</td>
<td>0.97</td>
<td>0.98</td>
<td>0.95</td>
<td>0.98</td>
<td>0.99</td>
<td>0.85</td>
<td>0.94</td>
<td>0.83</td>
</tr>
<tr>
<td>A10</td>
<td>0.98</td>
<td>0.97</td>
<td>0.98</td>
<td>0.98</td>
<td>0.97</td>
<td>0.82</td>
<td>0.95</td>
<td>0.86</td>
</tr>
<tr>
<td>A25</td>
<td>0.98</td>
<td>0.97</td>
<td>0.98</td>
<td>1.0</td>
<td>0.97</td>
<td>0.75</td>
<td>0.95</td>
<td>0.90</td>
</tr>
<tr>
<td>V05</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
<td>0.96</td>
<td>0.91</td>
<td>0.82</td>
</tr>
<tr>
<td>V10</td>
<td>1.0</td>
<td>1.0</td>
<td>0.99</td>
<td>1.0</td>
<td>0.98</td>
<td>0.96</td>
<td>0.97</td>
<td>0.95</td>
</tr>
<tr>
<td>V25</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.98</td>
<td>0.96</td>
<td>0.99</td>
<td>0.98</td>
</tr>
</tbody>
</table>

87
Figure 3-S2: Training/testing of ShortBRED’s default parameters. (a) and (b) report the sensitivity and specificity of the two methods for mapping reads to their correct families on six synthetic 454 metagenomes, spiked with 5%, 10%, and 25% of their material from the ARDB (panel a) and VFDB (panel b). (c) and (d) display scatterplots of protein family “predicted by mapping”, the abundance values calculated by ShortBRED and the centroids, vs. “expected from gold standard”, the abundance values of the protein families in the 10% synthetic metagenome. This figure is an analog of Figure 3-2 from the main text.
Figure 3-S3: Prevalence of all antibiotic resistance classes across bacterial isolate genomes. Phylogenetic tree of bacterial genomes from IMG [140] overlaid with presence/absence of ShortBRED antibiotic resistance protein families. The outermost ring indicates the share of genes in the species’ genome that mapped to any of the AR protein families.
Dysbiotic colonization of the preterm infant gut microbiota may have life-long effects on host health [154, 40, 45]. The normal progression of colonization of preterm infants has been shown to be patterned, yet chaotic, and thought to be driven primarily by host factors [120]. While preterm infants almost universally receive early and often repeated and/or prolonged intravenous antibiotic therapy [39], it remains unclear if and how specific antibiotics alter the developmental progression of the gut microbiota in this population [120]. By analyzing 401 fecal samples from 84 longitudinally-sampled preterm infants, we demonstrate that meropenem, cefotaxime, and ticarcillin-clavulanate significantly reduce species richness and significantly enrich or deplete specific bacterial species and AR genes. In contrast, vancomycin and gentamicin, two of the antibiotics most commonly ad-
ministered to preterm infants, produce a varied response in species richness. We show this response is predictable with 85% accuracy based on the relative abundance of only two bacterial species and two AR genes before treatment. To more fully investigate the role of the gut-associated resistome in the response to specific antibiotics, we performed functional metagenomic selections for resistance to 16 clinically relevant antibiotics from a set of 21 preterm infant gut microbiota. Of the 794 AR genes we identified, 79% had not previously been classified as providing any resistance function. Combined with deep shotgun sequencing of all 401 fecal samples, we find that MDR *Escherichia*, *Klebsiella*, and *Enterobacter* species commonly associated with nosocomial infections dominate the preterm infant gut microbiota. We show that AR genes enriched following specific antibiotic treatments are generally unique to the specific treatment and highly correlated (p<0.001) with the abundance of a single bacterial species. The most notable exception includes ticarcillin-clavulanate and ampicillin, which both result in a large number of overlapping AR genes being enriched, which are highly correlated with *Klebsiella pneumoniae* species in the preterm infant gut microbiota. In addition to AR genes relevant to specific antibiotics likely providing a protective function, we show that all antibiotic treatments also result in widespread collateral microbiome impact, through enrichment of AR genes with no known direct activity against the specific antibiotic treatment.

The developmental progression of the human infant gut microbiota is thought to have significant and lasting effects on health throughout life [154, 40, 45]. However, much of our knowledge about the dynamics of infant gut colonization has been based on limited longitudinal sampling and studies of term infants using 16S rRNA marker gene studies. Such sequencing strategies are restricted in their resolution for illuminating functional microbiome architectures [224, 120]. Further, it is becoming clear that the principles that govern infant gut microbiota colonization in term infants may differ from those of preterm infants [13]. Indeed, we show in the first two months of life, the preterm infant gut microbiota is colonized by approximately 10-fold fewer bacterial species than term infants.
(P<0.001, Wilcoxon; Figure 4-S1a), is compositionally distinct (P<0.001, PERMANOVA; Figure 4-S1b), and is not influenced by delivery mode as has been shown for term infants [82, 17] (P=0.454, PERMANOVA; Figure 4-S1c). As the number of preterm births remains high worldwide [26], the development of the preterm infant gut microbiota, and early life events that perturb this progression, are emerging as critical research and clinical priorities [81]. Antibiotics are the most prescribed medication in Neonatal Intensive Care Units (NICU) in the US [39]. Of the 460 preterm infants enrolled in our preterm infant neonatal microbiome cohort at St. Louis Children’s Hospital [120] between July 2009 and December 2013, 97% received antibiotics at birth. In many cases, this initial therapy is followed by substantial additional antibiotic exposure - each infant in this cohort received antibiotics for median (interquartile range, IQR) of 8 (3,19) days, corresponding to a median of 17% of the duration of their stay in the NICU. Accordingly, we sought to identify which, if any, of the specific antibiotics used in preterm populations influence gut microbiota development as well as how the commensal resistome mediates this response.

4.2 Results & Discussion

4.2.1 Cohort of 84 preterm infants sampled longitudinally for first 2.5 months of life

Using a powerful combination of metagenomic shotgun sequencing and functional metagenomic selections, we deeply interrogated the gut microbiota and resistome of a cohort of 84 preterm infants from this cohort sampled longitudinally (N=401 stools) during their stay in the NICU (Table 4-1, Figure 4-S2,4-S3). All infants included in our study were born prematurely (<33 weeks gestational age; median (IQR) of 27 (25, 29) weeks) and had low birth weights (<1800g; median (IQR) of 865 (718, 1141) g). To study the
effect of antibiotic therapy on preterm infant gut microbiota development, we specifically selected a cohort stratified by antibiotic use. While all but two infants out of the 84 in our cohort subset received antibiotic therapy within 24 hours of birth, 39% of preterm infants included in our study did not receive additional antibiotic treatment during their stay in the NICU after the first week of life. This subset of preterm infants were used as control individuals to study the effect of specific antibiotic treatments between 1 and 10 weeks of life, when the gut microbiota and resistome are still actively developing. Infants who did receive subsequent antibiotics received a median (IQR) of 21 (16, 32) total days of such therapy during their stay in the NICU, with the most common agents being gentamicin, vancomycin, and ampicillin (Table 4-S1). Samples were analyzed over time, ranging from 6.1 to 157.72 days of life with a median (IQR) of 3 (2, 6) samples analyzed per individual. Including recent antibiotic exposure, additional subject and clinical variables were recorded to control for impact on the developing preterm infant gut microbiota (e.g., overall health indications, postmenstrual age, hospital environment, enteral feeding, other medications, sequencing depth, etc.). In total, we sequenced and analyzed 551 GB of preterm infant gut metagenomes (1.37 ± 1.17 GB per sample) and a total of 107 GB of functionally selected preterm infant resistomes (5.1 ± 3.6 GB per sample).

4.2.2 Impact of early life antibiotics on preterm infant gut microbiota development

We profiled the bacterial composition of the preterm infant gut with metagenomic shotgun sequencing using unique clade-specific marker genes from approximately 17,000 reference genomes. Clade-specific marker genes allowed both species-level resolution and accurate relative abundance estimations of bacterial species composition of these communities, and offer advantages over traditional 16S rRNA marker gene sequencing [186] which often permits only classification to the class or genus level. We show
Table 4-1: Details of preterm infant cohort analyzed in this study.

<table>
<thead>
<tr>
<th></th>
<th>Early Antibiotic Exposure Only (N=33)</th>
<th>Early + Subsequent Antibiotic Exposure (N=51)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth weight, g, median (IQR)</td>
<td>1147 (1000, 1330)</td>
<td>790 (675, 855)</td>
</tr>
<tr>
<td>Gestational age at birth, weeks</td>
<td>29 (27, 30)</td>
<td>25 (24, 27)</td>
</tr>
<tr>
<td>Gender, M/F</td>
<td>42300</td>
<td>25/26</td>
</tr>
<tr>
<td>Route of delivery, C-section/vaginal</td>
<td>21/12</td>
<td>30/21</td>
</tr>
<tr>
<td>CRIB Score, median (IQR)</td>
<td>5 (3, 8)</td>
<td>11 (8, 13)</td>
</tr>
<tr>
<td>Total antibiotic exposure, days</td>
<td></td>
<td></td>
</tr>
<tr>
<td>median (IQR), n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gentamicin</td>
<td>2 (1, 2), 24</td>
<td>8 (4, 13), 49</td>
</tr>
<tr>
<td>Ampicillin</td>
<td>2 (2, 2), 30</td>
<td>2 (2, 2), 49</td>
</tr>
<tr>
<td>Vancomycin</td>
<td>0 (0, 0), 5</td>
<td>8 (5, 13), 48</td>
</tr>
<tr>
<td>Meropenem</td>
<td>none</td>
<td>0 (0, 6), 19</td>
</tr>
<tr>
<td>Cefotaxime</td>
<td>none</td>
<td>0 (0, 1), 15</td>
</tr>
<tr>
<td>Ticarcillin-clavulanate</td>
<td>none</td>
<td>0 (0, 2), 15</td>
</tr>
<tr>
<td>Clindamycin</td>
<td>none</td>
<td>0 (0, 0), 9</td>
</tr>
<tr>
<td>Cefazolin</td>
<td>none</td>
<td>0 (0, 0), 9</td>
</tr>
<tr>
<td>Trimethoprim-sulfamethoxazole</td>
<td>none</td>
<td>0 (0, 0), 4</td>
</tr>
<tr>
<td>Feedings+,median (IQR), n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human Milk (All)</td>
<td>37 (32, 47), 33</td>
<td>31 (21, 58), 51</td>
</tr>
<tr>
<td>Human Milk (Maternal)</td>
<td>35 (12, 47), 33</td>
<td>30 (20, 53), 50</td>
</tr>
<tr>
<td>Human Milk (Donor)</td>
<td>0 (0, 0), 5</td>
<td>0 (0, 0), 5</td>
</tr>
<tr>
<td>Formula</td>
<td>6 (0, 26), 23</td>
<td>15 (1, 35), 38</td>
</tr>
<tr>
<td>Fortification</td>
<td>44 (32, 50), 33</td>
<td>44 (22, 56), 48</td>
</tr>
<tr>
<td>Bacterial culture positive, n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gram+/Gram- (n - individuals)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blood</td>
<td>0/0 (0)</td>
<td>7/5 (10)</td>
</tr>
<tr>
<td>Tracheal</td>
<td>0/0 (0)</td>
<td>31/60 (23)</td>
</tr>
<tr>
<td>Urine</td>
<td>0/0 (0)</td>
<td>2/2 (4)</td>
</tr>
</tbody>
</table>
that only six bacterial species are both highly prevalent (at least one of these species is present in 99.8% of samples) and highly abundant with median (IQR) relative abundance of 51 (25, 72)% across all preterm infant gut microbiota (Figure 4-1a). The majority of these species belong to those commonly associated with MDR nosocomial infections [176, 179]. Of gut microbial communities where >50% of the population consists of one of these species, we observed a significant developmental trend in control individuals from communities dominated by skin-associated bacteria (e.g., *Staphylococcus epidermidis*) to communities dominated by Gram negative facultative species whose members are commonly associated with infection (e.g., *Escherichia coli* and *Enterobacter cloacae*) (Figure 4-1b; P<0.001, Kruskall-Wallis; *P<0.05, Dunn’s Test with Benjamini-Hochberg correction for multiple comparisons). Previous studies using 16S rRNA marker gene profiling have reported a bacterial Class-level developmental progression of the preterm infant gut microbiota composition from Bacilli to Gammaproteobacteria [120]. Our data not only confirm these predictions at species-level resolution, they demonstrate that the developmental trajectory is driven primarily by *Klebsiella*, *Escherichia*, and *Enterobacter* species replacement of *Enterococcus* and *Staphylococcus* species (Figure 4-1c).

The gut microbiota in our preterm infant cohort generally increases in species richness with age (P<0.001; ANOVA), while total antibiotic exposure results in a significant reduction in species richness (P<0.001; ANOVA). Decreased species richness and diversity of the gut microbiota in infancy and throughout life is associated with a number of host pathologies [59, 216, 75, 2]. To investigate the effect of specific antibiotic treatments on the species richness of the developing preterm infant gut microbiota, we calculated the log fold-change of species richness from immediately before to immediately after antibiotic treatment (within 48 hours of starting or stopping the antibiotic) compared to log fold-change of species richness in control individuals (no antibiotic treatment outside the first week of life) over intervals representing all antibiotic treatment durations (Figure 4-2a). With the exception of gentamicin, each specific antibiotic reduces species richness
Figure 4-1: Composition and development of the preterm infant gut microbiota.

(a) Six bacterial species are both prevalent (at least one of these species is present in 99.8% of samples) and highly abundant (at least one of these species comprise >50% of community in >50% of samples). (b) Significant developmental trend of abundant species composition over time in control individuals (P<0.001, Kruskall-Wallis; *P<0.05, Dunn's test with correction for multiple comparisons.). Significant for all pair-wise comparisons indicated. (c) Colonization progression and species composition of preterm infant gut microbiota summarized by postmenstrual age (gestational age plus chronological age) in weeks.
on average, while the gut microbiota of control individuals increase in species richness on average over similar intervals. To identify significant effects of specific antibiotics while controlling for factors that may contribute to preterm infant gut microbiota development, we used a generalized linear mixed model (GLMM) with individual as a random effect (Figure 4-2b). All experimental and clinical variables (including overall health indications, postmenstrual age, birth weight, enteral feeding, maternal intrapartum antibiotic exposure, infections, sequencing depth, etc.) were used in model selection and removed individually through backwards elimination at a significance level of p<0.05. We identified postmenstrual age and human breast milk as significant contributors to increased species richness over time (***P<0.001). In contrast, while controlling for these variables as well as overall health of the infant as determined by CRIB (Clinical Risk Index for Babies) score, our model identified three broad-spectrum antibiotic therapies that significantly decrease species richness of the preterm infant gut microbiota: meropenem, cefotaxime, and ticarcillin-clavulanate (**P<0.01,*P<0.05). Our model, with only six fixed-effects predictors, explains 33% of the variance in species richness and explains 67% of the variance when including individual as the random effect.

4.2.3 Functional profiling of the preterm infant gut resistome

To investigate the role of the preterm infant gut-associated resistome in response to antibiotic therapies, we first used functional metagenomic selections [195, 71, 151, 162, 72, 150] to deeply characterize the reservoir of AR genes associated specifically with preterm infant gut microbiota. Functional metagenomic selections allow culture unbiased interrogation of resistomes, identifying full-length, functionally verified AR genes without reliance on sequence similarity to previously established AR genes [89, 177]. We constructed metagenomic libraries from 21 representative preterm infant fecal samples, ranging in host age and total antibiotic exposure, totaling approximately 107 GB of bacterial
Figure 4-2: Antibiotic treatment significantly alters preterm infant gut microbiota development. (a) Species richness log fold-change of the preterm infant gut microbiota from directly before to directly after six different antibiotic treatments. Bars represent average species richness log fold-change. Control samples represent the species richness log fold-change over time in individuals with no antibiotic treatment outside the first week of life. Control samples represent similar mean and standard deviation of days between samples as in antibiotic treatment sets. (b) Postmenstrual age and human milk significantly increase gut microbiota species richness, while CRIB (Clinical Risk Index for Babies) score and three specific antibiotics, meropenem, cefotaxime, and ticarcillin-clavulanate, significantly decrease species richness (***P<0.001,**P<0.01;*P<0.05; GLMM with individual as random effect using 401 preterm infant gut samples). Error bars depict standard error.
DNA (Table 4-S2). We screened these libraries against 16 clinically relevant antibiotics (Table 4-S3), including antibiotics commonly used in the NICU [39] and in children in the US [35]. We assembled 2,004 contigs containing 794 AR genes as annotated using Resfams [78], identifying extensive resistance to beta-lactam, amphenicol, tetracycline, and polymyxin antibiotics (Figure 4-S4a). Culture unbiased approaches have previously emphasized the immense sequence novelty of bacterial AR genes in every habitat investigated, including the adult gut and soil [195, 72]. We demonstrate that the preterm infant gut resistome defies this trend: the majority of identified AR genes have high identity (median 95.8% amino acid identity) to previously sequenced proteins, compared to 58.5% and 50.8% median amino acid identity as previously reported for the human adult gut [195] and soil resistomes [72], respectively (Figure 4-3a). However, emphasizing the power of functional metagenomic selections [195, 71, 151, 162, 72, 150] combined with high-resolution annotation methods [78], these functionally identified AR genes from preterm infant microbiomes have low identity (median 25.4% amino acid identity) to AR databases [144] (Figure 4-3a). These results emphasize that while most proteins we identified through functional selections for AR had been previously sequenced, the proteins they encode had yet to be assigned any AR functionality.

Combining shotgun metagenomic sequencing with functional metagenomic selections, we present an unprecedented characterization of the abundance of functionally identified AR genes in the preterm infant gut microbiota. We find a substantial number of genes encoding resistance against antibiotics that are not used in NICUs [39], including chloramphenicol and tetracycline (Figure 4-4b). This is consistent with our observations of the gut resistomes of healthy term infants [151, 150], suggesting the resistome established very early in life likely represents antibiotic selection of the colonizing bacteria by exposure in other habitats, rather than direct antibiotic selection in infants. Early preterm infant gut resistome establishment is also characterized by a stable composition over time, with only 21% of AR gene classes significantly change during the first 10 weeks of
life (P<0.05, Figure 4-5). Specifically, MFS antibiotic efflux pumps and beta-lactamases significantly increase over time, while Qac multidrug efflux pumps and fluoroquinolone resistance genes significantly decrease over time (P<0.05, Figure 4-5). However, these could reflect the species in which they reside.

This diverse reservoir of AR genes is encoded largely by relatives of common MDR pathogen (Figure 4-3b-c). The three species (E. coli, E. cloacae, K. pneumoniae) encoding the highest number of AR genes that colonize the preterm infant gut microbiota include organisms belonging to the ESKAPE pathogen family, which pose urgent and severe clinical threats (as defined by the US Centers for Disease Control and Prevention) because of dwindling antibiotic options available to treat their infections [28, 164]. These species are widely distributed on hospital surfaces in NICUs [29] and were the source of infection for 25% of all culture positive suspected infections for the 461 premature infants enrolled in the cohort between July 2009 and December 2013. Of the total unique AR conferring fragments assembled from all antibiotic selections, 42% originated from one of these species, each of which are predicted to be resistant to over 50% of antibiotics tested (Figure 4-3c). We observe a high prevalence of multidrug resistance in the preterm infant gut microbiota. Of 794 functionally identified AR genes, 33% either conferred resistance to multiple antibiotics or were co-localized with another AR gene that encodes resistance to an independent antibiotic (Figure 4-3d). The most common co-selection relationship was exhibited with piperacillin, which shared >50% of identified MDR proteins or protein clusters with both ampicillin and amoxicillin (Figure 4-3d), indicating broad-spectrum beta-lactam resistance genotypes. Additionally, 21% exhibit MDR across multiple antibiotic classes, including co-selection of AR genes across all beta-lactams and both amphenicols and tetracyclines (Figure 4-3d).
Figure 4-3: Functional metagenomic selections identify novel antibiotic resistance gene clusters in Proteobacterial species in the preterm infant gut. (a) Amino acid identity between all functionally selected antibiotic resistance (AR) genes (identified using Resfams) from preterm infant gut resistomes and their top hit in the NCBI protein database (blue, n=710). Functionally AR genes for preterm infant gut resistomes and top hit in AR databases (red, n=707). Dashed lines represent mean amino acid identity for each group. While we identify known proteins, they have rarely been previously ascribed an AR function. (b) The number of unique resistance conferring fragments originating from MDR species identified from functional metagenomic selections. (c) Heatmap represent the number of unique resistance conferring fragments that originated from each bacterial species to each antibiotic (CH, chloramphenicol; TE, tetracycline; PE, penicillin; CX, ceftazidime; AP, ampicillin; PI, piperacillin; AXCL, amoxicillin-clavulanate; AX, amoxicillin; CZ, ceftazidime; COL, colistin). (d) Co-selection of MDR proteins and protein clusters. Circles represent unique AR proteins colored by resistance mechanism. Squares represent antibiotics that were included in functional metagenomic selections colored by antibiotic class. Lines connect antibiotic selections to the resistance proteins that conferred resistance to that antibiotic. The heatmap in the lower right corner depicts percent AR genes shared between antibiotic selections.
4.2.4 Preterm infant gut microbiota and resistome response to specific antibiotic therapies

To further investigate the compositional and functional response of the preterm infant gut microbiota to specific antibiotics, we calculated the average change in relative abundance of each of the six identified high prevalence/abundance bacterial species (Figure 4-1a) from directly before initiation to directly after termination of each antibiotic treatment (Figure 4-4a). For each treatment for which we observed a significant reduction in total species richness (Figure 4-2b), we find either a significant enrichment in specific species (\textit{S. epidermidis} enriched in meropenem treated individuals and \textit{K. pneumoniae} enriched in ticarcillin-clavulanate treated individuals) or significant depletion (\textit{E. coli} depleted in cefotaxime treated individuals) as compared to change in relative abundance of those specific species in control individuals over comparable time spans (Figure 4-4a, \text{*P}<0.05, \text{***P}<0.001, Wilcoxon Rank Sum, Huse-Bonferroni correction for multiple hypotheses). Notably, meropenem, a broad spectrum antibiotic with no AR genes detected in the microbiota of this population, appears to 'reset' the developmental trajectory of the preterm infant gut microbiota (Figure 4-1b), resulting in a significant increase in \textit{S. epidermidis}, the bacteria which dominate early preterm infant guts (Figure 4-4a). In addition to specific species significantly enriched or depleted after antibiotic treatment, we sought to identify the fraction of the resistome that is enriched following antibiotic treatment (Figure 4-4b). We identified 50 unique AR genes that were enriched greater than 10-fold from directly before initiation to directly after termination of a specific antibiotic treatment in the majority (>50%) of treated individuals. We show that the fraction of the preterm gut-associated resistome that is highly enriched following specific antibiotic treatments is both antibiotic and species specific (Figure 4-4b). The specific AR genes enriched following antibiotic treatment are largely exclusive to a specific antibiotic, with the exception of a number of overlapping genes enriched in ticarcillin-clavulanate and ampicillin treatments.
Large overlap in genes enriched following both gentamicin and vancomycin is expected as they are predominantly administered as a co-therapy in preterm infant populations. Further, the collection of AR genes enriched following each specific antibiotic treatment is largely contributed by a single bacterial species. For example, in meropenem and ticarcillin-clavulanate treated individuals, the majority of AR genes which are enriched are highly correlated ($p<0.001$, Pearson Correlation) with either *S. epidermidis* or *K. pneumoniae*, which are also significantly enriched following each antibiotic treatment ($P<0.05$, Wilcoxon Rank Sum), respectively (Figure 4-4). For cefotaxime treatment, we observe a significant depletion in the relative abundance of *E. coli*, resulting in enrichment of AR genes associated with *E. cloacae* (Figure 4-4).

While a small number of AR genes enriched following a specific antibiotic treatment provide a relevant AR function to that specific antibiotic (*e.g.*, beta-lactamases enriched in beta-lactam treatments), each treatment also substantially enriches collateral resistance genes against a wide range of antibiotics. For example, while we identify no AR genes that are known to encode resistance to meropenem (*e.g.*, carbapenemases), antibiotic treatment with meropenem results in relative enrichment of AR genes providing resistance to a wide range of fluoroquinolones, macrolides, tetracyclines, trimethoprim, and other beta-lactams (Figure 4-4b). We observe this type of collateral resistome change following every specific antibiotic treatment investigated, which is often a results from a diversity of AR genes being encoded by a single bacterial species whose abundance changes in response to the treatment.

We demonstrated that a subset of antibiotics (meropenem, cefotaxime, and ticarcillin-clavulanate) each result in consistent and significant reduction in species richness directly following termination of antibiotic treatment (Figure 4-2b). In contrast, one of the most common antibiotic treatment regimens in preterm infant populations, co-therapy with vancomycin and gentamicin, variably results in either substantial increases or substantial decreases in the species richness of different preterm infant gut microbiota (Figure 4-2a).
Figure 4-4: Species and antibiotic resistance gene enrichment following specific antibiotic treatments. (a) Average change in relative abundance of six high abundance species from directly before to directly after treatment with six different antibiotics. (*P<0.05, ***P<0.001; Wilcoxon Rank Sum, Holm-Bonferroni correction for multiple hypotheses; significance tested for averages with absolute value greater than 20%). (b) Grey shading represents antibiotic resistance (AR) genes enriched from directly before to directly after treatment at least 10-fold in over half of the individuals treated with the specified antibiotic. Pearson correlation coefficient was calculated for all species by all AR genes. Circles represent Pearson correlation for specified antibiotic by the six high abundance species across all 401 preterm infant gut microbiota where correlation is greater than 0.2 (P<0.001, Pearson Correlation, n=401). Black circles denote those where the Pearson correlation coefficient is greater than 0.8. AR genes considered include those identified in functional metagenomic selections. For AR genes identified by functional selection, each antibiotic selection for which the protein was identified is indicated directly to the left of the AR mechanism or class.
We sought to determine if key features of the preterm infant gut microbiota and/or resistome before treatment might predictably mediate this large variance in response. Using Random Forests classification [131], we can predict the species richness response to vancomycin and gentamicin treatment with only 15% out-of-bag error rate based on the relative abundance of only 2 species and 2 AR genes prior to treatment (Figure 4-5a). Based on this model, we hypothesize that this response is mediated by two bacterial species. Specifically, *Staphylococcus aureus* is likely responding to and susceptible to vancomycin, and *E. coli*, which likely encodes the two component CpxR/CpxA protein misfolding stress response system, is likely responding to and resistant to gentamicin [138] (Figure 4-5b). Across all preterm infant gut microbiomes, relative abundance of *cpxR* and *cpxA* is highly correlated (p<0.001, Pearson Correlation) with relative abundance of *E. coli* species (Pearson correlation coefficient: 0.93 and 0.74, respectively). Importantly, it appears that absence of *cpxR* in this two component stress response system abolishes the gentamicin resistance phenotype in *E. coli* as previously reported [138]. This is supported in our population: the single case in which our model incorrectly predicts species richness to decrease, we observe high abundance of *E. coli* and *cpxA* gene, but complete absence of the *cpxR* component (Figure 4-5c).

### 4.2.5 Discussion

Antibiotics are by far the most common drug prescribed in preterm infant populations [39]. We demonstrate that these agents have the potential for significant collateral damage to the commensal gut microbiota composition and function, resulting in short-term and long-term dysbiotic microbial states associated with human disease [211]. While some studies report the development of preterm infant gut microbiota and/or potential effect of early-life antibiotic therapy, these studies have been limited in cohort size, often just a single infant, include no functional characterization, or consider all antibiotic regimens
Figure 4-5: Predictable response of preterm infant gut microbiota species richness to gentamicin and vancomycin combination treatment. (a) Random forest classification confusion matrix for prediction of species richness response direction using relative abundance of 2 species and 2 antibiotic resistance (AR) genes before treatment. Model only includes treatments where a response was observed (i.e. species richness change not equal to zero) and both gentamicin and vancomycin were administered in combination. Random Forests classification resulted in 15% out-of-box error rate. (b) Two predictive species and two predictive AR genes identified by applying Random Forests classification and minimizing the out-of-box error with the fewest number of predictors. Inclusion of predictors was based on mean decrease in classification accuracy when the relative abundance of the species or AR genes were randomly permuted (mean decrease ± s.d., n=1000 replicates). (c) Relative abundance of predictive species and AR genes directly before antibiotic treatment with a combination of gentamicin and vancomycin.
equally [80, 120, 13, 172]. Here we present functional microbiome and resistome analysis of a large and diverse preterm infant cohort sampled longitudinally, stratified to include a subset of individuals with only very-early antibiotic exposure, who represent the baseline trajectory for normal microbiota development in this population, as well as a large subset of subjects with substantial subsequent antibiotic treatment. These analyses reveal significant antibiotic-specific microbial and compositional responses of the preterm infant gut microbiota. For the most commonly prescribed antibiotic regimen, co-therapy with vancomycin and gentamicin, we identified important species and AR gene biomarkers that can predict with high accuracy the short-term species richness response and therefore potential for disruption of the developing preterm infant gut microbiota. This work represents a key first step in advancing and informing precision, evidence-based recommendations for usage of antibiotics early in life in preterm infant populations that may limit disruptions in microbiota development.

4.3 Materials & Methods

4.3.1 Sample collection

All samples were collected as part of preterm infant neonatal microbiome cohort project at Washington University School of Medicine and approved by the Human Research Protection Office (approval number 201205152). Samples were obtained from infants hospitalized in the NICU at Saint Louis Children’s Hospital. Details outlining the human subjects and clinical study protocols can be found in [120]. Subjects and samples were selected to be included in our study to maximize ranges of both antibiotic exposure and of postmenstrual age. To achieve this, we stratified our selected cohort by antibiotic exposure, with a subset of individuals with early antibiotic exposure only (N=33) with no antibiotic exposure outside the first week of life and a subset of individuals with early and subsequent
antibiotic exposure (N=51) with a range of antibiotic exposures reflective of total antibiotic exposure in the greater neonatal microbiome cohort (Table 4-1). All individuals included in our study were born premature (<33 weeks) and 82 individuals have very low birth weight (VLBW; <1500 g). Two siblings of VLBW preterm infants were also included in our study with birth weights of 1530 g and 1710 g. Samples were selected from subjects included in the early and subsequent antibiotic exposure subset with stool samples available within 48 hours prior to antibiotic initiation as well as within 48 hours after completion of antibiotic therapy. See Table 4-1 for more details on individual antibiotic exposure for the entire cohort analyzed in this study, as well as the ‘Early Exposure Only’ and the ‘Early + Subsequent Exposure’ subsets. All stools produced were collected and stored as previously described [120]. In total, 401 samples collected longitudinally from 84 infants were included in this analysis.

4.3.2 Metagenomic DNA Extraction

Total metagenomic DNA was extracted from approximately 0.1g of preterm infant fecal samples using either the PowerMax Soil DNA Isolation Kit (MoBio Laboratories) following suggested protocols or using phenol:chlorophorm extraction methods following protocols described in [151]. Extraction method did not significantly affect species composition or species richness and was explicitly controlled for in all data analysis (see methods on statistical modeling). DNA was quantified using Qubit fluorometer BR assay kit (Life Technologies).
4.3.3 Comparison of term to preterm gut microbiota using 16S rRNA marker gene sequencing

We performed sequencing of the variable region (V4) of the 16S rRNA gene on a subset of 129 infants in our cohort in order to appropriately compare species richness and composition of the gut microbiota to age-matched term infants, which had previously been sequenced using similar methods. The 515F/806R PCR primers including Illumina flow-cell adapter sequences were used to amplify the V4 region using the Earth Microbiome Protocols [32] (described in more detail here: http://www.earthmicrobiome.org/emp-standard-protocols/16s):

1. Mix the following in 25 µl reaction prepared in 96-well plate: 13 µl H2O, 10 µl Five Primer Hot Master Mix (5Prime catalogue number 2200410), 0.5 µl forward primer (10 µM), 0.5 µl reverse primer mix (10 µM), 1.0 µl template DNA
2. PCR cycle temperature as follows: 94°C for 3 min, then 35 cycles of [94°C for 45s, 50°C for 60 s, 72°C for 90s], then 72°C for 10 min.

PCR products were cleaned using the Agencourt AMPure XP PCR Purification kit (Beckman Coulter catalogue number A63880) and quantified using Picogreen (Invitrogen catalogue number #P11496), both following manufacturers protocols. 16S rRNA gene amplicons were sequencing by 250bp paired-end sequencing on the Illumina MiSeq platform using custom primers (read 1: 5’-TAT GGT AAT TGT GTG CCA GCM GCC GCG GTA A-3’; read 2: 5’-AGT CAG TCA GCC GGA CTA CHV GGG TWT CTA AT-3’; and index: 5’-ATT AGA WAC CCB DGT AGT CCG GCT GAC TGA CT-3’) at a loading concentration of 8 pM with 25% PhiX spike-in.

Sequencing data was de-multiplexed by sample and operational taxonomic units (OTUs) were generated following the UPARSE pipeline [67] and USEARCH v7. Specifically, forward and reverse reads were merged (usearch -fastq_mergepairs -fastq_maxmergelen
258 -fastq_minmergelen 248 -fastq_truncqual 3 -fastq_maxdifs 0), the merged reads were quality filtered (usearch -fastq_filter -fastq_maxee 0.5), de-replicated (usearch -drep_fulllength -sizeout), sorted and singletons removed (usearch -sortbysize -minsize 2), clustered (usearch -cluster_otus), checked for chimeric sequences using the Gold database (usearch -uchime_ref -db gold.fa -strand plus -nonchimeras), OTUs were re-named (uparse/fasta_number.py), reads were mapped back to OTUs at 97% identity (usearch -usearch_global -strand both -id 0.97), and converted to the final OTU table (uparse/uc2otutab.py). Taxonomy was assigned using uclust against the Greengenes database (retrieved August 2013). All OTUs that were not identified by Greengenes were discarded from downstream analysis (as to compare appropriately to closed reference OTU picking described below).

For comparison with age-matched term infants, previously generated 16S rRNA gene sequencing of human gut microbiota was downloaded from MG-RAST (Project 98) [224]. A total of 10 gut microbiota samples collected from term infants in the Saint Louis Neonatal Microbiome Initiative at 1 and 2 months of age were used in our comparison. OTUs were picked as using the QIIME pipeline [32] and closed reference OTU picking (pick_closed_reference_otus.py). Both OTU tables were individually subsampled to 2,750 reads per sample (single_rarefaction.py). After subsampling, each table was summarized at both the family and genus level (summarize_taxa.py). As OTUs were picked independently between groups, this summarization is required for comparison. Significance of bacterial family enrichment between term and preterm infant gut microbiota in the first two months of life was performed using non-parameteric Student’s T-test with 9,999 permutations (group_significance.py).
4.3.4 Multiplexed shotgun metagenomic sequencing

Sequencing libraries were prepared from 500ng total metagenomic DNA per sample. DNA was sheared to a target size range of approximately 500-600bp using either the Covaris E210 sonicator with the following settings: intensity, 4; duty cycle, 10%; cycles per burst, 200; treatment time, 75s; temperature, 4°C; sample volume, 130 µl, or the Covaris E220 sonicator with the following settings: peak incident power, 140; duty cycle, 10%; cycles per burst, 200; treatment time, 75s; temperature, 7°C; sample volume, 130 µl. Sheared DNA was purified and concentrated using a MinElute PCR Purification Kit (Qiagen), eluting in 63 µl pre-warmed nuclease-free H₂O. Purified sheared DNA was then end-repaired and Illumina adapters were ligated using the following protocol:

1. Mix the following for a 25 µl reaction volume (performed 3 reactions per sample in order to have a 3:1 ratio of barcode:sample): 20 µl of purified sheared DNA, 2.5 µl T4 DNA ligase buffer with 10 mM ATP (10X, New England BioLabs), 1 µl dNTPs (1 mM, New England BioLabs), 0.5 µl T4 polymerase (3 U/µl, New England BioLabs), 0.5 µl T4 PNK (10 U/µl, New England BioLabs), and 0.5 µl Taq Polymerase (5 U/µl, New England BioLabs).

2. Incubate the reaction at 25°C for 30 min followed by 20 min at 75°C.

3. Prepare barcode mix: Forward and reverse sequencing adapters were stored in TES buffer (10 mM Tris, 1 mM EDTA, 50 mM NaCl, pH 8.0) and annealed by heating the 1 mM mixture to 95°C followed by a slow cool (0.1°C per second) to a final holding temperature of 4°C.

4. Add 2 µl of prepared barcode mix and 0.8 µl of T4 DNA ligase (New England Bio-Labs) to each end-repair reaction from step 1.

5. Incubate the reaction at 16°C for 40 min followed by 10 min at 65°C.
Reactions were pooled (3-8 samples/pool) and purified using a MinEluted PCR Purification Kit (Qiagen), eluting in 16 µl pre-warmed elution buffer. The pooled, adaptor-ligated, sheared DNA was then size-selected to a target range of 400-900 bp on a 2% agarose gel in 0.5X TBE, stained with GelGreen dye (Biotium) and extracted using a MinElute Gel Extraction Kit (Qiagen). The purified DNA was enriched using the following protocol:

1. Mix the following for a 25 µl reaction volume: 2 µl of purified DNA, 12.5 µl 2X Phusion HF Master Mix (New England BioLabs), 1 µl of 10 mM Illumina PCR Primer Mix (5’-AAT GAT ACG GCG ACC ACC GAG ATC-3’ and 5’-CAAGCAGA A GAC GGC ATA CGA GAT-3’), and 9.5 µl of nuclease-free H$_2$O.

2. PCR cycle temperature as follows: 98°C for 30 s, then 18 cycles of [98°C for 10s, 65°C for 30 s, 72°C for 30s], then 72°C for 5 min.

Amplified DNA was measured using the Qubit fluorometer HS assay kit (Life Technologies) and 10 nM of each sample were pooled for sequencing. Subsequently, samples were submitted for paired-end 150-bp sequencing using the Illumina HiSeq 2500 platform at GTAC (Genome Technology Access Center, Washington University in St Louis, USA). In total, 320 samples across 10 HiSeq lanes were sequenced to sufficient depth (1M reads) for microbiome and resistome analyses.

Prior to all downstream data analysis, Illumina paired-end shotgun metagenomic sequence reads were binned by barcode (exact match required), quality filtered using Trimmmomatic v0.30 [27] (java -Xms1024m -Xmx1024m -jar trimmomatic-0.30.jar PE -phred33 ILLUMINACLIP: TruSeq3-PE.fa:2:30:10 LEADING:6 TRAILING:6 SLIDINGWINDOW:4:15 MINLEN:60), and human DNA was removed using DeconSeq [182] using build 38 of the human genome using default parameters.
4.3.5 Calculating relative abundance of species from shotgun sequencing

Relative abundance of species was calculated from shotgun sequencing of samples using MetaPhlAn 2.0 [186] downloaded from Bitbucket with repository tag 2.0.0. MetaPhlAn 2.0 was run with the following parameters: --bowtie2db metaphlan2/db_v20/mpa_v20_m200 --bt2_ps sensitive --input_type multifastq --mpa_pkl metaphlan2/db_v20/mpa_v20_m200.pkl --ignore_viruses --ignore_eukaryotes --ignore_archea. Individual MetaPhlAn 2.0 relative abundance tables were merged using the metaphlan2/utils/merge_metaphlan_tables.py script. Species were classified as the major or dominant species in a sample if they comprised greater than 50% relative abundance (e.g., Figure 4-1b). The number of observed species in a sample was calculated from MetaPhlAn summary as the number of unique species-level bacterial organisms that were identified in a sample. For the subset of samples that we also sequenced the V4 region of the 16S rRNA gene, we calculated the number of observed species and calculated a significant correlation (P<0.0001) between these two methods.

4.3.6 Calculating relative abundance of antibiotic resistance genes from shotgun sequencing

Relative abundance of antibiotic resistance genes was calculated using ShortBRED (Short Better Read Extraction Dataset) (http://huttenhower.sph.harvard.edu/shortbred). A total of 6,731 proteins associated with antibiotic resistance were used as proteins of interest for identification of marker families using shortbred_identify.py with the following non-default parameters: –clustid 0.95. These proteins included all antibiotic resistance genes in the CARD [144] (retrieved 2014 October 20), a curated list of beta-lactamases from Lahey Clinic (http://www.lahey.org/Studies/), and antibiotic resistance proteins iden-
tified using functional metagenomic selections performed in the current study on preterm infant gut microbiota as well as from healthy infant microbiota [151], adult gut microbiota [195], and the soil [71, 72]. The reference protein database used for identification of antibiotic resistance specific markers included 24,874,711 proteins encoded in 6,629 microbial genomes downloaded from the IMG database [140]. A complete list of microbial proteomes included was previously reported [78]. Of the 2,425 antibiotic resistance protein families created after clustering at 95% identity, 2,419 had at least one marker for quantification of antibiotic resistance from short reads. In order to calculate relative abundance of resistance genes using short markers, shortbred_quantify.py was used with the following non-default parameters: –avgreadBP 142.

4.3.7 Construction of metagenomic libraries from preterm infant gut samples for functional selection

Between 11.7 ± 3.9 µg of purified extracted total metagenomic DNA was used as starting material for metagenomic library creation. To create small-insert metagenomic libraries, DNA was sheared to a target size of 3,000bp using the Covaris E210 sonicator following manufacturer’s recommended settings. Sheared DNA was size-selected to a range of 1,000-5,000bp by electrophoresis through a 1% low melting point agarose gel in 0.5X Tris-Borate-EDTA (TBE) buffer stained with GelGreen dye (Biotium). Size selected fragments were gel extracted using a QIAquick Gel Extraction Kit (Qiagen), eluting in 30 µl of warm nuclease-free H2O. Purified DNA was then end-repaired using the End-It DNA End Repair kit (Epicentre) with the following protocol:

1. Mix the following in a 50 µl reaction volume: 30 µl of purified DNA, 5 µl dNTP mix (2.5 mM), 5 µl 10X End-Repair buffer, 1 µl End-Repair Enzyme Mix and 4 µl nuclease-free H2O.
2. Mix gently and incubate at room temperature for 45 min.

3. Heat-inactivate the reaction at 70°C for 15 min.

End-repaired DNA was then purified using the QIAquick PCR purification kit (Qiagen) and quantified using the Qubit fluorometer BR assay kit (Life Technologies) and ligated into the pZE21-MCS-1 vector at the HincII site. The pZE21 vector was linearized at the HincII site using inverse PCR with PFX DNA polymerase (Life Technologies):

1. Mix the following in a 50 µl reaction volume: 10 µl of 10X PFX reaction buffer, 1.5 µl of 10 mM dNTP mix (New England Biolabs), 1 µl of 50 mM MgSO4, 5 µl of PFX enhancer solution, 1 µl of 100 pg µl-1 circular pZE21, 0.4 µl of PFX DNA polymerase, 0.75 µl forward primer (5' GAC GGT ATC GAT AAG CTT GAT 3'), 0.75 µl reverse primer (5' GAC CTC GAG GGG GGG 3') and 29.6 µl of nuclease free H2O to a final volume of 50 µl.

2. PCR cycle temperature as follows: 95°C for 5 min, then 35 cycles of [95°C for 45 s, 55°C for 45 s, 72°C for 2.5 min], then 72°C for 5 min.

Linearized pZE21 was size-selected (2,200bp) on a 1% low melting point agarose gel (0.5X TBE) stained with GelGreen dye (Biotium) and purified as described above. Pure vector was dephosphorylated using calf intestinal alkaline phosphatase (CIP, New England BioLabs) by adding 1/10th reaction volume of CIP, 1/10th reaction volume of New England BioLabs Buffer 3, and nuclease-free H2O to the vector elute and incubating at 37°C overnight before heat inactivation from 15 min at 70°C. End-repaired metagenomic DNA and linearized vector were ligated together using the Fast-Link Ligation Kit (Epicentre) at a 5:1 ratio of insert:vector using the following protocol:

1. Mix the following in a 15 µl reaction volume: 1.5 µl 10X Fast-Link buffer, 0.75 µl ATP (10 mM), 1 µl FastLink DNA ligase (2 U/µl), 5:1 ratio of metagenomic DNA to vector, and nuclease-free H2O to final reaction volume.
2. Incubate at room temperature overnight.

3. Heat inactivate for 15 min at 70°C.

After heat inactivation, ligation reactions were dialysed for 30 min using a 0.025 um cellulose membrane (Millipore catalogue number VSWP09025) and the full reaction volume used for transformation by electroporation into 25 µl E. coli MegaX (Invitrogen) according to manufacturer’s recommended protocols. Cells were recovered in 1 ml Recovery Medium (Invitrogen) at 37°C for one hour. Libraries were titered by plating out 0.1 µl and 0.01 µl of recovered cells onto Luria-Bertani (LB) agar (5 g yeast extract, 5 g NaCl, 10 g of tryptone, 12 g agar in 1 liter of water) plates containing 50 µg/ml kanamycin. For each library, insert size distribution was estimated by gel electrophoresis of PCR products obtained by amplifying the insert from 12 randomly picked clones using primers flanking the HincII site of the multiple cloning site of the pZE21 MCS1 vector (which contains a selectable marker for kanamycin resistance). The average insert size across all libraries was determined to be 3,000 bp, and library size estimates were calculated by multiplying the average PCR-based insert size by the number of titered colony forming units (CFUs) after transformation recovery. The rest of the recovered cells were inoculated into 50 ml of LB containing 50 µg/ml kanamycin and grown overnight. The overnight culture was frozen down with 15% glycerol and stored at -80°C for subsequent screening.

4.3.8 Functional selections for antibiotic resistance

For each preterm infant gut metagenomic library, selections for resistance to each of 16 antibiotics (at concentrations listed in Table 4-S3 plus 50 µg/ml kanamycin for plasmid library selection) was performed using Mueller-Hinton (MH) agar (2 g beef infusion solids, 1.5 g starch, 17 g agar, 17.5 g casein hydrolysate, pH 7.4, in a final volume of 1 liter). Of note, as our library host, E. coli, is intrinsically resistant to vancomycin, we
are unable to functionally screen for this antibiotic. Further, the use of kanamycin as the selective marker for the metagenomic plasmid library results in low-level cross-resistance with other aminoglycoside antibiotics, resulting in a higher required minimum inhibitory concentration for gentamicin. For each metagenomic library, the number of cells plated on each antibiotic selection represented 10x the number of unique CFUs in the library, as determined by titers during library creation. Depending on the titer of live cells following library amplification and storage, the appropriate volume of freezer stocks were either diluted to 100 µl using MH broth + 50 µg/ml kanamycin or centrifuged and reconstituted in this volume for plating. After plating (using sterile glass beads), antibiotic selections were incubated at 37°C for 18 h to allow the growth of clones containing an antibiotic resistant DNA insert. After overnight growth, all colonies from a single antibiotic plate (gut microbiota by antibiotic selection) were collected by adding 750 µl of 15% LB-glycerol to the plate and scraping with an L-shaped cell scraper to gently remove colonies from the agar. The liquid ‘plate scrape culture’ was then collected and this process was repeated a second time for a total volume of 1.5 mL to ensure that all colonies were removed from the plate. The bacterial cells were then stored at -80°C before PCR amplification of antibiotic-resistant metagenomic fragments and Illumina library creation.

4.3.9 Amplification and sequencing of functionally selected fragments

Freezer stocks of antibiotic-resistant transformants were thawed and 300 µl of cells pelleted by centrifugation at 13,000 revolutions per minute (r.p.m.) for two minutes and gently washed with 1 mL of nuclease-free H₂O. Cells were subsequently pelleted a second time and re-suspended in 30 µl of nuclease-free H₂O. Re-suspensions were then frozen at -20°C for one hour and thawed to promote cell lysis. The thawed re-suspension was pelleted by centrifugation at 13,000 r.p.m. for two minutes and the resulting supernatant was used as template for amplification of resistance-conferring DNA fragments by
PCR with Taq DNA polymerase (New England BioLabs):

1. Mix the following for a 25 µl reaction volume: 2.5 µl of template, 2.5 µl of ThermoPol reaction buffer (New England BioLabs), 0.5 µl of 10 mM deoxynucleotide triphosphates (dNTPs, New England Biolabs), 0.5 µl of Taq polymerase (5 U/µl), 3 µl of a custom primer mix, and 16 µl of nuclease-free H₂O.

2. PCR cycle temperature as follows: 94°C for 10 min, then 25 cycles of [94°C for 45 s, 55°C for 45 s, 72°C for 5.5 min], then 72°C for 10 min.

The custom primer mix consisted of three forward and three reverse primers, each targeting the sequence immediately flanking the HincII site in the pZE21 MCS1 vector, and staggered by one base pair. The staggered primer mix ensured diverse nucleotide composition during early Illumina sequencing cycles and contained the following primer volumes (from a 10 mM stock) in a single PCR reaction: (primer F1, CCGAATTCAATTAAAGAGGAGAAAG, 0.5 µl); (primer F2, CGAATTCAATTAAAGAGGAGAAAGG, 0.5 µl); (primer F3, GAATTCATTAAAGAGGAGAAAGGTAC, 0.5 µl); (primer R1, GATATCAAGCTTATCGATACCGTC, 0.21 µl); (primer R2, CGATATCAAGCTTATCGATACCG, 0.43 µl); (primer R3, TCGATATCAAGCTTATCGATACC, 0.86 µl). The amplified metagenomic inserts were then cleaned using the Qiagen QIAquick PCR purification kit and quantified using the Qubit fluorometer HS assay kit (Life Technologies).

For amplified metagenomic inserts from each antibiotic selection, elution buffer was added to PCR template for a final volume of 200 µl and sonicated in a half-skirted 96-well plate on a Covaris E210 sonicator with the following setting: duty cycle, 10%; intensity, 5; cycles per burst, 200; sonication time, 600s. Following sonication, sheared DNA was purified and concentrated using the MinElute PCR Purification kit (Qiagen) and eluted in 20 µl of pre-warmed nuclease-free H₂O. In the first step of library preparation, purified sheared DNA was end-repaired:
1. Mix the following for a 25 µl reaction volume: 20 µl of elute, 2.5 µl T4 DNA ligase buffer with 10 mM ATP (10X, New England BioLabs), 1 µl dNTPs (1 mM, New England BioLabs), 0.5 µl T4 polymerase (3 U/µl, New England BioLabs), 0.5 µl T4 PNK (10 U/µl, New England BioLabs), and 0.5 µl Taq Polymerase (5 U/µl, New England BioLabs).

2. Incubate the reaction at 25°C for 30 min followed by 20 min at 75°C.

Next, to each end-repaired sample, 5 µl of 1 µM pre-annealed, barcoded sequencing adapters were added (adapters were thawed on ice). Barcoded adapters consisted of a unique 7-bp oligonucleotide sequence specific to each antibiotic selection, facilitating the de-multiplexing of mixed-sample sequencing runs. Forward and reverse sequencing adapters were stored in TES buffer (10 mM Tris, 1 mM EDTA, 50 mM NaCl, pH 8.0) and annealed by heating the 1 µM mixture to 95°C followed by a slow cool (0.1°C per second) to a final holding temperature of 4°C. After the addition of barcoded adapters, samples were incubated at 16°C for 40 min and then for 10 min at 65°C. Before size-selection, 10 µl each of adapted-ligated samples were combined into pools of 12 and concentrated by elution through a MinElute PCR Purification Kit (Qiagen), eluting in 14 µl of elution buffer (10 mM Tris-Cl, pH 8.5). The pooled, adaptor-ligated, sheared DNA was then size-selected to a target range of 300-400 bp on a 2% agarose gel in 0.5X TBE, stained with GelGreen dye (Biotium) and extracted using a MinElute Gel Extraction Kit (Qiagen). The purified DNA was enriched using the following protocol:

1. Mix the following for a 25 µl reaction volume: 2 µl of purified DNA, 12.5 µl 2X Phusion HF Master Mix (New England BioLabs), 1 µl of 10 mM Illumina PCR Primer Mix (5’-AAT GAT ACG GCG ACC ACC GAG ATC-3’ and 5’-CAAGCAGA A GAC GGC ATA CGA GAT-3’), and 9.5 µl of nuclease-free H₂O.

2. PCR cycle temperature as follows: 98°C for 30 s, then 18 cycles of [98°C for 10s, 65°C for 30 s, 72°C for 30s], then 72°C for 5 min.
Amplified DNA was measured using the Qubit fluorometer HS assay kit (Life Technologies) and 10 nM of each sample were pooled for sequencing. Subsequently, samples were submitted for paired-end 101-bp sequencing using the Illumina HiSeq 2500 platform at GTAC (Genome Technology Access Center, Washington University in St Louis, USA). In total, three sequence runs were performed at 10 pM concentration per lane.

### 4.3.10 Functional metagenomic assembly and annotation

Illumina paired-end sequence reads were binned by barcode (exact match required), such that independent selections were assembled and annotated in parallel. Assembly of the resistance-conferring DNA fragments from each selection was achieved using PAR-FuMS [71] (Parallel Annotation and Re-assembly of Functional Metagenomic Selections), a tool developed specifically for the high-throughput assembly and annotation of functional metagenomic selections. Assembly with PARFuMS consists of: (1) three iterations of variable job size with the short-read assembler Velvet, (2) two iterations of assembly with Phrap, and (3) custom scripts to clean sequence reads, remove chimeric assemblies, and link contigs by coverage and common annotation, as previously described. Of the 336 antibiotic selections performed, 183 yielded antibiotic-resistant *E. coli* transformants (Figure 4-S4) and from each selection we assembled fragments larger than 500 bp and annotated antibiotic resistance genes (ARGs).

ORFs were predicted in assembled contigs using the gene finding algorithm MetaGeneMark [231] and annotated by searching amino acid sequences against an ARG specific pHMM database, Resfams [78] (http://www.dantaslab.org/resfams), with HMMER3 [70]. MetaGeneMark was run using default gene-finding parameters while hmmscan (HMMER3) was run with the option --cut_ga as implemented in the script annotate_functional_selections.py. Proteins were classified as ARGs if they had a significant hit to a Resfams pHMM using profile specific gathering thresholds. Of the 3,133
unique predicted proteins, 667 (21%) were classified as ARGs.

4.3.11 Percentage identity of selected ORFs against NCBI and AR gene-specific databases

Percentage identity of all ARGs were conducted via a BlastP query against both the NCBI protein Non-Redundant (NR) database (retrieved 20 August 2013) or a combined database of all proteins from CARD [144] and a curated list of beta-lactamases from Lahey Clinic (http://www.lahey.org/Studies/) to identify the top local alignment. Once the top local alignment was identified using BlastP, it was the used for a global alignment using the EMBOSS v.6.3.1 implementation of the Needleman-Wunsch global alignment algorithm using the needle program with the following non-default parameters: -gapopen-10 -gapextend=0.5.

4.3.12 Taxonomic classification of functionally selected fragments

All bacterial genomes (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.fna.tar.gz) and bacterial plasmids (ftp://ftp.ncbi.nlm.nih.gov/genomes/Plasmids/plasmids.all.fna.tar.gz) were downloaded from NCBI (retrieved 26 November 2014). BlastN was used to align all contigs assembled from functional metagenomic selections to bacterial genomes and plasmids. Fragments were assigned to a bacterial genome or plasmid if it had greater than 95% nucleotide identity over 95% of the fragment.

4.3.13 Multidrug resistance in preterm infant gut microbiota

In order to identify ARGs which were selected on multiple antibiotics (as in Figure 4-3d), all identified ARGs were clustered at 100% identity using CD-HIT [128] with the
Protein clusters that included antibiotic resistance proteins identified on multiple antibiotics were classified as MDR and could arise one of two ways: (1) the protein itself could be bi-functional (across antibiotic classes) or have a broad-spectrum of activity (within antibiotic classes) or (2) the antibiotic resistance protein could be co-localized on the same 1-5kb fragment with another protein that provides resistance to an independent antibiotic. Multidrug resistance was visualized in Figure 4-3d using Cytoscape [187]. Percent sharing of antibiotic resistance genes across antibiotics (as in Figure 4-3d inset) was calculated as:

\[
\left( \frac{2 \times \text{Number of shared ARGs between antibiotic 1 and antibiotic 2}}{\text{Number of ARGs selected on antibiotic 1} \times \text{Number of ARGs selected on antibiotic 2}} \right) \times 100.
\]

### 4.3.14 Statistical modeling of preterm infant gut microbiota species richness and ARG abundance

In order to formally model the trends in preterm infant gut microbiota species richness and ARG abundance, a generalized linear mixed model (GLMM) with a Poisson family was fit by maximum likelihood (Laplace Approximation) using the lme4 package in R. Individual preterm infant was defined as the random effect to control for longitudinal sampling. As the scales for predictor variables were diverse, all predictors were centered and scaled using the generic function, scale, in R with default parameters. The final model included five fixed effects variables following backward elimination of non-significant fixed effects: postmenstrual age (P<0.001), total days receiving fortified maternal human milk nutrition (P<0.001), total days receiving ticarcillin-clavulanate antibiotic therapy (P<0.01), total days receiving meropenem antibiotic therapy (P<0.05), and the total number of sequencing reads remaining after quality filtering and removal of human reads (P<0.01) in order to control for sequencing depth. Other experimental details, such as extraction
method, were initially included in the model; however, they did not significantly affect preterm infant gut microbiota species richness. The same modeling approach was applied to beta-lactamase resistance gene abundance as quantified by short shotgun reads being mapped to ShortBRED markers. After step-wise removal of fixed effects variables, three significant fixed effects variables were included in the final model: postmenstrual age (P<0.001), total days receiving meropenem antibiotic therapy (P<0.001), and the total number of sequencing reads remaining after quality filtering and removal of human reads (P<0.001) in order to control for sequencing depth. Marginal and conditional R² values were calculated using an implementation of Schielzeth and Nakagwa’s R² for generalized linear mixed effects models in R.

4.3.15 Random Forests classification of species richness response for vancomycin and gentamicin treated individuals

Random Forests classification was used to predict the species richness response following treatment with vancomycin and gentamicin based on the species (based on MetaPhlAn 2.0 relative abundance) and ARG composition (based on ShortBRED relative abundance) prior to treatment. The default parameters of the R implementation of the algorithm (R package ‘randomForest’) were used with the following non-default parameters: ntree=10,000, importance=T. The full model with all species and all ARGs known to provide resistance to gentamicin (no relevant vancomycin resistance was identified) was used to identified the most informative predictors and then predictors were removed in order to minimize both out-of-box error rate and the number of predictors.
4.3.16 Correlation of species and antibiotic resistance genes across all 401 preterm infant metagenomes

Pearson correlation coefficients reported in Figure 4-4b were calculated for all 546 antibiotic resistance genes identified in the preterm infant gut resistome by all samples (n=401) analyzed using the corrcor function from the Numpy package in Python. Only those with $r>0.2$ are depicted in Figure 4-4.

4.3.17 Determination of appropriate sequencing depth

In order to determine the appropriate sequencing depth necessary to fully characterize the composition and function of the preterm infant gut microbiota, we rarefied all samples and identified the sequencing depth for which we did not identify any new species using clade-specific markers in the majority of samples. First all quality-filtered samples with human reads removed were rarefied 10 times using the random_subsample.py script from the QIIME package [31] including all sampling depths included in the following range that were less than the total sequences in the sample of interest: 10000, 100000, 1000000, 2000000, 3000000, 5000000, 10000000, 15000000. For each subsample replicated and depth, the species richness was calculated from MetaPhlAn 2.0 output and averaged across all replicates for a particular sample and sequencing depth. Using this analysis, we show that the average sequencing depth ± s.d. for all of our samples occurs well into the plateaued portion of the rarefaction curve (Figure 4-S6a) and the majority of samples (77%) do not identify any additional species beyond the average sequencing depth (Figure 4-S6b).
4.4 Acknowledgements

The work presented in this chapter was done in collaboration with Bin Wang, Sara Ahmadi, Carey-Ann D. Burnham, Phillip I. Tarr, Barbara B. Warner, and Gautam Dantas and is currently under review for publication. P.I.T. and B.B.W. assembled the cohort and developed protocols and infrastructure to obtain the biologics and clinical metadata from the cohort. M.K.G. and G.D. designed and conceived the study. M.K.G. performed 16S rRNA gene sequencing for comparison to term infants. B.W. prepared shotgun metagenomic sequencing libraries. S.A. created metagenomic libraries, performed functional selections, and prepared sequencing libraries. M.K.G. performed functional metagenomic data analysis, shotgun metagenomic data analysis, statistical modeling. M.K.G. wrote the manuscript with contributions from C.A.B., P.I.T., B.B.W., and G.D.

We thank A. Moore for initial study conception and clinical insight, C. Hall-Moore for assistance in access to cohort samples, M. Wallace for cultured strain characterization, and members of the Dantas lab for thoughtful discussions of the results and analyses. We also thank families and clinical staff in the St. Louis Children’s Hospital Neonatal Intensive Care Unit for their cooperation with study, and Laura Linneman and Julie Hoffmann for their efforts in enrollment and data accrual. This work is supported in part by awards to G.D. through the Children’s Discovery Institute (MD-II-2011-117 and 127), the March of Dimes Foundation (6-FY12-394), and the National Institute of General Medical Sciences of the National Institutes of Health (R01-GM099538). This work is also supported in part by an award to Washington University School of Medicine through a Clinical and Translational Science Award (CTSA) Grant (UL1 TR000448). This collection was also supported by the US National Institutes of Health Grants UH3AI083265 and P30DK052574 (Biobank Core), along with funding from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, and Foundation for the National Institutes of Health (made possible by support from the Gerber Foundation). M.K.G. is a
Mr. and Mrs. Spencer T. Olin Fellow at Washington University and a National Science Foundation (NSF) graduate research fellow (DGE-1143954).

4.5 Supplementary Figures & Tables
Figure 4-S1: The preterm infant gut microbiota is lower in species richness and compositionally distinct from age-matched term infants. (a) Preterm infant gut microbiota in the first two months of life are significantly lower in species richness than age-matched term infants (***(P<0.001; Wilcoxon) as determined by 16S rRNA marker gene sequencing. (b) Gut microbiota composition of preterm infants (N=129) and age-matched term infants (N=10) colored by bacterial family. Arrows represent the bacterial families that are significantly enriched in term infant gut microbiota (P<0.05, non-parametric Student’s T-test with 10,000 permutations). (c) Preterm infant gut microbiota composition is not influenced by delivery mode (P=0.452; PERMANOVA) as determined by species profiling of shotgun metagenomic sequencing data. Only the first sample from each individual is included (N=70).
Figure 4-S2: 84 preterm infants and 401 samples included in this study with antibiotic treatments. (a) Day of life depicted along the x-axis for each preterm infant along the y-axis represented by Individual ID. Colored lines represent duration of a specific antibiotic treatment. Circles represent samples that were analyzed using shotgun sequencing. (b) Number of specific discrete antibiotic treatments where a fecal sample was collected and analyzed directly before initiation and directly after termination of antibiotic treatment (within 48 hours).
Figure 4-S3: 84 preterm infants and 401 samples included in this study with feeding. (Day of life depicted along the x-axis for each preterm infant along the y-axis represented by Individual ID. Colored lines represent duration of a specific feed type. Circles represent samples that were analyzed using shotgun sequencing.)
Figure 4-S4: Functional selections of 21 preterm infant gut metagenomes for resistance against 16 antibiotics. (a) Phenotypic results of selections. A dark grey cell means that a resistance phenotype was observed whereas white cells indicate the absence of any drug-tolerant transformants. (b) Relative abundance (RPKM) of functionally selected antimicrobial resistance genes with observed resistance to 11 antimicrobials across all 401 preterm infant gut metagenomes.
Figure 4-S5: Resistome development in preterm infants early in life. Total antibiotic resistance gene relative abundance binned by postmenstrual age. Rare antibiotic resistance genes are defined as resistance genes that comprise <3% of all postmenstrual age bins. Four classes significantly change over time (*P<0.05, **P<0.01, ***P<0.001; ANOVA).
Figure 4-S6: Rarefaction curves for species identified by sequencing depth (after quality-filtering and removal of human-sequence reads). (a) Rarefaction curve for the number of unique species identified by rarefied sequencing depth. Red bar represents mean sequencing depth of quality filtered sequences with grey shading representing ± s.d. (b) The percentage of samples where no additional species are identified with deeper sequencing depth than the indicated sequencing depth on the x-axis. Red bar represents mean sequencing depth of quality-filtered sequences with grey shading representing ± s.d.
Table 4-S1: Antibiotic exposure for only those infants who received a particular antibiotic.

<table>
<thead>
<tr>
<th>Antibiotic</th>
<th>Early Antibiotic Exposure (N=33) - median days (IQR), n</th>
<th>Early + Subsequent Antibiotic Exposure (N=51) - median days (IQR), n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gentamicin</td>
<td>2 (2, 4), 24</td>
<td>8 (5, 13), 49</td>
</tr>
<tr>
<td>Vancomycin</td>
<td>3 (3, 5), 5</td>
<td>8.5 (5, 13), 48</td>
</tr>
<tr>
<td>Ampicillin</td>
<td>2 (2, 2), 30</td>
<td>5 (2, 7), 49</td>
</tr>
<tr>
<td>Meropenem</td>
<td>n.a.</td>
<td>9 (6, 11), 19</td>
</tr>
<tr>
<td>Cefotaxime</td>
<td>n.a.</td>
<td>6 (2, 10), 15</td>
</tr>
<tr>
<td>Ticarcillin-Clavulanate</td>
<td>n.a.</td>
<td>6 (4, 9), 15</td>
</tr>
<tr>
<td>Clindamycin</td>
<td>n.a.</td>
<td>2 (1, 7), 12</td>
</tr>
<tr>
<td>Cefazolin</td>
<td>n.a.</td>
<td>1 (1, 2), 9</td>
</tr>
<tr>
<td>Sulfamethoxazole-Trimethoprim</td>
<td>n.a.</td>
<td>8.5 (6, 12), 4</td>
</tr>
</tbody>
</table>
Table 4-S2: Metagenomic libraries constructed from preterm infant stool samples.

<table>
<thead>
<tr>
<th>Library ID</th>
<th>Individual Postmenstral Age (Weeks)</th>
<th>Delivery Mode</th>
<th>Birth Weight</th>
<th>Library Size (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-A2</td>
<td>40.01</td>
<td>30.3</td>
<td>V</td>
<td>700</td>
</tr>
<tr>
<td>1-E2</td>
<td>55.01</td>
<td>25.6</td>
<td>V</td>
<td>760</td>
</tr>
<tr>
<td>1-F2</td>
<td>55.01</td>
<td>25.8</td>
<td>V</td>
<td>760</td>
</tr>
<tr>
<td>1-G3</td>
<td>107.01</td>
<td>30.5</td>
<td>CS</td>
<td>740</td>
</tr>
<tr>
<td>1-A5</td>
<td>107.02</td>
<td>32.0</td>
<td>CS</td>
<td>610</td>
</tr>
<tr>
<td>1-H5</td>
<td>120.01</td>
<td>30.4</td>
<td>CS</td>
<td>720</td>
</tr>
<tr>
<td>1-A6</td>
<td>122.01</td>
<td>32.2</td>
<td>CS</td>
<td>840</td>
</tr>
<tr>
<td>1-B6</td>
<td>122.01</td>
<td>32.6</td>
<td>CS</td>
<td>840</td>
</tr>
<tr>
<td>1-C6</td>
<td>126.01</td>
<td>32.4</td>
<td>V</td>
<td>740</td>
</tr>
<tr>
<td>4-C8</td>
<td>180.01</td>
<td>28.9</td>
<td>CS</td>
<td>1180</td>
</tr>
<tr>
<td>4-B7</td>
<td>180.01</td>
<td>35.6</td>
<td>CS</td>
<td>1180</td>
</tr>
<tr>
<td>1-F7</td>
<td>184.01</td>
<td>34.4</td>
<td>CS</td>
<td>690</td>
</tr>
<tr>
<td>1-G7</td>
<td>184.01</td>
<td>34.8</td>
<td>CS</td>
<td>690</td>
</tr>
<tr>
<td>4-D5</td>
<td>197.01</td>
<td>31.0</td>
<td>CS</td>
<td>1330</td>
</tr>
<tr>
<td>5-B1</td>
<td>197.01</td>
<td>33.9</td>
<td>CS</td>
<td>1330</td>
</tr>
<tr>
<td>1-H8</td>
<td>209.01</td>
<td>35.3</td>
<td>CS</td>
<td>670</td>
</tr>
<tr>
<td>5-D1</td>
<td>234.01</td>
<td>27.1</td>
<td>CS</td>
<td>830</td>
</tr>
<tr>
<td>5-A5</td>
<td>234.01</td>
<td>29.5</td>
<td>CS</td>
<td>830</td>
</tr>
<tr>
<td>5-F2</td>
<td>234.02</td>
<td>30.5</td>
<td>CS</td>
<td>510</td>
</tr>
<tr>
<td>2-B7</td>
<td>286.02</td>
<td>34.8</td>
<td>CS</td>
<td>990</td>
</tr>
<tr>
<td>2-B8</td>
<td>303.01</td>
<td>31.4</td>
<td>CS</td>
<td>830</td>
</tr>
</tbody>
</table>
Table 4-S3: Antibiotics and minimum inhibitory concentration used for functional selections.

<table>
<thead>
<tr>
<th>Antibiotic</th>
<th>Abbreviation</th>
<th>Minimum Inhibitory Concentration (MIC) - ug/ml</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amoxicillin</td>
<td>AX</td>
<td>16</td>
</tr>
<tr>
<td>Amoxicillin + Clavulanate</td>
<td>AXCL</td>
<td>16-8</td>
</tr>
<tr>
<td>Ampicillin</td>
<td>AP</td>
<td>64</td>
</tr>
<tr>
<td>Aztreonam</td>
<td>AZ</td>
<td>8</td>
</tr>
<tr>
<td>Cefepime</td>
<td>CP</td>
<td>8</td>
</tr>
<tr>
<td>Cefoxitin</td>
<td>CX</td>
<td>64</td>
</tr>
<tr>
<td>Ceftazidime</td>
<td>CZ</td>
<td>16</td>
</tr>
<tr>
<td>Chloramphenicol</td>
<td>CH</td>
<td>8</td>
</tr>
<tr>
<td>Ciprofloxacin</td>
<td>CI</td>
<td>0.5</td>
</tr>
<tr>
<td>Colistin</td>
<td>COL</td>
<td>8</td>
</tr>
<tr>
<td>Gentamicin</td>
<td>GE</td>
<td>16</td>
</tr>
<tr>
<td>Meropenem</td>
<td>ME</td>
<td>16</td>
</tr>
<tr>
<td>Penicillin G</td>
<td>PE</td>
<td>128</td>
</tr>
<tr>
<td>Piperacillin</td>
<td>PI</td>
<td>16</td>
</tr>
<tr>
<td>Tetracycline</td>
<td>TE</td>
<td>8</td>
</tr>
<tr>
<td>Tigecycline</td>
<td>TG</td>
<td>2</td>
</tr>
</tbody>
</table>
Chapter 5

Conclusions and Future Directions

5.1 Comparison of methods for studying antibiotic resistance encoded in microbial communities

In Chapters 2 & 3, I present two computational methods for improved identification and quantification of AR determinants in microbial communities and genomes. The first is an AR-specific pHMM database, Resfams. pHMMs are statistical models, or 'profiles', of multiple sequence alignments and can incorporate more information about functionally similar sequences than any single representative sequence [64]. pHMMs have been extensively utilized for high-throughput protein annotation and identification of general protein functions encoded in microbial communities [21, 84, 85, 69]. However, while there is an increasing emphasis on studying the entire reservoirs of AR genes encoded in diverse and uncultured microbial communities [73, 96], BLAST remains the primary method for identification and annotation of AR determinants [229, 135, 18]. In Chapter 2, I show that pairwise sequence alignment methods (e.g., BLAST) substantially bias towards well-studied environments, such as the human-associated communities and organisms, and by incorporating statistical models such as pHMMs, we can begin to correct this bias and
study antibiotic resistance across and between diverse microbial communities.

While pHMMs represent a substantial improvement over pairwise-sequence alignments, pHMMs require near full-length protein sequences, limiting their scope in metagenomic studies. In order to address the challenge of studying AR in short read datasets, I optimized ShortBRED, a short marker database, for AR functions, allowing the high-throughput study of AR across diverse and dynamic microbial communities. In Chapter 3, we show that this method both increases speed and sensitivity for identification of AR genes in metagenomic datasets over typical centroid-based homology search methods.

Both identification and quantification of functions of interest across diverse and dynamic microbial communities are significant challenges in studying microbial communities central to human health. Throughout my thesis, I demonstrate that by combining functional metagenomic selections and high-resolution annotation with Resfams, we can extensively catalogue and identify novel AR genes encoded in diverse microbial communities. Further, combining high-throughput shotgun metagenomic sequencing and ShortBRED markers built from these newly identified AR proteins increases power for highly specific quantification of AR reservoirs across a large number of microbial communities. While this powerful combination of experimental and computational methods significantly improves our ability to study the problem of AR, there are on-going challenges that warrant more study. First, while improvements over standard methods, both Resfams and ShortBRED still rely on previously identified AR determinants. This necessitates an iterative process of updates and manual curation of both the pHMMs and short marker databases as new AR genes are discovered. Second, we show that Resfams specificity varies across different antibiotic mechanisms, with ABC Transporters resulting in the highest number of false positive annotations (Figure 2-S7). Further work to identify key sequence signatures central to efflux of specific antibiotic substrates within this class of AR genes is necessary.
5.2 Impact of dissemination of antibiotic resistance on treatment of infectious disease

In 2013, the CDC published four core actions for the prevention of AR, the first of which is preventing the spread of resistance [34]. AR can be disseminated through the transmission of resistant bacteria or through HGT of AR genes. It has been hypothesized that there is a vast network of AR gene exchange between bacteria associated with people, animals, and the environment. This highly connected network makes it easy for AR genes to disseminate through microbial communities and ultimately be acquired by pathogenic bacteria [49]. While there is evidence for recent exchange of AR genes between soil bacteria and human pathogens [71], the rate at which this occurs and the mechanisms driving the exchange of genetic material through various nodes in this network remain unclear. For example, while there has been much focus on environmental AR as a major source of AR in the clinic [220, 222, 143], a recent study suggests that the rate of HGT in soil bacteria is minimal compared to in clinical environments [72]. Models to predict the dissemination patterns of AR genes through this network are expected to be of high value [11], however, these models are ultimately limited by our ability to identify and quantify AR genes in diverse microbial communities, including environmental sources. In Chapter 2, I demonstrate that by using pHMMs we are able quantitatively compare diverse reservoirs of AR genes. Further work enabled by methods like Resfams to characterize and model networks of AR gene dissemination through diverse environmental and clinical environments, including wastewater treatment plants [178], farmland exposed to antibiotic-treated manure [93], concentrated agriculture feeding operations [79], animal meat supply [217], etc. is necessary.
5.3 Clinical implications of early life antibiotic therapy

Antibiotics are the most prescribed medications in neonatal and pediatric populations in the US [39, 35, 228]. It is becoming increasingly evident that disruption of the gut microbiota during important developmental timepoints has significant metabolic and physiological implications [157, 45], however, the characterization of impact of antibiotics on microbial communities and function during development have largely been limited to 16S rRNA gene surveys or small cohorts with no ability for predictive or statistical modeling [172, 120, 29, 80, 13, 43, 188, 153]. In Chapter 4, I quantify and model the effect of specific antibiotic therapies on the gut microbiota and resistome in a large, longitudinally-sampled preterm infant cohort (84 infants, 401 samples). I show that three specific antibiotics, meropenem, cefotaxime, and ticarcillin-clavulanate all significantly reduce species richness, and significantly enrich or deplete specific bacterial species and AR genes. In contrast, vancomycin and gentamicin, two of the antibiotics most commonly administered to preterm infants, result in varied response in species richness. We show that this response is predictable with 85% accuracy based on the relative abundance of only two bacterial species and two AR genes prior to treatment. This predictability based on the prior microbiota and resistome state underscores the opportunity for precision microbiome therapies and opportunities for restorative measures directly following specific antibiotic regimens in specific individuals. Future controlled clinical studies and animal models are necessary in order to extend and validate our modeling framework.
Bibliography


147


