METAGENOMIC METHODS
TO INVESTIGATE MOBILE ELEMENT CONTEXT
AND NASCENT TRANSCRIPTION
IN THE HUMAN GUT MICROBIOME

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
In Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by
Albert Charles Vill
May 2022

METAGENOMIC METHODS
TO INVESTIGATE MOBILE ELEMENT CONTEXT
AND NASCENT TRANSCRIPTION
IN THE HUMAN GUT MICROBIOME

Albert Charles Vill, PhD
Cornell University 2022

Microbial systems are in continuous flux. Considering the human gut microbiome, changes in community composition are associated with differences in host immune regulation, metabolic function, and a litany of physiological metrics. A growing number of diseases can be etiologically explained by the presence of some strains or the absence of others, though myriad conditions associated with the microbiome cannot be cleanly attributed to the actions of single organisms. Indeed, the true nature of the human microbiome is analogized best as a tangled web of ecological co-dependency, a careful balance between resource antagonism and symbiotic negotiations at the Maginot Line of the gut epithelium.

Short-read sequencing is an invaluable tool for examining the nucleic acid content of microbiome samples, and the tools of metagenomic assembly are getting ever-better at partitioning reads into near-complete pseudo genomes. However, many important microbial genes are found on genetic constructs that are readily shared between bacterial cells. These constructs, called mobile genetic elements (MGEs), are difficult to assemble, and their promiscuity confounds reference-based mapping of taxa to functions. Getting to the ground truth of gene-taxa pairings requires that we extend classic metagenomic sequencing to retain information about *in situ* MGE context. Of course, the carriage of a particular gene does not tell us to what extent a gene is expressed in the gut niche, so metagenomic techniques must be paired with tailored transcriptomics methods to ultimately draw causal links from genes to bacteria to human cells. In this dissertation, I present the application of two sequencing techniques, Hi-C

and PRO-seq, to human microbiome samples, with the goal of contributing a partial framework for gaining greater insight into the tripartite interaction between bacterial cells, mobile genetic elements, and the human that encapsulates it all.

The human gut microbiome is a reservoir of antibiotic resistance genes (ARGs) that can be accessed by pathogens via horizontal gene transfer, leading to multidrug-resistant infections. Metagenomic short-read sequencing can reveal community composition and the presence of ARGs, but assembly alone is insufficient to link ARGs on extrachromosomal elements with their host strains, and culture of ARG-containing gut microbes is complicated by specific nutritive requirements and low oxygen tolerance. In Chapter 2 of this dissertation, I will discuss the application of metagenomic proximity ligation to probe the microbiomes of neutropenic patients with hematologic malignancies. Broadly, we observe individualistic networks of mobile gene carriage and increased exchange of antibiotic resistance genes in the guts of hospitalized patients, with implications for understanding the emergence of multi-drug resistant Enterobacteriaceae.

Transcriptional analyses of mixed bacterial communities can give valuable insights into the ecological and metabolic interactions of neighboring species. However, RNAseq of microbiomes is confounded by variable efficiency of ribosomal RNA depletion across organisms and the short half-lives of most bacterial mRNAs, meaning that only robust transcriptional changes are typically observed in bulk meta-transcriptomic experiments. In Chapter 3, I discuss the application of a minimally modified precision run-on sequencing protocol (PRO-seq) for run-on transcription from engaged prokaryotic RNA polymerase, allowing for the biotinylation and capture of nascent bacterial transcripts. We show that PRO-seq is replicable in both *E. coli* and diverse members of the human gut microbiome, and that PRO-seq gives information beyond that of RNAseq concerning RNA polymerase dynamics at metagenomic loci.

BIOGRAPHICAL SKETCH

Albert Vill is from South Plainfield, New Jersey, home to world-class wrestlers, superb pizza, and, coincidentally, the Cornell-Dubilier Superfund Site. He completed his Bachelor of Science in Biochemistry and Molecular Biology at Gettysburg College in 2016. As an undergraduate, he worked in the “Phages Rock” Lab at Gettysburg, jointly advised by Dr. Véronique Delesalle and Dr. Greg Krukonis, studying the genetic determinants of bacteriophage host range for various *Bacillus* species. In 2016, Albert enrolled in the Genetics, Genomics & Development Ph.D. program at Cornell University, and soon after joined the lab of Dr. Ilana Brito. As a graduate student, he focused on the development and application of genomics methods for microbial communities, primarily the human gut microbiome. Albert’s scientific interests include bacterial molecular evolution, host-microbe interactions, and the innumerable mechanisms by which phage and bacteria antagonize each other. His Ph.D. work has resulted in the publication of a first-author paper in *Nature Communications*, co-authored with Dr. Alyssa Kent, as well as a first-author manuscript in review at *Nature Microbiology*. Albert hopes that the contribution of his scientific insights to the welfare of mankind may one day outweigh the combined environmental impact of his server use and discarded pipette tips.

This dissertation is dedicated to my family, friends, and mentors:
To my parents and sisters for their endless encouragement.
To Ming and Pop for always advocating for my education.
To Mr. Shah and Mrs. Timko for their disdain for mediocrity.
To Véronique and Greg for giving me a chance.
And to Carolyn, my wife, for her love, support, and patience.

ACKNOWLEDGMENTS

First, I would like to thank my advisor, Prof. Ilana Brito, for always rewarding ambition and encouraging creativity. She is a model scientist and a caring mentor, and I have greatly enjoyed my time in her lab, even when she mistakes me for Peter. "*Thick skin!"*

Thank you to Prof. Eric Alani and Prof. Andrew Clark for their service on my Special Committee. Their collective insight has been invaluable in guiding my research and expanding my knowledge of the fields of genomics and molecular biology.

I give my heartfelt thanks to my Brito Lab compatriots. I don't take for granted the blessing of being surrounded by coworkers that are keen to converse and eager to help. In particular, I would like to acknowledge Dr. Qiaojuan Shi for her tireless work to support our research and keep the lab operational. Also, I want to thank Peter, my desk buddy and a self-proclaimed Cloning Wizard, for his experimental acuity, sardonic humor, and selfless friendship.

I thank my GGD cohort for their camaraderie, especially Dashiell, who has kept me well-acquainted with local wines and cheeses, and Felicia, without whom I could barely tell a Unix shell from a clam shell.

Lastly, I would like to acknowledge my funding sources during my Ph.D., including the CDC and NIH NHLBI for funding much of our work concerning antibiotic resistance genes in the human microbiome. I want to also thank the Cornell Center for Vertebrate Genomics for selecting me to be a Distinguished Scholar, and for fostering a diverse and supportive research community at Cornell.

TABLE OF CONTENTS

BIOGRAPHICAL SKETCH iv

ACKNOWLEDGMENTS vi

CHAPTER 1: Introduction 1

The genomic fluidity and functional flexibility of bacteria 1

The benefits and limitations of metagenomic sequencing 4

A brief but enlightening analysis of the genes at the nodes of short-read assembly graphs 5

Proximity ligation sequencing permits more complete metagenomes 8

The promise of nascent transcriptomics for understanding microbial functions 10

References 11

CHAPTER 2: Widespread transfer of mobile antibiotic resistance genes within individual gut microbiomes revealed through bacterial Hi-C 17

Abstract 17

Introduction 18

Results 25

Hi-C substantially improves antibiotic resistance gene-taxa associations 25

Horizontal gene transfer networks vary across individuals' gut microbiomes 39

Horizontal gene transfer is frequent and is elevated in neutropenic patients 43

Emergence of antibiotic resistance in pathogens and commensals 46

Remaining challenges linking bacteria with their mobile genetic elements 47

Discussion 48

Methods 49

Sample collection 49

Metagenomic sequencing 50

Proximity ligation 50

Mock community methods 52

Quality filtering and assembly 52

Metagenomic binning 53

Taxonomic Identification 53

Antibiotic resistance gene annotation 53

Mobile genetic element annotation 54

Sequence mapping 55

Cleanliness comparisons 55

Mobile and antibiotic resistance gene associations 55

Comparison of horizontal gene transfer between individual taxa 56

Antibiotic resistance gene and phage machinery gene host specificity 56
Network density 57
Measuring novel horizontal gene transfer during individuals' timecourses 57
Data and code availability 58
References 58
CHAPTER 3: Run-On Sequencing Reveals Transcriptional Dynamics Within The Human Microbiome 63
Abstract 63
Introduction 64
Results 66
Paired PRO-seq and RNA-seq discriminate promoters by sigma factors in E. coli 66
PRO-seq is suitable for diverse species of human-associated microbiota 70
PRO-seq captures concurrent transcription and cleavage of CRISPR RNAs 72
Concurrent transcription and cleavage also occurs at tRNA loci 76
RNAP pause-site motifs annotated in diverse species 88
Discussion 90
Materials and methods 95
E. coli heat-shock experiment 95
Human gut microbiome sample collection 95
RNAseq sample preparation and sequencing 95
PRO-seq sample preparation and sequencing 97
Metagenomic library preparation, sequencing, assembly, binning, and annotation 99
Transcriptomics data processing and analysis 100
Data and code availability 101
References 101
CHAPTER 4: Concluding remarks, and the future of meta'omics 105
References 106

CHAPTER 1: Introduction

The genomic fluidity and functional flexibility of bacteria

In 1997, the first complete *E. coli* genome was published. The single 4.6 Mb chromosome of *E. coli* strain K12 contained 4,288 protein-coding genes, more than a third of which had no known function [1]. Five years later, a comparison of uropathogenic *E. coli* genomes showed that only 39% of nonredundant proteins were shared across three strains [2]. In 2008, comparative genomics of 17 *E. coli* isolates identified more than 13,000 unique genes, with a set of nearly 2,200 genes common between all strains [3]. Then, in 2019, an analysis of 4,401 *E. coli* and *Shigella* genomes uncovered an astonishing 128,193 genes, of which just 2608 were shared by at least 99% of the genomes analyzed [4]. This trend, while best exemplified by the genomes of well-sampled pathogens, is common across all bacteria. For a given species, plotting the number of genomes against the proportion of genes gives a characteristic U-shaped distribution [5–7]; *i.e.* many genes are present in all or most strains, and many genes are present in only one or two strains. The genes shared across all strains comprise the so-called *core* genome, which includes genes essential to cellular function and propagation. The rest of the genes are supplementary, including genes promoting niche-specific metabolism [8], virulence [9–11], and resistance to chemical insults [12,13]; these are collectively dubbed the *accessory* or *peripheral* genome. Together, the core and accessory genomes define a *pangenome*, which is all the genes that have been found in any representative genome of that species. Over the last two decades, pangenomes have been curated and published for dozens of bacterial species and higher taxa [14,15], revealing extensive genomic mosaicism across the prokaryotic tree of life. Though pangenomics is a useful frame to ask questions about the genetic architecture unifying specific clades, more data seems to only beget more questions. Analyses of ever-larger databases of bacterial genomes have blurred the distinction between *core* and *accessory* elements: ostensibly essential genes can be functionally displaced by non-orthologous proteins [16–18], and genes labeled “non-essential” can be surprisingly persistent [19,20] or contribute to the adaptive evolution of emerging pathogens [21]. New approaches go beyond orthologous gene groups and account for genomic structure by encoding pangenome features at the nodes of weighted networks [22] or by building composite core genomes from strain genomes embedded in reference graphs [5]. While such methods enable estimation of gene retention *within* the set of related strains representing a population, they tell us very little about how genes are initially acquired by a species. Even in the era of genomics, the bacterial species is a concept that continues to defy rigid definition.

At the crux of this genomic fluidity is the phenomenon of horizontal gene transfer (HGT). In single- and multi-celled eukaryotes, inter-organismal gene transfer is taxonomically constrained and relatively rare [23]. In bacteria, however, HGT is common and may be carried out by diverse mechanisms including plasmid conjugation [24], transduction [25], transformation [26], extracellular vesicles [27], and virus-like particles [28]. The genetic payloads exchanged during HGT are often called mobile genetic elements (MGEs), and these elements themselves engage in ecological interplay that dictates their proliferation. For example, plasmids interfere with the replication of genetically similar plasmids through the expression of antisense RNAs that tightly control copy number by inhibition of replication machinery translation [29] or by directly binding the origin of replication [30]. Plasmids may also inhibit the proliferation of other plasmids through incompatibility of DNA-binding proteins necessary for partitioning plasmid copies during the separation of daughter cells [31,32]. Likewise, temperate bacteriophage are classified by immunity groups, within which a phage may preclude co-infection by another group member due to similarities in their genetic circuitry controlling lytic growth [33,34]. From the perspective of the bacterial cell, genetic similarity is the best predictor of whether two organisms can engage in HGT [35–37]. Shared ecology, as well, is highly correlated with HGT frequency across phyla [38,39]. Further complicating the study of bacterial HGT is the propensity of certain species for the transformation of naked DNA; the genus *Acinetobacter*, a pathogen of increasing prevalence [40], is a paragon of transformation ability and has been shown to indiscriminately incorporate both highly damaged DNA [41] and DNA with very small regions of homology [42–44].

Of course, the movement of genes and their persistence in cells can only be meaningfully explained through consideration of their functions. The survival and proliferation of bacteria requires that they respond quickly and robustly to environmental stimuli. As discussed, genetic heterogeneity is part of the bacterial strategy – different cells within a population will have different peripheral genomes that may provide a selective advantage in the presence of stressors or in specific environments. Notably, plasmids are common vectors for genes conferring resistance to antibiotics [45–47] heavy metals [48–51], and there are numerous examples of pathogens acquiring phage-encoded toxins [52–58]. Separate from the products of horizontally transferred genes, MGEs may be co-opted by bacteria for transcriptional control of niche-specific or lifecycle-critical functions. Examples include phagosome escape by competence system activation via prophage excision in *Listeria monocytogenes* [59,60], mutation rate control by reversible integration of a chromosomal island into the *mutSL* operon of *Streptococcus pyogenes* [61], and the timed excision of inactive phage remnants from the *sigK* locus of various Gram-positive bacteria to express the sigma factors required for endospore maturation [62–65]. In these ways, MGEs may become indispensable to their host chromosomes. Conversely, “static” components of bacterial chromosomes may also be transferred by lateral transduction [66]. Truly, the borders between bacterial species are little more than picket fences – readily hopped. It is helpful, then, to conceptualize bacterial chromosomes and MGEs not as discrete entities, but rather as consortia of genetic components along a continuum of mobilizability, whose movement within and between cells is a function of evolutionary distance, environmental co-habitation, and functional compatibility.

The benefits and limitations of metagenomic sequencing

As cited, studies on the contribution of MGEs to bacterial functional capacity and transcriptional regulatory networks have been largely constrained to cultured pathogens. From an anthropocentric perspective, this makes sense: we want to know about the bugs that make us sick. However, the fluidity of bacterial genomes implies that many pathogenic functions of interest may be derived from non-pathogenic bacteria. Concerning human pathogens, the human gut microbiome is frequently referenced as a persistent source from which pathogens may acquire genetic material via HGT [67–70]. It is paramount, then, that we have sensitive methods by which to interrogate microbiomes so that we may better understand the selective pressures and organismal interactions that promote HGT.

While culture is a powerful tool to interrogate microbial communities, the full diversity of the human gut microbiome has been difficult to recapitulate *in vitro* due to the complex nutrient requirements [71,72] and oxygen sensitivity [73,74] of its constituent species. Amplicon sequencing of 16S ribosomal rRNA genes from microbiome samples can be used to reliably estimate taxonomic relative abundance in microbial communities [75], and leveraging a multi-region framework can greatly increase phylogenetic resolution [76]. However, despite the availability of ever-better databases to predict microbiome metabolic capacity from taxonomic marker genes [77,78], inference of total microbiome genotypes from amplicon sequencing will always miss some genes, especially those encoded in peripheral pangenomes. Genotypic gaps can be partially filled by PCR-based methods that fuse taxonomic marker genes with sequences of interest [79,80], though such methods require target-specific primers and are thus low-throughout.

Currently, the most useful tool for untargeted assessment of the total genetic content of complex microbial communities is metagenomic sequencing. In brief, this involves the isolation of DNA from a sample, the preparation of short-read Illumina sequencing libraries, and the assembly of short reads into contiguous sequences (contigs) [81–83]. As a field, metagenomics has shed light on the genetic composition of countless environmental and host-associated microbiomes. However, one obvious drawback to assembly by contiguity is that DNA sequences that are not co-molecular *in situ* cannot be associated after sequencing, such as a chromosome and a plasmid that occupy the same cell. Metagenomic assemblies can be made better by binning, which is the process of associating contigs by alternative sequence-based metrics. Most binning programs use a combination of two metrics: (1) sequence composition, like GC content or *k*-mer frequencies, and (2) co-abundance measurements to group contigs that show similar sequence coverage variation [84–89]. However, binning programs disproportionately fail for plasmids [90], whose sequence composition can be greatly diverged from their hosts' chromosomes. Even MGEs like prophage and transposons often fail to assemble with their host chromosomes [91,92], despite the fact that they are contiguous with bacterial genomes.

A brief but enlightening analysis of the genes at the nodes of short-read assembly graphs

As a case in point, I sought to understand why assemblies fail for MGEs by interpreting the source material: short-read assembly graphs. Assembly graphs consist of segments of sequences connected by links, which are collapsed into contigs for downstream analyses. In the parlance of graph theory, segments are *nodes* and links are *edges*. Taking the graphical fragment assembly files output by metaSPAdes [82], I count the links associated with each segment and extract sequences incident at high-degree nodes. After filtering by coverage to preclude segments that are poorly assembled due to low depth, the remaining segments represent the set of sequences whose contig assignment is maximally ambiguous, from the perspective of the assembly software. I formalized this procedure as a small program (github.com/acvill/nodeSeqs) and applied it to six graphs, each recovered from a human oral microbiome assembly that I generated using ~50 million paired-end reads and default metaSPAdes parameters. The resulting sets of sequences include some stretches of simple repeats, but most of the extracted sequences lack any discernible repeat structure. To evaluate the putative functions of these sequences, I ran each sequence set through eggNOG-mapper [93], which is built to annotate novel sequences using clustered orthologous groups of proteins (COGs). With COG annotations, I then asked which functions were highly represented among high-degree segments. Plotting the top functions as proportions of the total number of segments for each sample, we see that transposases and recombinases, both common components of integrated mobile elements [94], are overrepresented among high-degree nodes in short-read assembly graphs (Figure 1.1). This is unsurprising, given the ubiquity of these elements and their sequence conservation across phylogenetic space [95], but this result gives empirical justification for the assertion that MGEs are underrepresented in metagenome-assembled genomes.

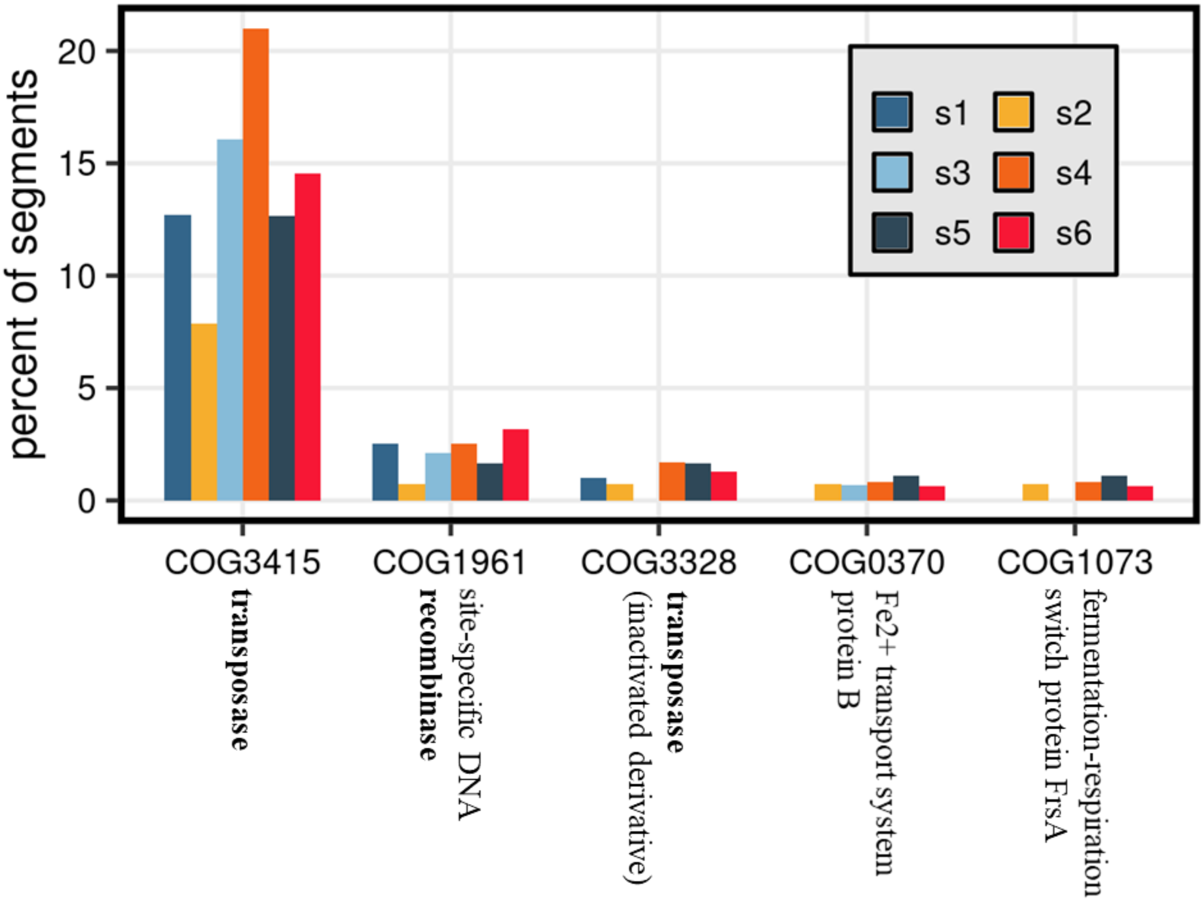


Figure 1.1. Bar graph showing the proportion of high-degree segments from six human oral microbiome assembly graphs (s1 – s6) that encode proteins belonging to certain COGs. The 5 COGs with the highest median percentage across samples are shown. Annotations of selected COGs whose functions are common in mobile elements are shown in bold.

Proximity ligation sequencing permits more complete metagenomes

So, what can be done to reliably link all classes of mobile genetic elements – extrachromosomal and integrated – with their bacterial host cells? One method that has shown great potential is all-against-all proximity ligation sequencing, commonly called Hi-C. In short, whole cells are chemically crosslinked to preserve protein-DNA interactions, then DNA is enzymatically fragmented and subject to dilute ligation. DNA molecules captured in the same crosslinked globule are ligated end-to-end, and sequencing across ligation junctions implicates different DNA molecules as residents in the same cellular compartment (Figure 1.2). Initially developed as a method to explore the spatial architecture of eukaryotic genomes [96], Hi-C has since been applied to microbiomes in order to assemble more complete bacterial genomes [97–99], link plasmids with host chromosomes [100–102], and probe phage-host infection networks [103].

Chapter 2 of this dissertation describes our application of metagenomic Hi-C to the microbiomes of cancer patients. These patients receive multiple courses of antibiotics to prevent nosocomial infections and are therefore vulnerable to multidrug-resistant pathogens. We find antibiotic resistance genes that are shared across diverse taxa and distinct networks of mobile gene transfer within individuals. Inclusion of proximity ligation libraries representing healthy microbiomes reveals that resistance genes and MGEs are dispersed across more taxa in neutropenic patients compared to healthy individuals. Gene exchange is most frequent between bacterial species in the same phylum, though transfer between Proteobacteria and Firmicutes is increased in our patient population during treatment, establishing a likely route for ARG transmission between enteric pathogens and commensal microbiota.

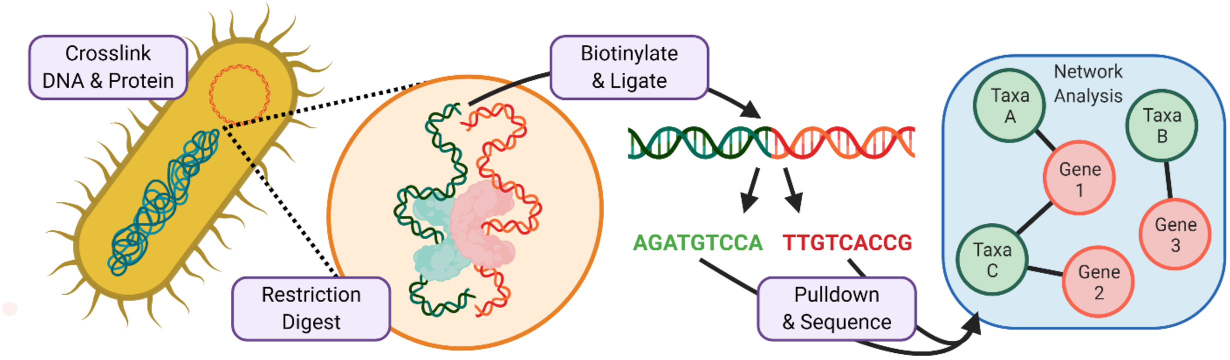


Figure 1.2. Metagenomic proximity ligation sequencing can associate plasmids with bacterial chromosomes, without the need to isolate and culture the host. Figure made with Biorender.com.

The promise of nascent transcriptomics for understanding microbial functions

The construction of complete metagenome-assembled genomes with concomitant mobile elements is a vital step in understanding the biology of microbiomes and their impacts on human health. However, the genetic contents of a microbiome only divulge the *potential* functions of a community. RNA sequencing (RNAseq) has enabled a deeper understanding of microbial communities and their overall effects on host health [104–106]. Yet, despite these gains, vanilla RNAseq analysis performed on microbiomes provides incomplete information about the transcriptional landscape. Bacterial RNAseq requires negative selection of ribosomal RNA, which comprises 85% of the total transcripts in bacterial cells [107,108], but rRNA depletion can be costly and introduce bias into sensitive datasets. Bacteria also perform extensive post-transcriptional modification [109], especially of non-coding RNAs [110,111], and these extra chemical moieties can interfere with reverse transcription during vanilla RNAseq protocols. Moreover, measurement of mature RNA abundance via RNAseq gives little information about real-time RNA polymerase activity, specifically with respect to transcriptional dynamics like initiation, pausing, and aborted polymerization [112–115]. Some bacterial transcripts are regulated by modulating their stability [116–119], though RNAseq alone cannot decouple changes in transcript stability from changes in the rate of polymerization. These dynamics are understudied outside common laboratory strains, even though further insight into the transcriptional regulatory mechanisms employed by diverse bacteria will enable deeper understanding of the functional plasticity of the human microbiome.

Chapter 3 of this dissertation discusses my application of precision run-on sequencing (PRO-seq) to assess the nascent transcriptomes of cultured and human-associated bacteria. Using an *E. coli* heat shock model as a proof-of-concept, we show that PRO-seq gives reproducible results for bacterial monocultures and captures transcriptional dynamics that are not apparent in paired RNA-seq libraries. Extending this technique to human gut microbiome samples, we observe concordance between PRO-seq and RNAseq signals across metagenome-assembled genomes. PRO-seq, however, is sensitive to transient transcriptional events at metagenomic features that are lost in RNA-seq, including transcription across CRISPR arrays and tRNA clusters that are co-transcriptionally processed. Altogether, nascent prokaryotic transcriptomics is a technique that can give a deeper understanding of the transcriptional dynamics of microbiomes.

In Chapter 4, I discuss the current state of cutting-edge metagenomics methods and end with a brief discussion about how these methods may be combined to comprehensively understand the genetic structure and function of human-associated microbial communities.

References

1. Blattner, F. R. *et al.* The complete genome sequence of Escherichia coli K-12. *Science (80-.).* **277**, 1453–1462 (1997).
2. Welch, R. A. *et al.* Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 17020–17024 (2002).
3. Rasko, D. A. *et al.* The pangenome structure of Escherichia coli: Comparative genomic analysis of E. coli commensal and pathogenic isolates. *J. Bacteriol.* **190**, 6881–6893 (2008).
4. Park, S. C., Lee, K., Kim, Y. O., Won, S. & Chun, J. Large-scale genomics reveals the genetic characteristics of seven species and importance of phylogenetic distance for estimating pan-genome size. *Front. Microbiol.* **10**, 1–12 (2019).
5. Colquhoun, R. M. *et al.* Pandora: nucleotide-resolution bacterial pan-genomics with reference graphs. *Genome Biol.* **22**, 1–30 (2021).
6. Gordienko, E. N., Kazanov, M. D. & Gelfand, M. S. Evolution of pan-genomes of Escherichia coli, Shigella spp., and Salmonella enterica. *J. Bacteriol.* **195**, 2786–2792 (2013).
7. Sela, I., Wolf, Y. I. & Koonin, E. V. Assessment of assumptions underlying models of prokaryotic pangenome evolution. *BMC Biol.* **19**, 1–15 (2021).
8. Vieira, G. *et al.* Core and panmetabolism in Escherichia coli. *J. Bacteriol.* **193**, 1461–1472 (2011).
9. Oliver, H. F., Orsi, R. H., Wiedmann, M. & Boor, K. J. Listeria monocytogenes σB has a small core regulon and a conserved role in virulence but makes differential contributions to stress tolerance across a diverse collection of strains. *Appl. Environ. Microbiol.* **76**, 4216–4232 (2010).
10. Georgiades, K. & Raoult, D. Comparative genomics evidence that only protein toxins are tagging bad bugs. *Front. Cell. Infect. Microbiol.* **1**, 7 (2011).
11. Georgiades, K. & Raoult, D. Genomes of the most dangerous epidemic bacteria have a virulence

repertoire characterized by fewer genes but more toxin-antitoxin modules. *PLoS One* **6**, (2011).

12. Li, L. *et al.* Comparative Genomic Analysis Reveals the Distribution, Organization, and Evolution of Metal Resistance Genes in the Genus Acidithiobacillus. *Appl. Environ. Microbiol.* **85**, 1–22 (2019).
13. Her, H. L., Lin, P. T. & Wu, Y. W. PangenomeNet: a pan-genome-based network reveals functional modules on antimicrobial resistome for Escherichia coli strains. *BMC Bioinformatics* **22**, 1–19 (2021).
14. Vernikos, G., Medini, D., Riley, D. R. & Tettelin, H. Ten years of pan-genome analyses. *Curr. Opin. Microbiol.* **23**, 148–154 (2015).
15. Anani, H., Zgheib, R., Hasni, I., Raoult, D. & Fournier, P. E. Interest of bacterial pangenome analyses in clinical microbiology. *Microb. Pathog.* **149**, 104275 (2020).
16. Charlebois, R. L. & Doolittle, W. F. Computing prokaryotic gene ubiquity: Rescuing the core from extinction. *Genome Res.* **14**, 2469–2477 (2004).
17. Wong, A. C. N. *et al.* The host as the driver of the microbiota in the gut and external environment of Drosophila melanogaster. *Appl. Environ. Microbiol.* **81**, 6232–6240 (2015).
18. Martínez-Carranza, E. *et al.* Variability of bacterial essential genes among closely related bacteria: The case of Escherichia coli. *Front. Microbiol.* **9**, 1–7 (2018).
19. Fang, G., Rocha, E. & Danchin, A. How essential are nonessential genes? *Mol. Biol. Evol.* **22**, 2147–2156 (2005).
20. Rancati, G., Moffat, J., Typas, A. & Pavelka, N. Emerging and evolving concepts in gene essentiality. *Nat. Rev. Genet.* **19**, 34–49 (2018).
21. Álvarez, V. E. *et al.* Crucial Role of the Accessory Genome in the Evolutionary Trajectory of Acinetobacter baumannii Global Clone 1. *Front. Microbiol.* **11**, (2020).
22. Gautreau, G. *et al.* PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. *PLoS Comput. Biol.* **16**, 1–27 (2020).
23. Van Etten, J. & Bhattacharya, D. Horizontal Gene Transfer in Eukaryotes: Not if, but How Much? *Trends Genet.* **36**, 915–925 (2020).
24. Cabezón, E., Ripoll-Rozada, J., Peña, A., de la Cruz, F. & Arechaga, I. Towards an integrated model of bacterial conjugation. *FEMS Microbiol. Rev.* **39**, 81–95 (2015).
25. Fillol-Salom, A. *et al.* Bacteriophages benefit from generalized transduction. *PLoS Pathog.* **15**, 1–22 (2019).
26. Johnston, C., Martin, B., Fichant, G., Polard, P. & Claverys, J. P. Bacterial transformation: Distribution, shared mechanisms and divergent control. *Nat. Rev. Microbiol.* **12**, 181–196 (2014).
27. Fischer, S. *et al.* Indication of horizontal DNA gene transfer by extracellular vesicles. *PLoS One* **11**, 1–22 (2016).
28. McDaniel, L. D. *et al.* High frequency of horizontal gene transfer in the oceans. *Science (80-.).* **337**, 911 (2012).
29. Novick, R. P. Plasmid Incompatibility. *Microbiol. Rev.* **51**, 381–395 (1987).
30. del Solar, G., Giraldo, R., Ruiz-Echevarría, M. J., Espinosa, M. & Díaz-Orejas, R. Replication and Control of Circular Bacterial Plasmids. *Microbiol. Mol. Biol. Rev.* **62**, 434–464 (1998).
31. Ebersbach, G., Sherratt, D. J. & Gerdes, K. Partition-associated incompatibility caused by random assortment of pure plasmid clusters. *Mol. Microbiol.* **56**, 1430–1440 (2005).
32. Schumacher, M. A. Bacterial plasmid partition machinery: A minimalist approach to survival. *Curr. Opin. Struct. Biol.* **22**, 72–79 (2012).
33. Mavrich, T. N. & Hatfull, G. F. Evolution of superinfection immunity in cluster A mycobacteriophages. *MBio* **10**, (2019).
34. Bondy-Denomy, J. *et al.* Prophages mediate defense against phage infection through diverse mechanisms. *ISME J.* **10**, 2854–2866 (2016).
35. Adato, O., Ninyo, N., Gophna, U. & Snir, S. Detecting Horizontal Gene Transfer between Closely Related Taxa. *PLoS Comput. Biol.* **11**, 1–23 (2015).
36. Bolotin, E. & Hershberg, R. Horizontally acquired genes are often shared between closely related

bacterial species. *Front. Microbiol.* **8**, 1–10 (2017).

37. Azad, R. K. & Lawrence, J. G. Towards more robust methods of alien gene detection. *Nucleic Acids Res.* **39**, 1–11 (2011).
38. Smillie, C. S. *et al.* Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**, 241–244 (2011).
39. Caro-Quintero, A. & Konstantinidis, K. T. Inter-phylum HGT has shaped the metabolism of many mesophilic and anaerobic bacteria. *ISME J.* **9**, 958–967 (2015).
40. Ayobami, O. *et al.* The incidence and prevalence of hospital-acquired (carbapenem-resistant) Acinetobacter baumannii in Europe, Eastern Mediterranean and Africa: a systematic review and meta-analysis. *Emerg. Microbes Infect.* **8**, 1747–1759 (2019).
41. Overballe-Petersen, S. *et al.* Bacterial natural transformation by highly fragmented and damaged DNA. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 19860–19865 (2013).
42. De Vries, J. & Wackernagel, W. Integration of foreign DNA during natural transformation of Acinetobacter sp. by homology-facilitated illegitimate recombination. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 2094–2099 (2002).
43. Hülter, N. & Wackernagel, W. Double illegitimate recombination events integrate DNA segments through two different mechanisms during natural transformation of Acinetobacter baylyi. *Mol. Microbiol.* **67**, 984–995 (2008).
44. Harms, K. *et al.* Substitutions of short heterologous DNA segments of intragenomic or extragenomic origins produce clustered genomic polymorphisms. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 15066–15071 (2016).
45. Li, Q., Chang, W., Zhang, H., Hu, D. & Wang, X. The role of plasmids in the multiple antibiotic resistance transfer in ESBLs-producing Escherichia coli isolated from wastewater treatment plants. *Front. Microbiol.* **10**, 1–8 (2019).
46. Rajer, F. & Sandegren, L. The Role of Antibiotic Resistance Genes in the Fitness Cost of Multiresistance Plasmids. *MBio* **13**, (2022).
47. McMillan, E. A. *et al.* Antimicrobial resistance genes, cassettes, and plasmids present in salmonella enterica associated with United States food animals. *Front. Microbiol.* **10**, 1–18 (2019).
48. Hernández-Ramírez, K. C. *et al.* A plasmid-encoded mobile genetic element from Pseudomonas aeruginosa that confers heavy metal resistance and virulence. *Plasmid* **98**, 15–21 (2018).
49. Wu, C. *et al.* The β-lactamase gene profile and a plasmid-carrying multiple heavy metal resistance genes of enterobacter cloacae. *Int. J. Genomics* **2018**, (2018).
50. Billman-jacobe, H. *et al.* pSTM6-275, a Conjugative IncHI2 Plasmid of Salmonella enterica That Confers Antibiotic and Heavy-Metal Resistance under Changing Physiological Conditions. *Antimicrob. Agents Chemother.* **62**, 1–6 (2018).
51. Bukowski, M. *et al.* Prevalence of Antibiotic and Heavy Metal Resistance Determinants and Virulence-Related Genetic Elements in Plasmids of Staphylococcus aureus. *Front. Microbiol.* **10**, 1–14 (2019).
52. Waldor, M. K. & Mekalanos, J. J. Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science (80-.).* **272**, 1910–1913 (1996).
53. Mirold, S. *et al.* Isolation of a temperate bacteriophage encoding the type III effector protein SopE from an epidemic Salmonella typhimurium strain. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 9845–9850 (1999).
54. Coleman, D. C. *et al.* Staphylocuccus aureus Bacteriophages Mediating the Simultaneous Lysogenic Conversion of P-Lysin, Staphylokinase and Enterotoxin A: Molecular Mechanism of Triple Conversion. *J. Gen. Microbiol.* **135**, 1679–1697 (1989).
55. Freeman, V. J. Studies on the virulence of bacteriophage-infected strains of Corynebacterium diphtheriae. *J. Bacteriol.* **61**, 675–688 (1951).
56. Fujii, N., Oguma, K., Yokosawa, N., Kimura, K. & Tsuzuki, K. Characterization of bacteriophage nucleic acids obtained from Clostridium botulinum types C and D. *Appl. Environ. Microbiol.* **54**, 69–73 (1988).

57. Barksdale, L. & Arden, S. B. Persisting bacteriophage infections, lysogeny, and phage conversions. *Annu. Rev. Microbiol.* **28**, 265–299 (1974).

58. Plunkett, G., Rose, D. J., Durfee, T. J. & Blattner, F. R. Sequence of Shiga toxin 2 phage 933W from Escherichia coli O157:H7: Shiga toxin as a phage late-gene product? *J. Bacteriol.* **181**, 1767–1778 (1999).

59. Rabinovich, L., Sigal, N., Borovok, I., Nir-Paz, R. & Herskovits, A. A. Prophage excision activates listeria competence genes that promote phagosomal escape and virulence. *Cell* **150**, 792–802 (2012).

60. Pasechnek, A. *et al.* Active Lysogeny in Listeria Monocytogenes Is a Bacteria-Phage Adaptive Response in the Mammalian Environment. *Cell Rep.* **32**, 107956 (2020).

61. Scott, J., Nguyen, S. V., King, C. J., Hendrickson, C. & McShan, W. M. Phage-like Streptococcus pyogenes chromosomal islands (SpyCi) and mutator phenotypes: Control by growth state and rescue by a SpyCi-encoded promoter. *Front. Microbiol.* **3**, 1–15 (2012).

62. Takemaru, K., Mizuno, M., Sato, T., Takeuchi, M. & Kobayashi, Y. Complete nucleotide sequence of a skin element excised by DNA rearrangement during sporulation in Bacillus subtilis. *Microbiology* **141**, 323–327 (1995).

63. Abe, K. *et al.* Developmentally-Regulated Excision of the SPβ Prophage Reconstitutes a Gene Required for Spore Envelope Maturation in Bacillus subtilis. *PLoS Genet.* **10**, (2014).

64. Abe, K. *et al.* Regulated DNA rearrangement during sporulation in bacillus weihenstephanensis KBAB4. *Mol. Microbiol.* **90**, 415–427 (2013).

65. Kim, K.-P. *et al.* Inducible Clostridium perfringens bacteriophages ΦS9 and ΦS63: Different genome structures and a fully functional sigK intervening element. *Bacteriophage* **2**, 89–97 (2012).

66. Hall, J. P. J. Is the bacterial chromosome a mobile genetic element? *Nat. Commun.* **12**, 12–15 (2021).

67. Penders, J., Stobberingh, E. E., Savelkoul, P. H. M. & Wolffs, P. F. G. The human microbiome as a reservoir of antimicrobial resistance. *Front. Microbiol.* **4**, 87 (2013).

68. Relman, D. A. & Lipsitch, M. Microbiome as a tool and a target in the effort to address antimicrobial resistance. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 12902–12910 (2018).

69. Yan, W., Hall, A. B. & Jiang, X. Bacteroidales species in the human gut are a reservoir of antibiotic resistance genes regulated by invertible promoters. *npj Biofilms Microbiomes* **8**, 1–9 (2022).

70. Baron, S. A., Diene, S. M. & Rolain, J. M. Human microbiomes and antibiotic resistance. *Hum. Microbiome J.* **10**, 43–52 (2018).

71. Lagier, J.-C. *et al.* Culturing the human microbiota and culturomics. *Nat. Rev. Microbiol.* **16**, 540–550 (2018).

72. Diakite, A. *et al.* Optimization and standardization of the culturomics technique for human microbiome exploration. *Sci. Rep.* **10**, 9674 (2020).

73. Singhal, R. & Shah, Y. M. Oxygen battle in the gut: Hypoxia and hypoxia-inducible factors in metabolic and inflammatory responses in the intestine. *J. Biol. Chem.* **295**, 10493–10505 (2020).

74. Jalili-Firoozinezhad, S. *et al.* A complex human gut microbiome cultured in an anaerobic intestine-on-a-chip. *Nat. Biomed. Eng.* **3**, (2019).

75. Caporaso, J. G. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 4516–4522 (2011).

76. Jones, C. B., White, J. R., Ernst, S. E., Sfanos, K. S. & Peiffer, L. B. Incorporation of Data From Multiple Hypervariable Regions when Analyzing Bacterial 16S rRNA Gene Sequencing Data. *Front. Genet.* **13**, 1–15 (2022).

77. Langille, M. G. I. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* **31**, 814–821 (2013).

78. Douglas, G. M. *et al.* PICRUSt2 for prediction of metagenome functions. *Nat. Biotechnol.* **38**, 669–673 (2020).

79. Spencer, S. J. *et al.* Massively parallel sequencing of single cells by epicPCR links functional genes with phylogenetic markers. *ISME J.* 1–10 (2015). doi:10.1038/ismej.2015.124

80. Diebold, P. J., New, F. N., Hovan, M., Satlin, M. J. & Brito, I. L. Linking plasmid-based beta-lactamases to their bacterial hosts using single-cell fusion pcr. *Elife* **10**, 1–29 (2021).

81. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).

82. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. MetaSPAdes: A new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).

83. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).

84. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).

85. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).

86. Lu, Y. Y., Chen, T., Fuhrman, J. A., Sun, F. & Sahinalp, C. COCACOLA: Binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge. *Bioinformatics* **33**, 791–798 (2017).

87. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).

88. Graham, E. D., Heidelberg, J. F. & Tully, B. J. Binsanity: Unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation. *PeerJ* **2017**, 1–19 (2017).

89. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).

90. Maguire, F. *et al.* Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic islands. *Microb. Genomics* **6**, 1–12 (2020).

91. Nelson, W. C., Tully, B. J. & Mobberley, J. M. Biases in genome reconstruction from metagenomic data. *PeerJ* **8**, 1–26 (2020).

92. Lapidus, A. L. & Korobeynikov, A. I. Metagenomic Data Assembly – The Way of Decoding Unknown Microorganisms. *Front. Microbiol.* **12**, (2021).

93. Cantalapiedra, C. P., Herņandez-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).

94. Khedkar, S. *et al.* Landscape of mobile genetic elements and their antibiotic resistance cargo in prokaryotic genomes. *Nucleic Acids Res.* **50**, 3155–3168 (2022).

95. Johansson, M. H. K. *et al.* Detection of mobile genetic elements associated with antibiotic resistance in Salmonella enterica using a newly developed web tool: MobileElementFinder. *J. Antimicrob. Chemother.* **76**, 101–109 (2021).

96. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598–606 (2015).

97. Burton, J. N., Liachko, I., Dunham, M. J. & Shendure, J. Species-Level Deconvolution of Metagenome Assemblies with Hi-C-Based Contact Probability Maps. *G3* **4**, 1339–1346 (2014).

98. Beitel, C. W. *et al.* Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ* **2**, e415 (2014).

99. Stewart, R. D. *et al.* Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat. Commun.* **9**, 870 (2018).

100. Marbouty, M. *et al.* Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. *Elife* **3**, e03318 (2014).

101. Press, M. O. *et al.* Hi-C deconvolution of a human gut microbiome yields high-quality draft

genomes and reveals plasmid-genome interactions. (2017). doi:10.1101/198713

102. Stalder, T., Press, M. O., Sullivan, S., Liachko, I. & Top, E. M. Linking the Resistome and Plasmidome to the Microbiome. *ISME J.* 484725 (2019). doi:10.1101/484725

103. Marbouty, M., Baudry, L., Cournac, A. & Koszul, R. Scaffolding bacterial genomes and probing host-virus interactions in gut microbiome by proximity ligation (chromosome capture) assay. *Sci. Adv.* **3**, e1602105 (2017).

104. Bikel, S. *et al.* Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: towards a systems-level understanding of human microbiome. *Comput. Struct. Biotechnol. J.* **13**, 390–401 (2015).

105. Benítez-Páez, A., Belda-Ferre, P., Simón-Soro, A. & Mira, A. Microbiota diversity and gene expression dynamics in human oral biofilms. *BMC Genomics* **15**, 1–13 (2014).

106. Franzosa, E. A. *et al.* Relating the metatranscriptome and metagenome of the human gut. *Proc. Natl. Acad. Sci. U. S. A.* **111**, (2014).

107. He, S. *et al.* Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nat. Methods* **7**, 807–812 (2010).

108. Stewart, F. J., Ottesen, E. A. & Delong, E. F. Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. *ISME J.* **4**, 896–907 (2010).

109. Antoine, L. *et al.* Rna modifications in pathogenic bacteria: Impact on host adaptation and virulence. *Genes (Basel).* **12**, (2021).

110. Marbaniang, C. N. & Vogel, J. Emerging roles of RNA modifications in bacteria. *Curr. Opin. Microbiol.* **30**, 50–57 (2016).

111. de Crécy-Lagard, V. & Jaroch, M. Functions of Bacterial tRNA Modifications: From Ubiquity to Diversity. *Trends Microbiol.* **29**, 41–53 (2021).

112. Belogurov, G. A. & Artsimovitch, I. Regulation of Transcript Elongation. *Annu. Rev. Microbiol.* **69**, 49–69 (2015).

113. Larson, M. H. *et al.* A pause sequence enriched at translation start sites drives transcription dynamics in vivo. *Science (80-.).* **344**, 1042–1047 (2014).

114. Mirkin, E. V. & Mirkin, S. M. Mechanisms of Transcription-Replication Collisions in Bacteria. *Mol. Cell. Biol.* **25**, 888–895 (2005).

115. Kang, J. Y., Mishanina, T. V., Landick, R. & Darst, S. A. Mechanisms of Transcriptional Pausing in Bacteria. *J. Mol. Biol.* **431**, 4007–4029 (2019).

116. Steglich, C. *et al.* Short RNA half-lives in the slow-growing marine cyanobacterium Prochlorococcus. *Genome Biol.* **11**, (2010).

117. Ho, K. & Ja, A. Epitranscriptomics: RNA Modifications in Bacteria and Archaea. *Regul. with RNA Bact. Archaea* 399–420 (2018). doi:10.1128/microbiolspec.rwr-0015-2017

118. Selinger, D. W., Saxena, R. M., Cheung, K. J., Church, G. M. & Rosenow, C. Global RNA half-life analysis in Escherichia coli reveals positional patterns of transcript degradation. *Genome Res.* **13**, 216–223 (2003).

119. Vargas-Blanco, D. A., Zhou, Y., Zamalloa, L. G., Antonelli, T. & Shell, S. S. MRNA Degradation Rates Are Coupled to Metabolic Status in Mycobacteria. *bioRxiv* 1–15 (2019). doi:10.1101/595199

CHAPTER 2: Widespread transfer of mobile antibiotic resistance genes within individual gut microbiomes revealed through bacterial Hi-C

This chapter is adapted from a 2020 paper in *Nature Communications*. Alyssa Kent and Albert Vill contributed equally to this publication:

> AG Kent, AC Vill, Q Shi, MJ Satlin, IL Brito. Widespread transfer of mobile antibiotic resistance genes within individual gut microbiomes revealed through bacterial Hi-C. *Nature Communications* **11**, 4379 (2020).

M.J.S. and I.L.B. conceived of the project and collected samples. A.K., A.C.V., and I.L.B. designed the experiments and performed the analyses. A.C.V. and Q.S. processed samples. A.K. wrote and compiled code. All authors contributed to the manuscript.

Abstract

The gut microbiome harbors a ‘silent reservoir’ of antibiotic resistance (AR) genes that is thought to contribute to the emergence of multidrug-resistant pathogens through horizontal gene transfer (HGT). To counteract the spread of AR, it is paramount to know which organisms harbor mobile AR genes and which organisms engage in HGT. Despite methods that characterize the overall abundance of AR genes in the gut, technological limitations of short-read sequencing have precluded linking bacterial taxa to specific mobile genetic elements (MGEs) encoding AR genes. Here, we apply Hi-C, a high-throughput, culture-independent method, to surveil the bacterial carriage of MGEs. We compare two healthy individuals with seven neutropenic patients undergoing hematopoietic stem cell transplantation, who receive multiple courses of antibiotics, and are acutely vulnerable to the threat of multidrug-resistant infections. We find distinct networks of HGT across individuals, though AR and mobile genes are associated with more diverse taxa within the neutropenic patients than the healthy subjects. Our data further suggest that HGT occurs frequently over a several-week period in both cohorts. Whereas most efforts to

understand the spread of AR genes have focused on pathogenic species, our findings shed light on the role of the human gut microbiome in this process.

Introduction

The acquisition of antibiotic resistance (AR) genes has rendered important pathogens, such as multidrug-resistant (MDR) Enterobacteriaceae and *Pseudomonas aeruginosa*, nearly or fully unresponsive to antibiotics. It is widely accepted that these so-called 'superbugs' acquire AR genes through the process of horizontal gene transfer (HGT) with members of the human microbiome with whom they come into contact[1]. The emergence of these MDR bacteria threatens our ability to perform life-saving interventions, such as curative hematopoietic cell transplants for patients with hematologic malignancies[2]. Furthermore, antibiotic use, required for vital prophylaxis in these patients, has been proposed as a trigger for HGT. Although tools are available to identify AR genes within the gut microbiome, and characterize their function [3], abundance[4,5] and their host-associations[6], no studies have attempted to monitor the bacterial host associations of AR genes and mobile elements during relatively short periods, such as during these patients' hospitalizations.

To determine the bacterial hosts of mobile AR genes, we utilized a high-throughput chromatin conformation capture (Hi-C) method aimed at sampling long-range interactions within single bacterial genomes[7,8,9]. Briefly, while cells are still intact, DNA within individual cells is crosslinked by formaldehyde. Cells are then lysed and the DNA is cut with restriction enzymes, biotinylated, and subjected to dilute ligation to promote intra-molecular linkages between crosslinked DNA. Crosslinking is reversed and then ligated DNA molecules are pulled-down and made into DNA libraries for sequencing. As is, this protocol has been used to improve metagenomic assemblies of bacterial genomes[10] and has identified a handful of strong plasmid-

and phage-bacterial host associations[11,12,13,14], suggesting that this technique could be applied to link mobile genes with specific taxa more broadly and to observe the process of HGT over time.

Here, we develop a modified version of current Hi-C protocols and analytical pipelines (Figure 2.1) in conjunction with metagenomic shotgun sequencing to surveil the bacterial taxa harboring specific mobile AR genes in the gut microbiomes of two healthy individuals and seven patients undergoing hematopoietic stem cell transplantation. These patients have prolonged hospitalizations during their transplant (21 ± 4 days) and often receive multiple courses of antibiotic therapy, increasing the likelihood of an MDR infection. As a result of their condition and treatment, these patients face mortality rates of 40-70% when bacteremic with carbapenem-resistant Enterobacteriaceae (CRE) or carbapenem-resistant *Pseudomonas aeruginosa*[15], and therefore represent a salient population for surveillance and one in which MDR pathogens may emerge and/or amplify under antibiotic selection. Gut microbiome samples for patients and healthy subjects were collected over a 2-3-week period, which, for the neutropenic patients started upon admission for transplant and continued during their hospitalization until neutrophil engraftment (Figure 2.2A).

We introduce a number of improvements to current bacterial Hi-C protocols to obtain gene-taxa associations. We change sample storage and optimize the choice of restriction enzymes to improve the congruence between the composition of metagenomic and Hi-C sequencing libraries (Figure 2.3). We also integrate Nextera XT sequencing library preparation directly into the Hi-C experimental protocol, streamlining operations and decreasing sample preparation time. Importantly, within diverse bacterial communities such as the gut microbiome, MGEs may be highly promiscuous and recombinogenic, complicating both assembly[16] and linkage analyses[17]. Therefore, we implement a computational workflow to assemble genomes, separating large integrated phage onto their own contigs, and allowing them to associate with genomes via binning or Hi-C connections. In a mock community of three organisms, each harboring an identifiable plasmid, we are able to confidently link each plasmid to its nascent genome (Figure 2.4).

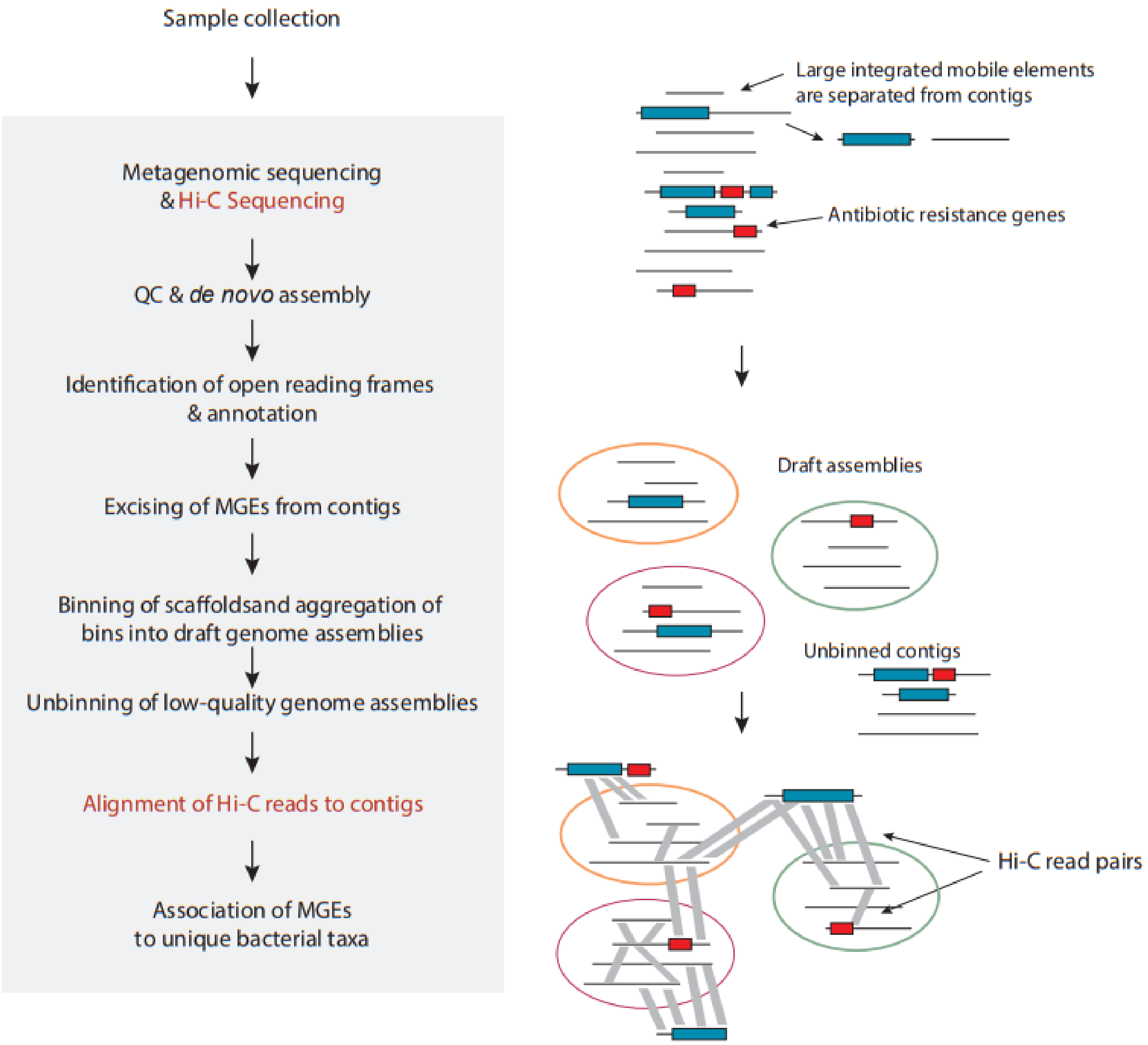

Sample collection
Metagenomic sequencing
& Hi-C Sequencing
QC & *de novo* assembly
Identification of open reading frames
& annotation
Excising of MGEs from contigs
Binning of scaffoldsand aggregation of
bins into draft genome assemblies
Unbinning of low-quality genome assemblies
Alignment of Hi-C reads to contigs
Association of MGEs
to unique bacterial taxa
Large integrated mobile elements
are separated from contigs
Antibiotic resistance genes
Draft assemblies
Unbinned contigs
Hi-C read pairs

Figure 2.1. Experimental and computational pipeline. Our pipeline for assigning mobile and AR genes to bacterial taxa utilizes data from metagenomic (black) and Hi-C (red) sequencing libraries. In brief, metagenomic samples are assembled using standard approaches. The resulting contigs (circles) are binned into draft assemblies and quality filtered. Bins are taxonomically annotated at the lowest level with >50% of bps assigned to a taxon using a weighted Kraken approach. Contigs containing AR or mobile genes are associated with metagenomic assemblies by residency or by Hi-C linkages requiring at least 2 readpairs linking the contig with the metagenomic assemblies. Associations of mobile or AR genes with specific taxa are made by clustering the genes of interest at 99% identity and counting each unique taxon once.

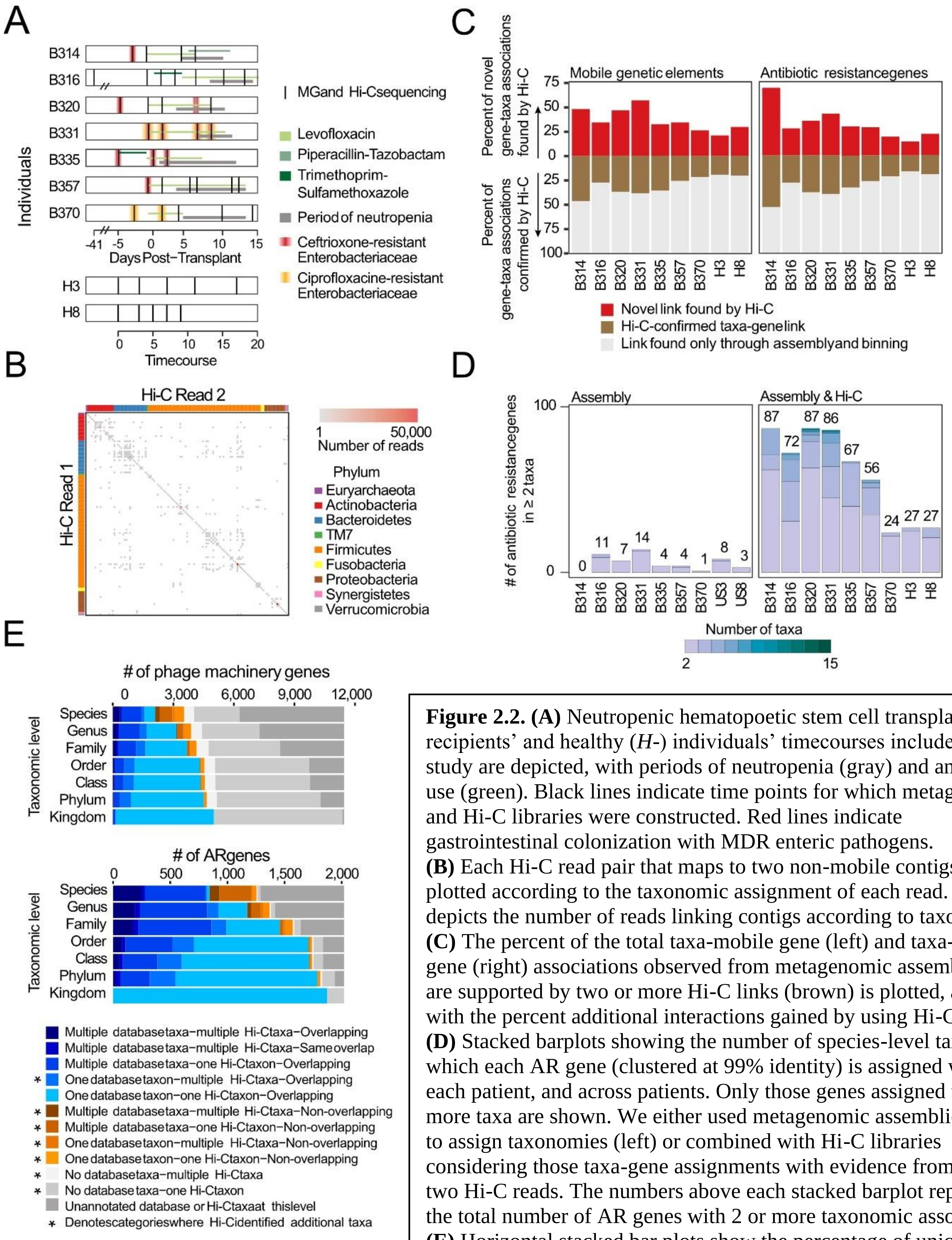


Figure 2.2. (A) Neutropenic hematopoetic stem cell transplant (*B-*) recipients' and healthy (*H-*) individuals' timecourses included in the study are depicted, with periods of neutropenia (gray) and antibiotic use (green). Black lines indicate time points for which metagenomic and Hi-C libraries were constructed. Red lines indicate gastrointestinal colonization with MDR enteric pathogens. **(B)** Each Hi-C read pair that maps to two non-mobile contigs is plotted according to the taxonomic assignment of each read. Color depicts the number of reads linking contigs according to taxonomy. **(C)** The percent of the total taxa-mobile gene (left) and taxa-AR gene (right) associations observed from metagenomic assembly that are supported by two or more Hi-C links (brown) is plotted, along with the percent additional interactions gained by using Hi-C (red). **(D)** Stacked barplots showing the number of species-level taxa to which each AR gene (clustered at 99% identity) is assigned within each patient, and across patients. Only those genes assigned to 2 or more taxa are shown. We either used metagenomic assemblies alone to assign taxonomies (left) or combined with Hi-C libraries considering those taxa-gene assignments with evidence from at least two Hi-C reads. The numbers above each stacked barplot represent the total number of AR genes with 2 or more taxonomic associations. **(E)** Horizontal stacked bar plots show the percentage of unique phage genes (defined as 95% similar) (above) or AR genes (below) in the metagenomic assemblies and the origins of their taxonomic associations, identified either by BLAST to NCBI's NT (for phage) or PATRIC's reference database (for AR genes) or through Hi-C linkages.

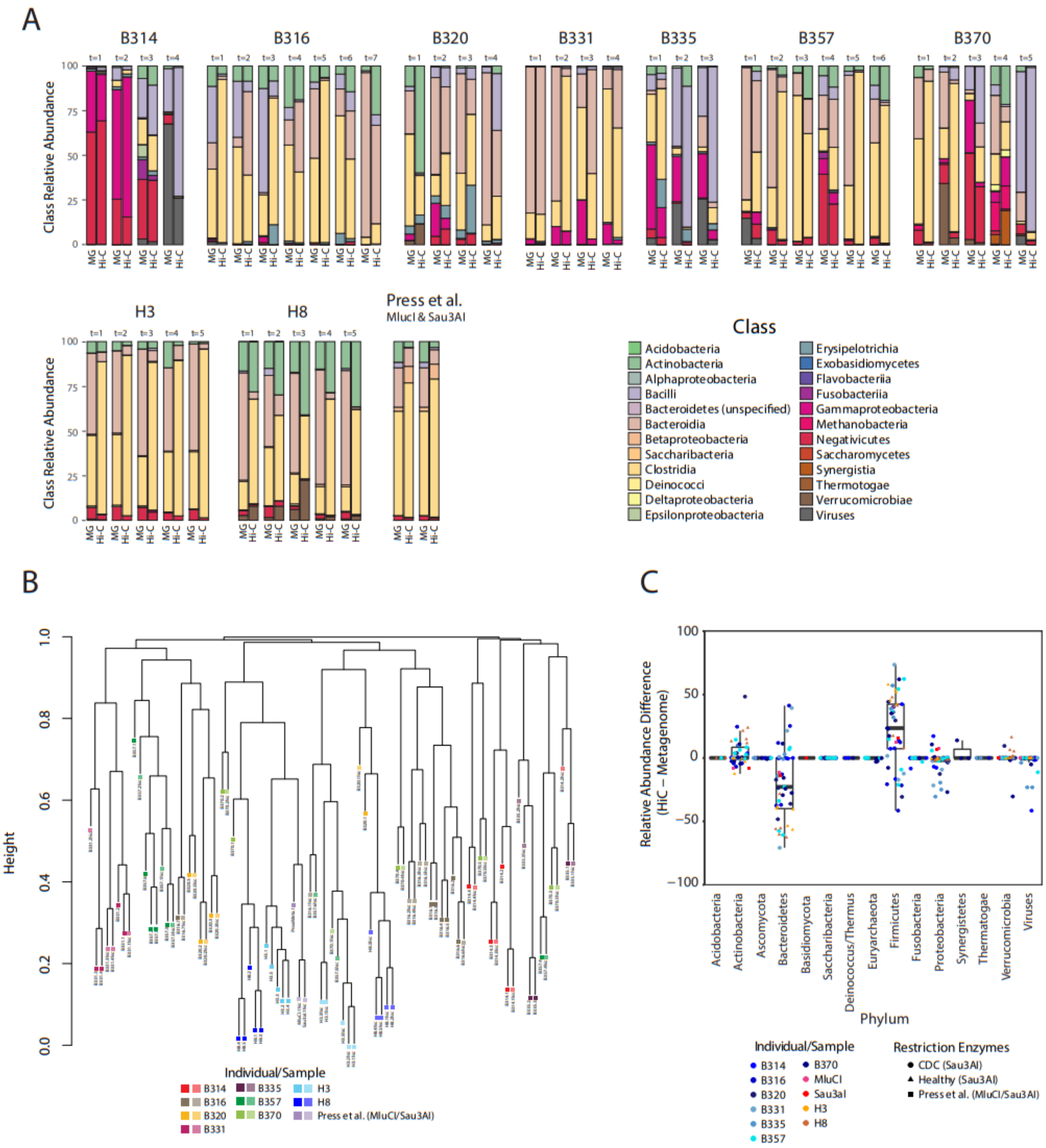


Figure 2.3. Congruence between metagenomic and Hi-C sample composition. **(A)** Class-level compositions of individuals' gut metagenomes and Hi-C libraries as determined by MetaPhlAn2. The human microbiome sample from Press *et al.* was included in our analysis. **(B)** Dendrogram of metagenome and Hi-C library compositions. Sample compositions (class-level) were hierarchically clustered according to their Bray-Curtis distances. **(C)** Compositional differences between metagenomic and Hi-C libraries in samples processed according to the restriction enzyme(s) used (numbers of comparisons are 4, 7, 4, 4, 3, 6, 5, and 5 for B314, B316, B320, B331, B335, B357, B370, H3, H8, respectively. In addition, data from two Hi-C samples were compared with 1 metagenome from Press *et al*.). The bounds of the box represent the first and third quartiles with the center value the median.

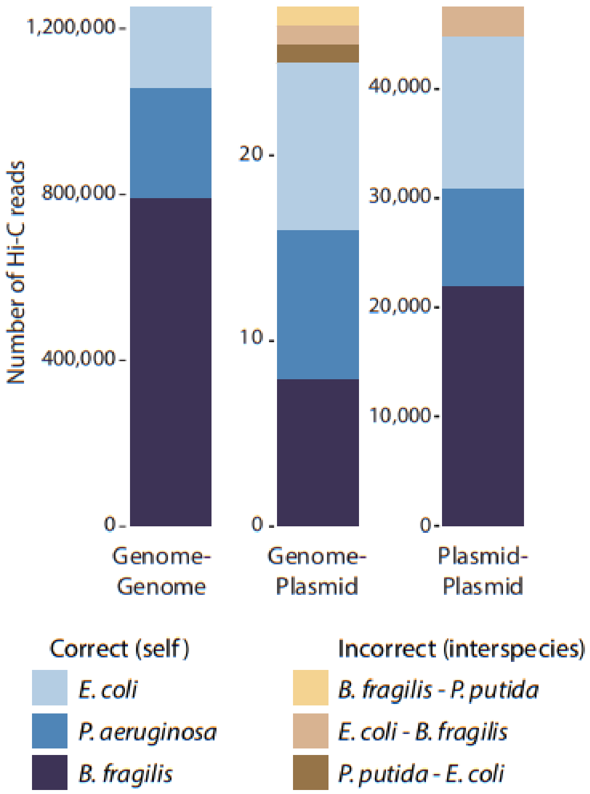


Figure 2.4. Number of Hi-C reads linking genomic regions to themselves (left), to their plasmids (middle) and within each plasmid (right). Blue linkages are correct, whereas brown hues are incorrect associations. Note that there is a region of homology between the plasmid backbones of the RP5 and pKJK5 plasmids carried by *Pseudomonas putida* and *Escherichia coli*, respectively. Nevertheless, none of the incorrect host-plasmid linkages would have been surpassed our threshold for assigning gene-taxa associations.

Results

Hi-C substantially improves antibiotic resistance gene-taxa associations

Our Hi-C experimental and computational approach results in robust linkages between contigs in human microbiome samples. Hi-C read pairs linking non-mobile contigs with contradictory taxonomic annotations are rarely observed (3.4% at the genus-level) and likely represent homologous sequence matches, highlighting the purity of our Hi-C libraries (Figure 2.2B). Hi-C read pairs linking two contigs are preferentially recruited to contigs that are longer and more abundant, but to a lesser degree than expected, reducing potential bias in our dataset toward highly abundant organisms (Figure 2.5). We binned contigs using several tools (Maxbin[18], MetaBat and Concoct), and applied a binning aggregation strategy, DAS Tool[19], to obtain a set of draft genomic assemblies. As misassembly can resemble HGT, we removed assemblies with greater than 10% contamination, as determined by CheckM, resulting in taxonomically coherent assemblies (Figure 2.6), albeit a greater number of unbinned contigs (24.6% of the total). We then apply conservative criteria to link mobile and mobile AR-containing contigs with the genomic draft assemblies, considering an MGE part of a genome assembly only if it is directly linked to it by at least two uniquely-mapped Hi-C read-pairs. As MGEs are known to recombine, this mitigates the potential for falsely linking contigs that merely share common mobile genes. However, this also potentially reduces our ability for overall detection, especially for larger MGEs, since mobile contigs are often fragmented in metagenomic assemblies[20]. Nevertheless, we restricted our analysis to those AR-organism and MGE-organism linkages derived from high-confidence read mappings.

Hi-C significantly improved our ability to detect mobile gene-bacterial host linkages beyond standard metagenomic assembly alone. Hi-C confirms many of the AR gene-taxa and mobile gene-taxa associations observed in the metagenomic assemblies (30.49% ± 11.49% of the AR genes; 30.1% ± 9.52% of the mobile genes), but importantly adds on average 31.81% ± 16.28% AR gene associations and 36.64% ± 11.56% mobile gene associations to those observed by metagenome assembly alone (Figure 2.2C). Furthermore, whereas metagenomic assembly methods can generally link a single mobile gene cluster to one or two organisms, our Hi-C method was able to identify up to 15 bacterial hosts harboring the same AR or mobile gene, requiring two or more Hi-C linkages within a single individual (mean = 3.53 ± 5.69 bacterial hosts per AR gene, Figure 2.2D; mean = 6.85 ± 10.88 bacterial hosts per gene, Figure 2.7). A larger percentage of AR and mobile genes overall (8.1%±5.2% vs. 0.9%±0.9% for AR genes and 6.1%±4.8% vs. 1.9%±1.7% for mobile genes) can be assigned to multiple taxonomies. These results were consistent with more stringent thresholds for Hi-C associations (Figure 2.8).

Our data increases mobile and AR gene-taxa assignments above those observed using publicly available reference genomes, while focusing on those immediately relevant to the individual patient. We first investigated phage-host associations identified through Hi-C and compared them with those in NCBI, as many phage are host-specific[21]. Indeed, 43.5% of the phage genes with Hi-C genera-level assignments recapitulate known interactions (Figure 2.2E). However, broader genera-level associations are obtained for 64.2% of the unique phage genes in our database, reflecting apparent selection biases within our reference databases and the promiscuity of certain phage[22,23,24]. A greater percentage of AR genes with Hi-C genera-level taxonomic assignments, 82.8%, were evident in reference genomes. Yet, Hi-C expands genera-level assignments for 37.6% of the AR genes. Despite having a limited number of reads linking each mobile or AR gene to a particular taxa, our annotations are supported by the fact that Hi-C reads preferentially map near to these genes on the overall contig (Figure 2.9).

We next sought to determine the extent to which we could capture associations using Hi-C. First, we performed a modified rarefaction analysis to determine whether the number of AR gene-taxa associations and mobile gene-taxa associations saturated with increased sequencing depth of our Hi-C libraries. Most of our samples saturated within our target sequencing depth (roughly 15 million paired reads), and sequencing samples to roughly four-fold this amount did not significantly increase the number of gene-taxa associations (Figure 2.10). The number of contigs that recruited Hi-C reads (on average 18.3 ± 10.9%) was not dependent on sequencing depth, yet 88.2% ± 9.5 of our genome bins recruited two or more Hi-C reads, which amounts to 90.7% ± 9.0% of the taxa recruiting reads. This breadth is supported by the congruence of Hi-C libraries and metagenomic libraries (Figure 2.3). We suspect that the variation in recruitment of Hi-C reads across the genome reflects either recurrent structural patterning of DNA[25], differences in DNA-binding proteins available for cross-linking, and the distribution of restriction enzyme cut sites[26]. We next measured our ability to detect the same mobile genes across timepoints. If we consider only the AR gene-taxa associations we observe at least once, and we conservatively assume that should continue to observe the gene-taxa association, *i.e.* that the lack of repeated observation was due to the stochastic sampling process of Hi-C rather than HGT, we repeatedly detect an average of 66% of all possible associations where both the organism and AR genes were detectable in the draft assemblies but were not linked through Hi-C (Figure 2.11).

Overall, within each person's microbiome, mobile genes, including AR genes and HGT machinery genes, were distributed across a wide range of taxa (Figures 2.12 & 2.13). Less than 10% of unique mobile genes and 19% of unique AR genes (clustered at 99% identity) were found across multiple patients, a finding consistent with previous surveys of MGEs across individuals[27], indicating limited inter-personal or nosocomial transmission. Furthermore, for these MGEs found across patients, few of their host associations were conserved. We speculate that HGT may result in their dispersal within individual's gut microbiomes and that selection may affect MGE-taxa associations at the level of individuals[27]. Despite heavy administration of antibiotics, the abundances of AR genes, even those conferring resistance to administered antibiotics, did not correspond with patient-specific therapeutic courses, a finding consistent with other patient-timecourses of mobile AR genes[28], and possibly reflective of the low plasmid-based resistance to levofloxacin or combination antibiotic therapies (Figure 2.14).

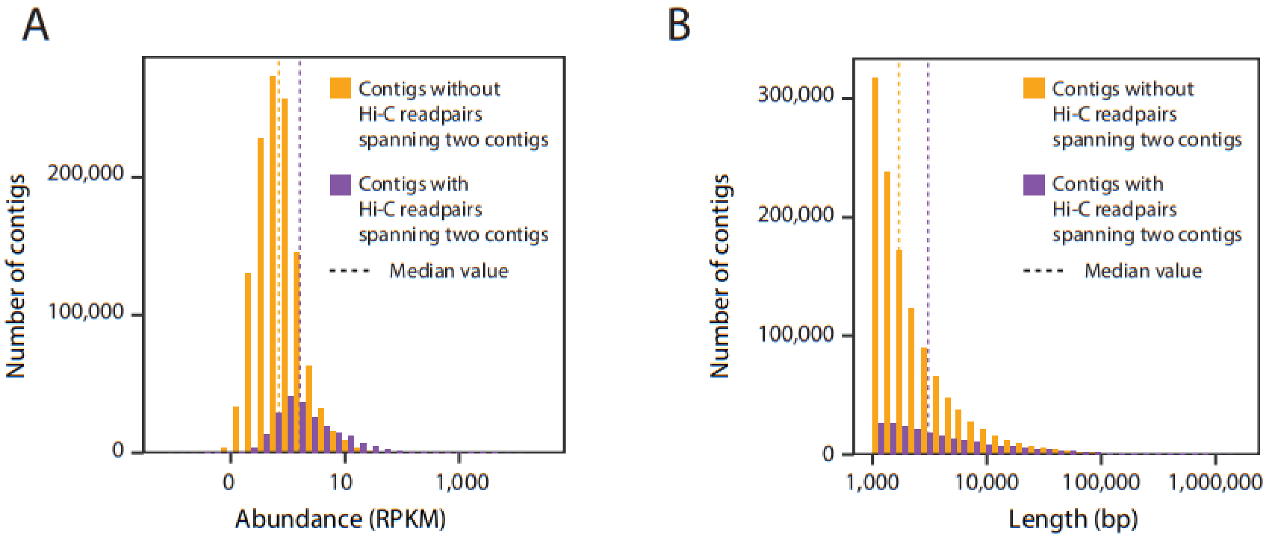


Figure 2.5. (A) A histogram showing the distribution of abundances (RPKM) of contigs that recruit (purple) or do not recruit (orange) Hi-C contig-connecting read pairs. **(B)** A histogram showing the distribution of lengths (bp) of contigs that recruit (purple) or do not recruit (orange) Hi-C contig-connecting read pairs.

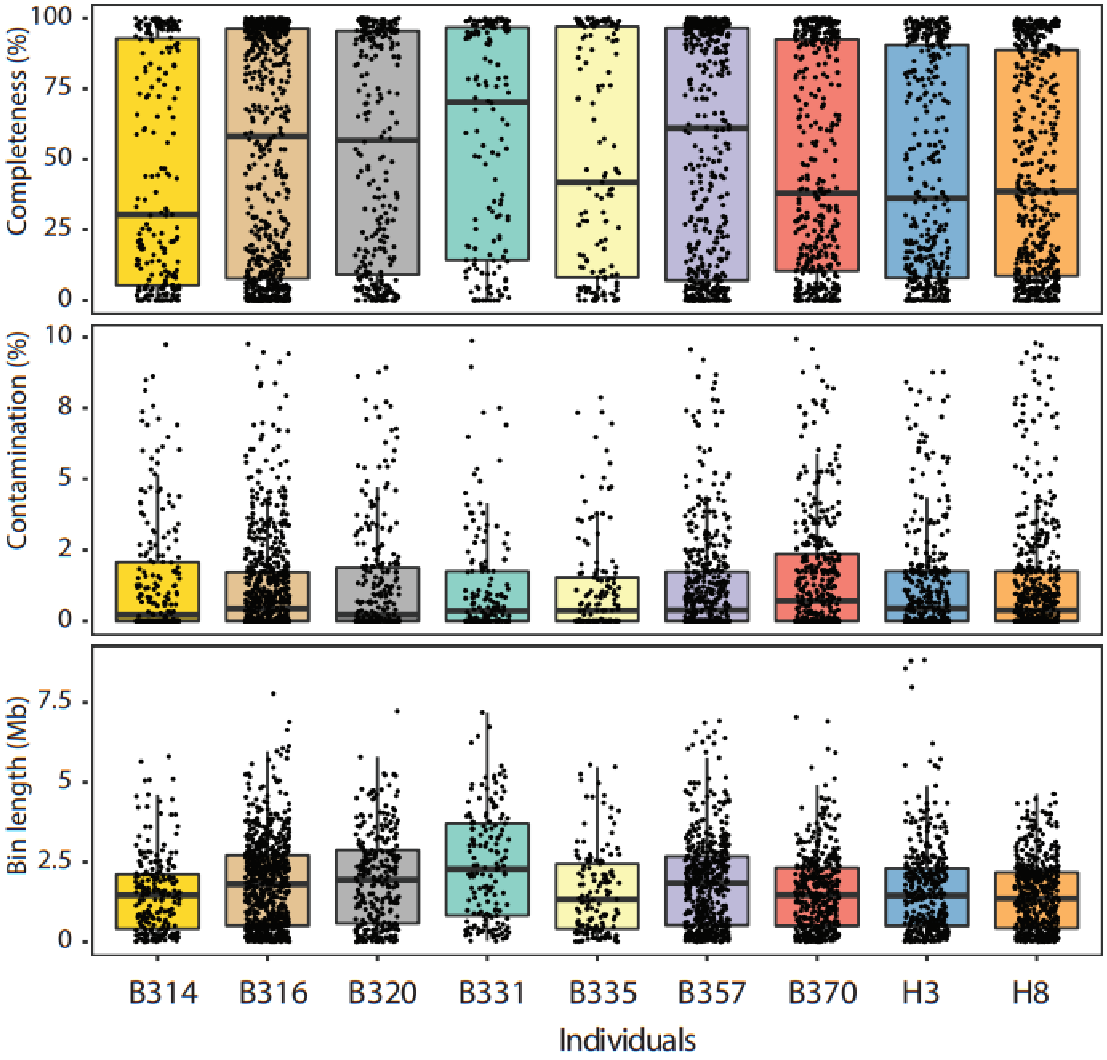


Figure 2.6. **(A)** Completeness of each assembled genome bins from each patients' samples as scored by CheckM. Boxplots show 25th, median and 75th percentile. (total number of genomes per person (n) = 261 (B314), 750 (B316), 297 (B320), 161 (B331), 156 (B335), 519 (B357), 469 (B370), 430 (H3), and 573 (H8)). **(B)** Contamination of each assembled genome bins from each patients' samples as scored by CheckM. Boxplots show 25th, median and 75th percentile. Same n as in (A). **(C)** Length of each assembled genome bins within each patients' samples. Boxplots show 25th, median and 75th percentile. Same n as in (A).

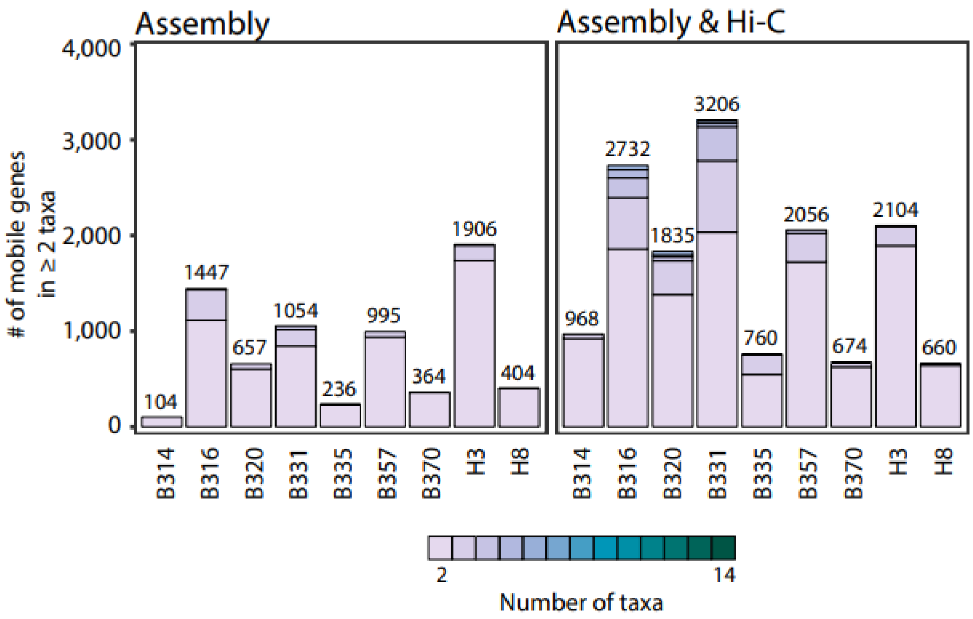


Figure 2.7. Stacked barplots showing the number of species-level taxa to which each mobile gene (clustered at 99% identity) is assigned within each patient, and across patients. Only those genes assigned to 2 or more taxa are shown. We either used metagenomic assemblies alone to assign taxonomies (left) or Hi-C libraries considering those taxa-gene assignments with evidence from at least two Hi-C read pairs. The numbers above each stacked barplot represent the total number of mobile genes with 2 or more taxonomic associations.

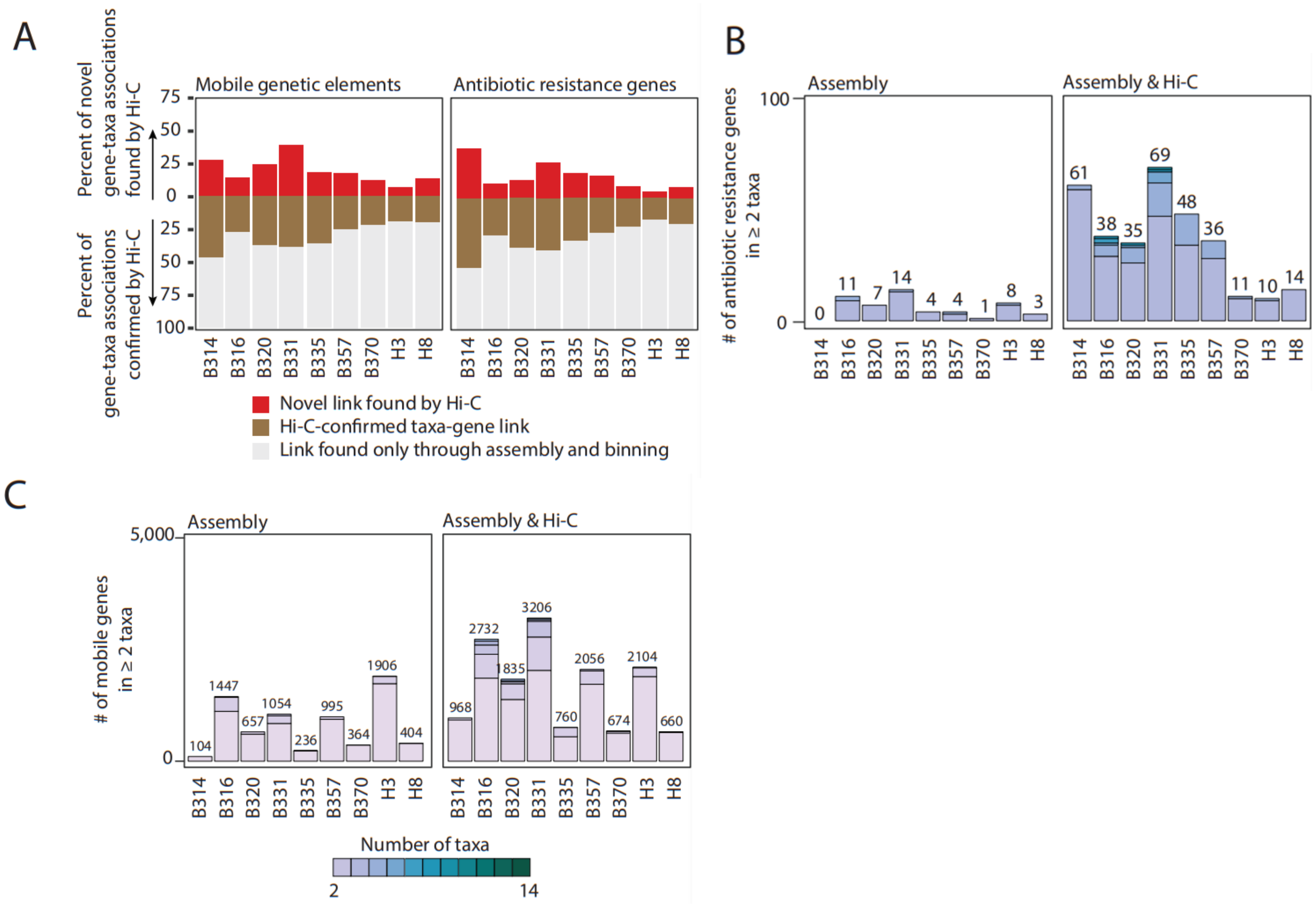


Figure 2.8. **(A)** The percent of the total taxa-mobile gene (left) and taxa-AR gene (right) associations observed from metagenomic assembly that are supported by five or more Hi-C links (brown) is plotted, along with the percent additional interactions gained by using Hi-C (red). **(B)** Stacked barplots showing the number of species-level taxa to which each AR gene (clustered at 99% identity) is assigned within each patient, and across patients as determined by 5 or more Hi-C links. Only those genes assigned to 2 or more taxa are shown. We either used metagenomic assemblies alone to assign taxonomies (left) or combined with Hi-C libraries considering those taxa-gene assignments with evidence from at least two Hi-C reads. The numbers above each stacked barplot represent the total number of AR genes with 2 or more taxonomic associations. **(C)** Stacked barplots showing the number of species-level taxa to which each mobile gene (clustered at 99% identity) is assigned within each patient, and across patients as determined by 5 or more Hi-C links. Only those genes assigned to 2 or more taxa are shown. We either used metagenomic assemblies alone to assign taxonomies (left) or combined with Hi-C libraries considering those taxa-gene assignments with evidence from at least two Hi-C reads. The numbers above each stacked barplot represent the total number of AR genes with 2 or more taxonomic associations.

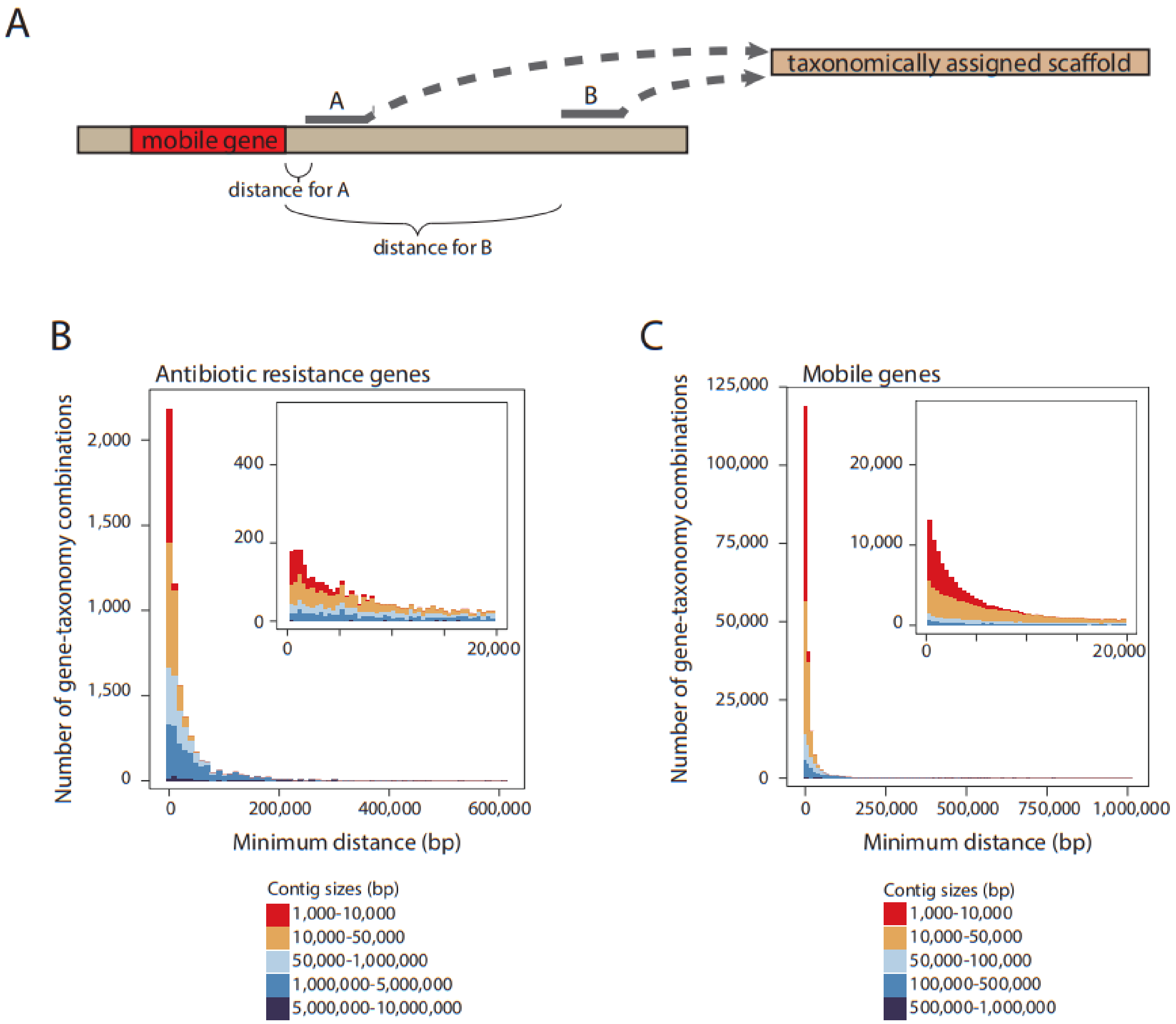


Figure 2.9. **(A)** As illustrated, Hi-C read pairs may map anywhere on a contig containing a mobile or AR gene. The boundaries of an MGE may be elusive and MGEs may integrate into contigs that have incomplete annotations. We assessed the linear distance (bp) between where Hi-C read pairs aligned and the positions of mobile or AR genes used for taxon-gene associations on the contigs to ensure that Hi-C read pairs were mapping at distances relevant for their assignments. In the example, both Hi-C reads A and B align to the same taxonomically annotated contig, yet read A maps at a minimum distance that is closer to the mobile gene, and therefore more confidently links the mobile gene with the contig. **(B)** For each taxon-mobile gene connection, we plot the minimum distance between a Hi-C read pair and the start/end of the mobile gene on that contig, according to contig length. The inset shows distances of less than 200,000bp broken down more finely. **(C)** The same analysis as (B) but for AR genes.

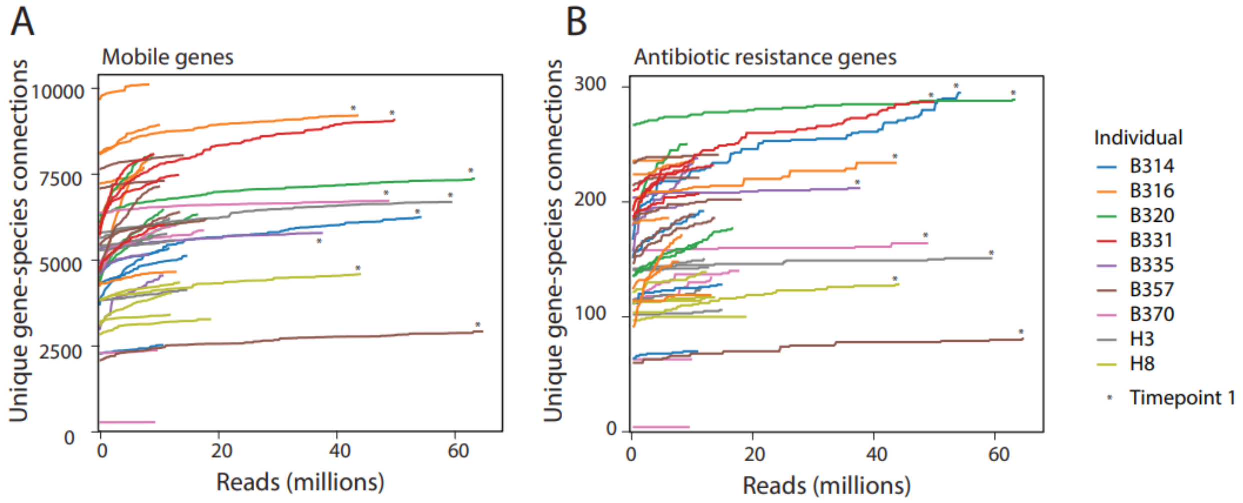


Figure 2.10. **(A)** The number of unique mobile gene-species connections within each sample after subsampling reads from each Hi-C dataset. The first Hi-C sample from each timecourse (noted with an asterisk) was sequenced significantly more deeply than the rest of the timecourse so that we could better assess whether there were any new gene-species connections that arose in any subsequent samples. **(B)** The number of unique AR gene-species connections within each sample after subsampling reads from each Hi-C dataset.

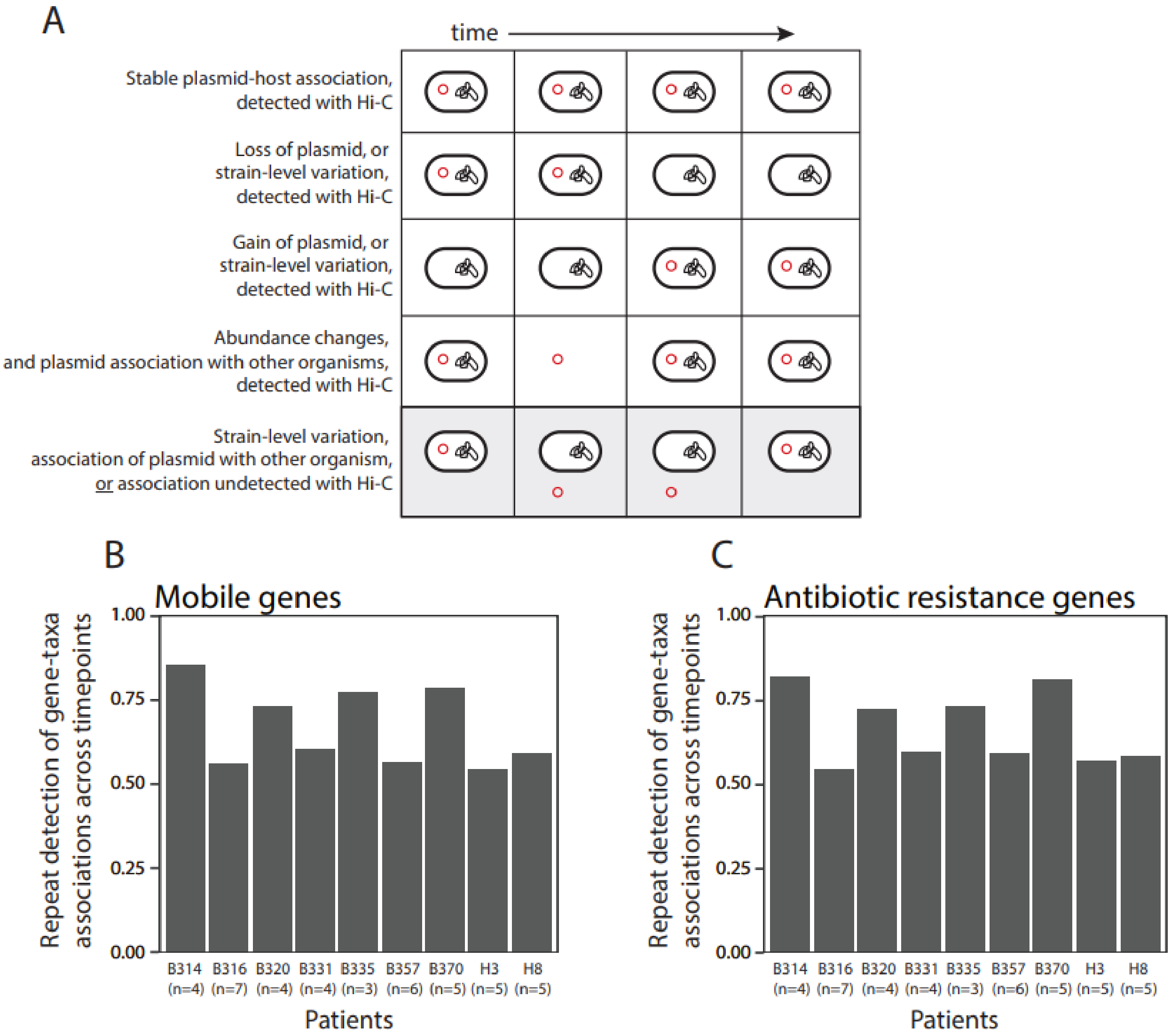


Figure 2.11. **(A)** Examples of the detection of AR genes, bacterial hosts and their linkage during patients' timecourses. The last scenario depicts an instance where a mobile or AR gene is linked with a specific taxon with Hi-C during at least one time point, but is detected in the metagenomic data at other time points but not linked with Hi-C. Although this may be explained by changes in strain-level composition or gene loss, we assessed the repeatability of detecting associations, assuming that these genes are truly linked in any instance when the mobile or AR gene and the bacterial taxon are both present. **(B)** A bar chart showing the extent to which we repeatedly detect specific gene-taxa associations observed within each patients' microbiomes. Assuming that we should observe associations present in one timepoint in all timepoints (i.e. that there is no HGT), we define the true positives (TPs) as the number of unique mobile gene-bacterial taxon connections observed; and the false negatives (FNs) as the total number of instances where both the bacterial taxon and the mobile gene are detected in the metagenomic assemblies. Repeat detection is calculated as TPs/(TPs+FNs), with the caveat that a portion of genome mobile gene-taxon linkages that did not depend on Hi-C sequencing read pairs are included here. **(C)** A bar chart showing the amount of repeat detection of AR gene-taxon connections observed within each patients' microbiomes. This was calculated as described in (B), with the same caveat applied.

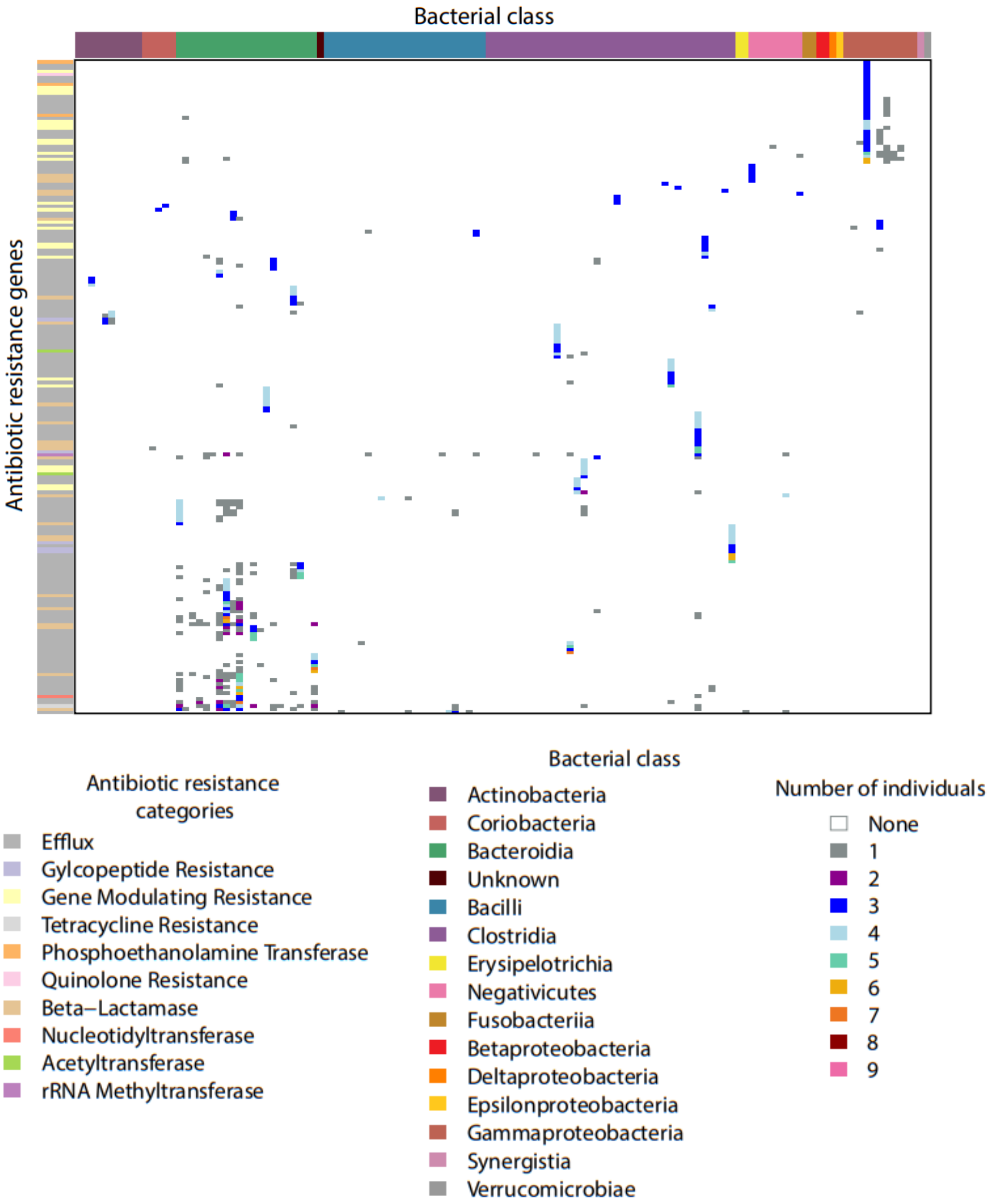


Figure 2.12. A heatmap of taxa-specific assignments, colored by class, for AR genes that are present in three or more patients.

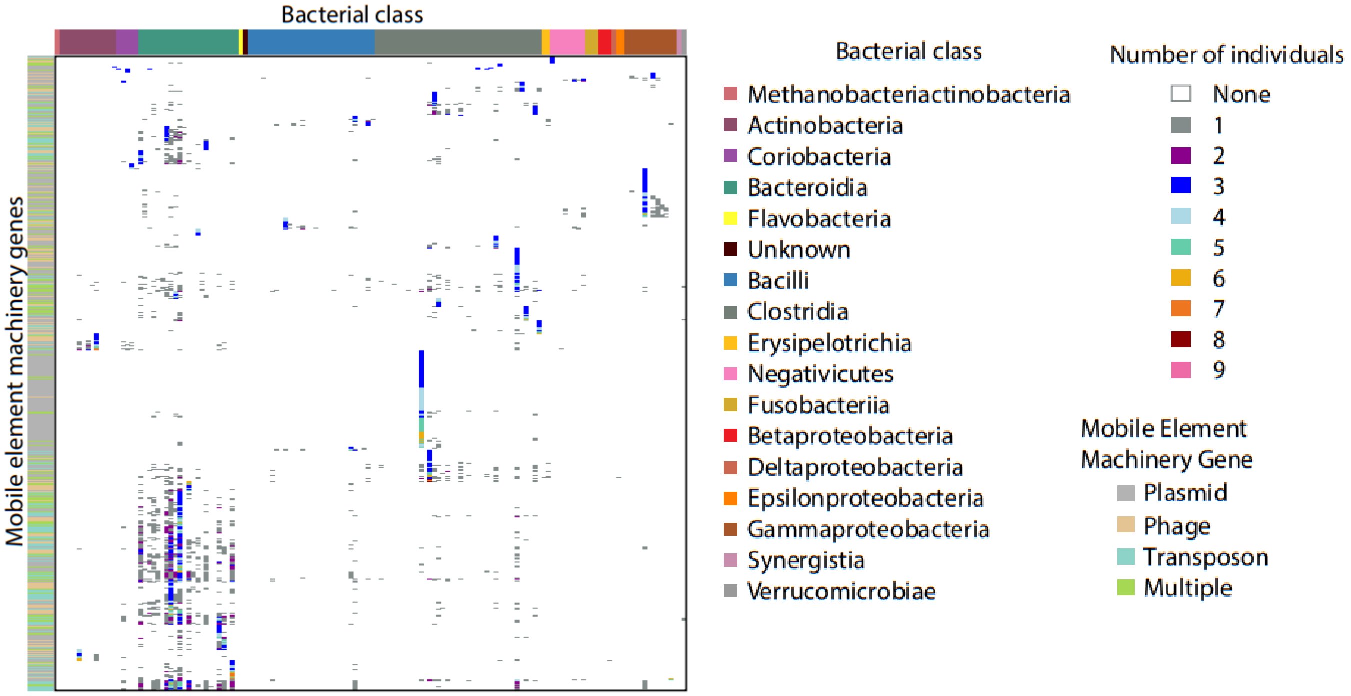


Figure 2.13. A heatmap of taxa-specific assignments, colored by class, for mobile genes that are present in three or more patients.

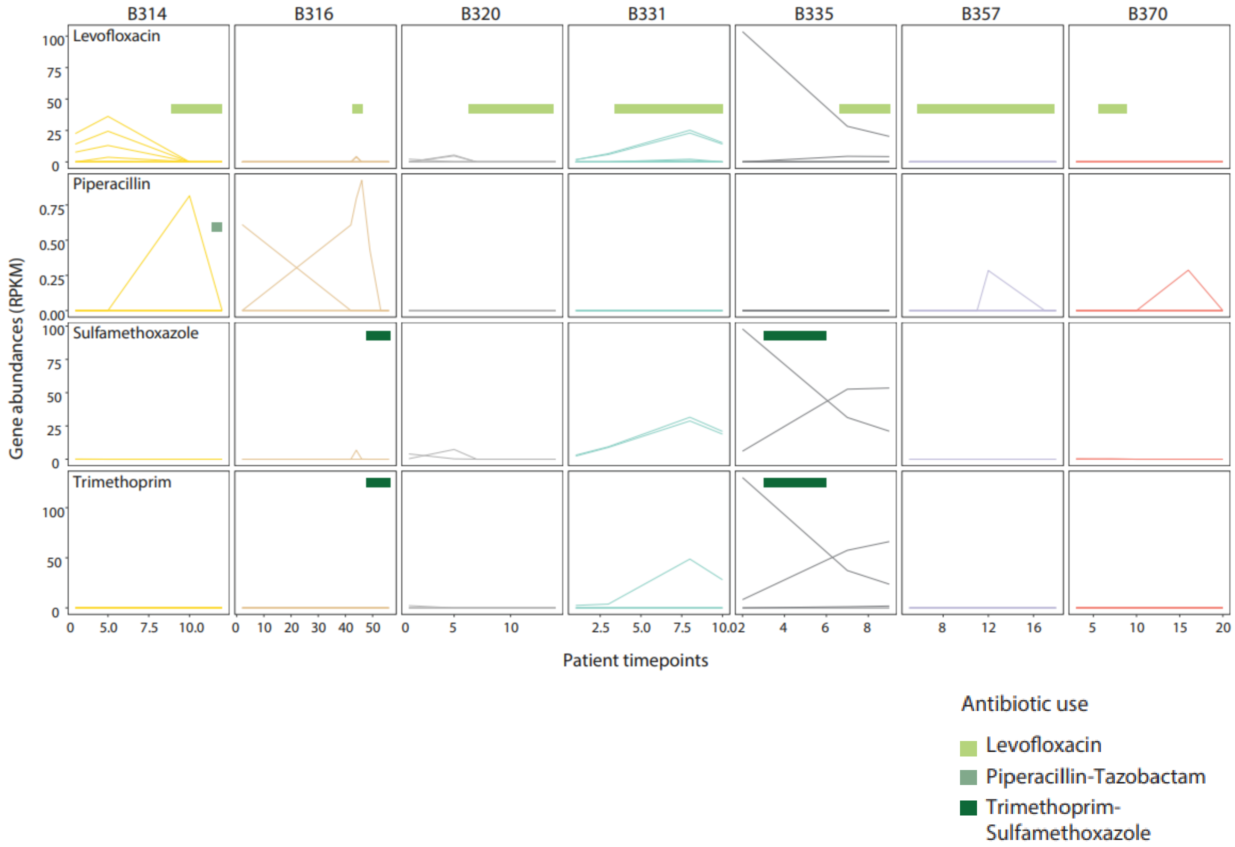


Figure 2.14. For each patient-timecourse (columns), AR gene abundances (RPKM) are plotted for each gene according to the antibiotic to which it confers resistant. The antibiotics administered to each patient over their timecourse is denoted.

Horizontal gene transfer networks vary across individuals' gut microbiomes

When comparing the networks of HGT within each individual's gut microbiome, we expected to observe a strong preference for gene exchange between more closely related organisms, as previously observed when comparing exchange networks using reference genomes[29]. Whether HGT occurs more frequently in an individual's gut is an essential question to understand the development and maintenance of the reservoir of AR genes in the gut microbiome, yet it has been difficult to answer for technical reasons. Using Hi-C, we find that the spread of AR genes and other mobile genes is significantly higher within an individual's gut microbiome than between different individuals' gut microbiomes (Figures 2.15A & 2.16). Beyond closely related pairs of organisms, there was considerable variation in the networks of shared AR and mobile genes across individuals (Figure 2.17). Despite this, we find that those microbiomes similar in composition shared more of the same connections among the organisms present in both microbiomes (Figure 2.15B), most notably between the two healthy individuals.

Given their clinical importance, we focused on the gene-sharing networks of Proteobacteria, and more specifically, Enterobacteriaceae. Within all patients, gene exchange was most frequent within members of the same phylum (Figure 2.15C,D). In neutropenic patients, Proteobacteria shared genes outside their phylum most often with Firmicutes. The main transfer partners with Enterobacteriaceae were different across patients, but notably included both opportunistic pathogens (*i.e. Veillonella parvula* and *Enterococcus faecium*), commensals that may flourish post-antibiotic use (*i.e. Erysipelotrichaceae* sp.[30]), and even those organisms that have been considered as probiotic (*i.e. Faecalibacterium prausnitzii*[31] and *Roseburia intestinalis*[32]).

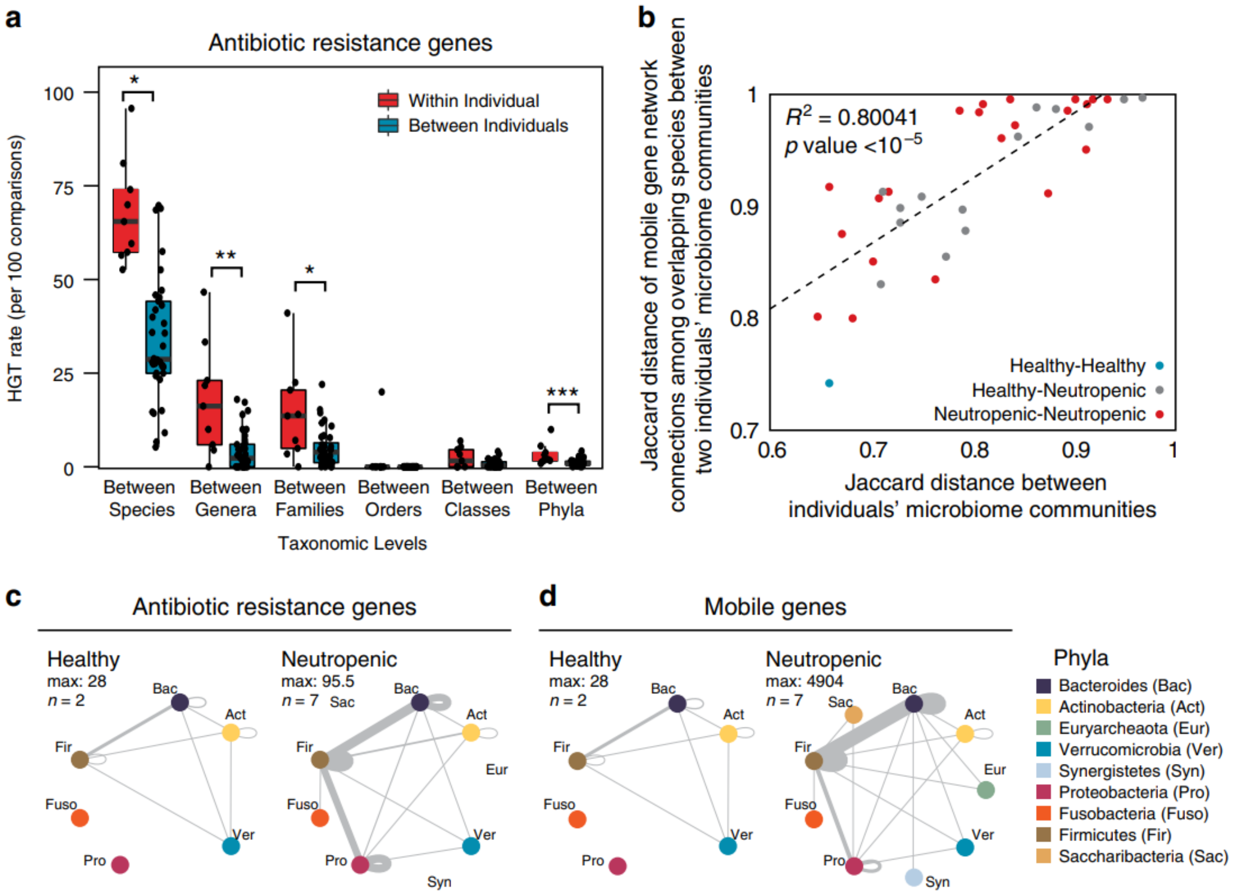


Figure 2.15. **(A)** HGT rates (per 100 comparisons) of AR genes between organisms within each individual (n=9) versus between individuals (n=36), according to those that share the same species, genus, family, order, class, and phylum are plotted for comparison. Significance was measured with Mann Whitney U-tests (two-sided; *, p<0.05; **, p<0.01; ***, p<0.005, ****, p<0.001;*****, p<0.0005. p-values are 0.0468, 0.0039, 0.0259, 0.0518,0.1929,0.0008, from species to phyla). Boxplot represents the interquartile range where ends of the whiskers represent ±1.5*interquartile range and median value is indicated. **(B)** For each pair of individuals, the Jaccard distance of their composite microbiome compositions are plotted against the average Jaccard distance of the HGT network connections of mobile genes exchanged between organisms present in both individuals. Points are colored according to the health status of the donors being compared. **(C)** Network plots showing bacterial AR gene exchange according to phyla within healthy (left) and neutropenic (right) individuals' microbiomes. n refers to the number of people included in the plot. **(D)** Network plots showing bacterial mobile gene exchange according to phyla within healthy (left) and neutropenic (right) individuals' microbiomes. n refers to the number of people included in the plot.

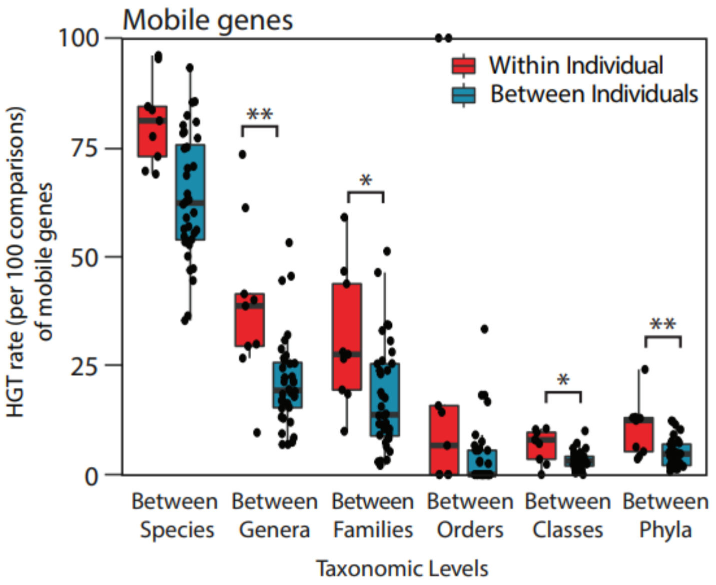


Figure 2.16. HGT rates (per 100 comparisons) of AR genes between organisms within each individual (n=9) versus between individuals (n=36), according to those that share the same genus, family, order, class, phylum and kingdom are plotted for comparison. Significance was measured with Mann Whitney U-tests (two-sided; *, p<0.05; **, p<0.01; ***, p<0.005, ****, p<0.001;*****, p<0.0005. p-values are 0.1186,0.00032, 0.01411, 0.1376, 0.0215, 0.0043 from species to phyla). The bounds of the box represent the first and third quartiles with the centre value the median. The ends of the whiskers represent either the smallest and largest values or at most ±1.5× interquartile range.

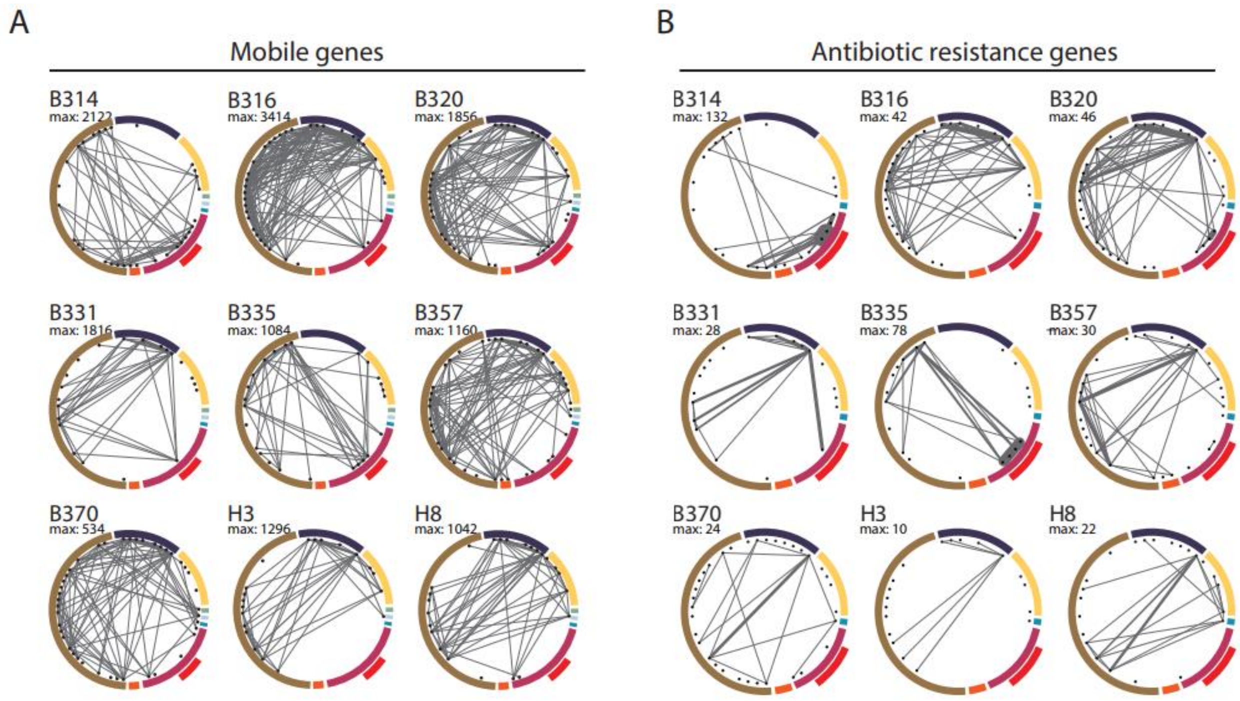


Figure 2.17. **(A)** Circle relationship plots showing networks of bacterial mobile gene exchange in the gut microbiomes across individuals (top left) and within each individual. Taxa present in each individual's microbiome are depicted by black circles. The thickness of the lines corresponds to the number of unique mobile genes associating the two taxa. **(B)** Circle relationship plots showing networks of bacterial AR gene exchange in the gut microbiomes across individuals (top left) and within each individual. Taxa present in each individual's microbiome are depicted by black circles. The thickness of the lines corresponds to the number of unique AR genes associating the two taxa.

Horizontal gene transfer is frequent and is elevated in neutropenic patients

Antibiotic treatment[33] and inflammation[34] are putative triggers for HGT, through the production of reactive oxygen species and DNA damage. We hypothesized that mucositis caused by cytotoxic chemotherapy, along with the selective pressures imposed by antibiotics and inflammation, would create conditions amenable to HGT in these neutropenic patients. We noticed that the average density of connections (percentage of actual connections of the total possible connections) between taxa and AR or mobile genes is greater in the neutropenic patients than the healthy individuals (Figure 2.18A). Several patients, B316, B320, B335 and B370, experienced increases in the proportion of overall gene-taxa connections, referred to as network density, during their timecourses. This was unrelated to the abundance of Enterobacteriaceae in the samples, which have been proposed as mediators of HGT[35], the total abundance of AR genes, or the number of Hi-C reads (Figure 2.19). Rather, we found that the only correlate was the number of taxa in a sample: as patients' microbiomes became less diverse, the gene-taxa network density increased (Figure 2.18B). We hypothesize that this is caused either by an undefined selective pressure acting to preserve more connected organisms; or that once selection has occurred, organisms in less diverse populations will have increased contact rates and therefore greater opportunity for transfer.

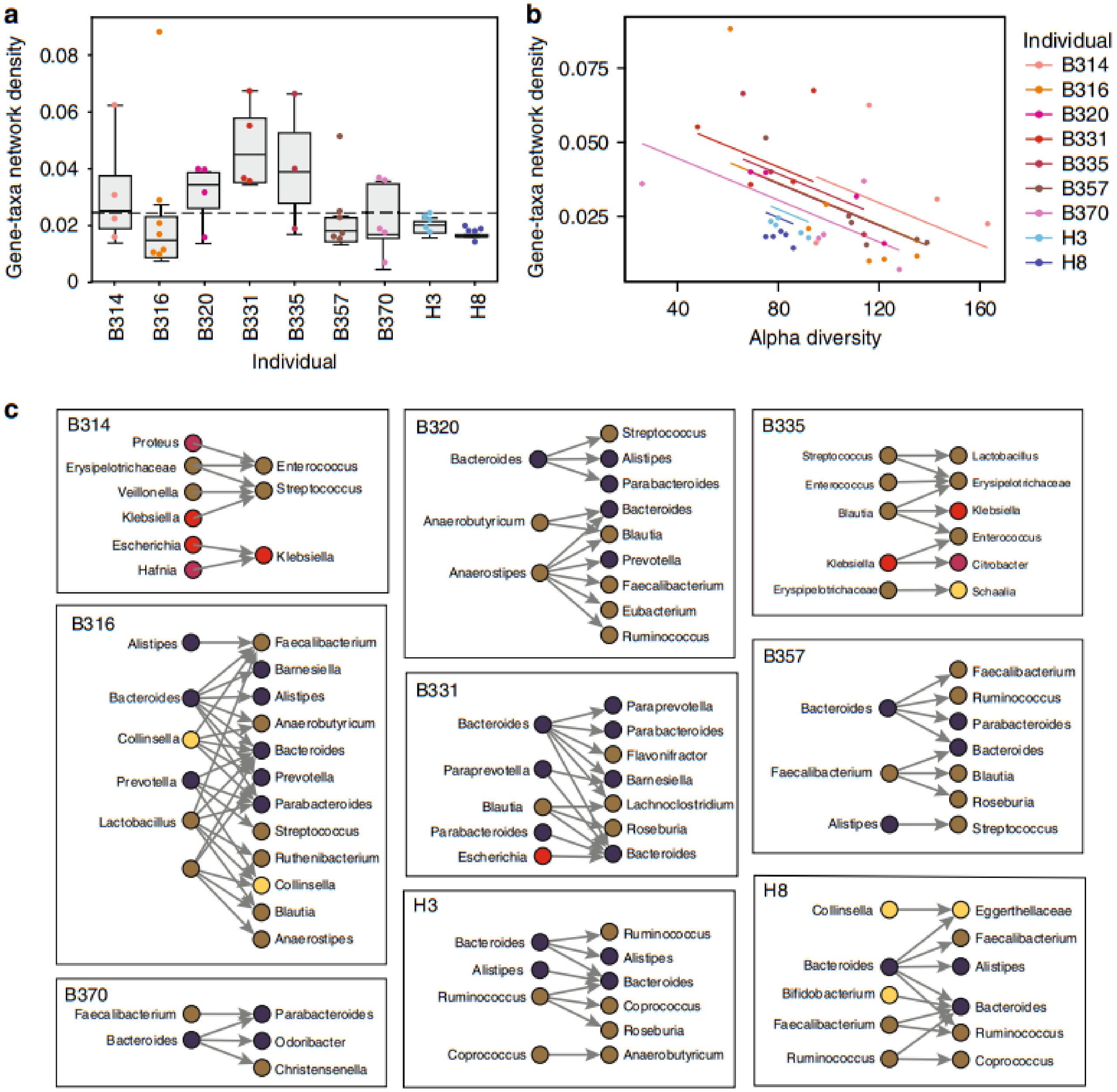


Figure 2.18. **(A)** Boxplots showing the gene-taxa network linkage densities, or the proportion of total possible gene-taxa links that are observed to be linked, for each individuals' samples (n=4, 7, 4, 4, 3, 6, 5, and 5 for B314, B316, B320, B331, B335, B357, B370, H3, H8, respectively). A dotted line is shown at the maximum network density observed in the healthy samples. The bounds of the box represent the first and third quartiles with the centre value the median. The ends of the whiskers represent either the smallest and largest values or at most ±1.5× interquartile range. **(B)** Individual patient samples are plotted according to the alpha diversity, assessed using Metaphlan, and their gene-taxa network density. An ANOVA showed that gene-taxa network density was related to alpha diversity (F(1,39), p = 3.3x10^{-5}) and health status (F(1,6), p=0.01501). **(C)** All observed HGT events across different genera are plotted for each individual. Each genus is colored according to its phylum.

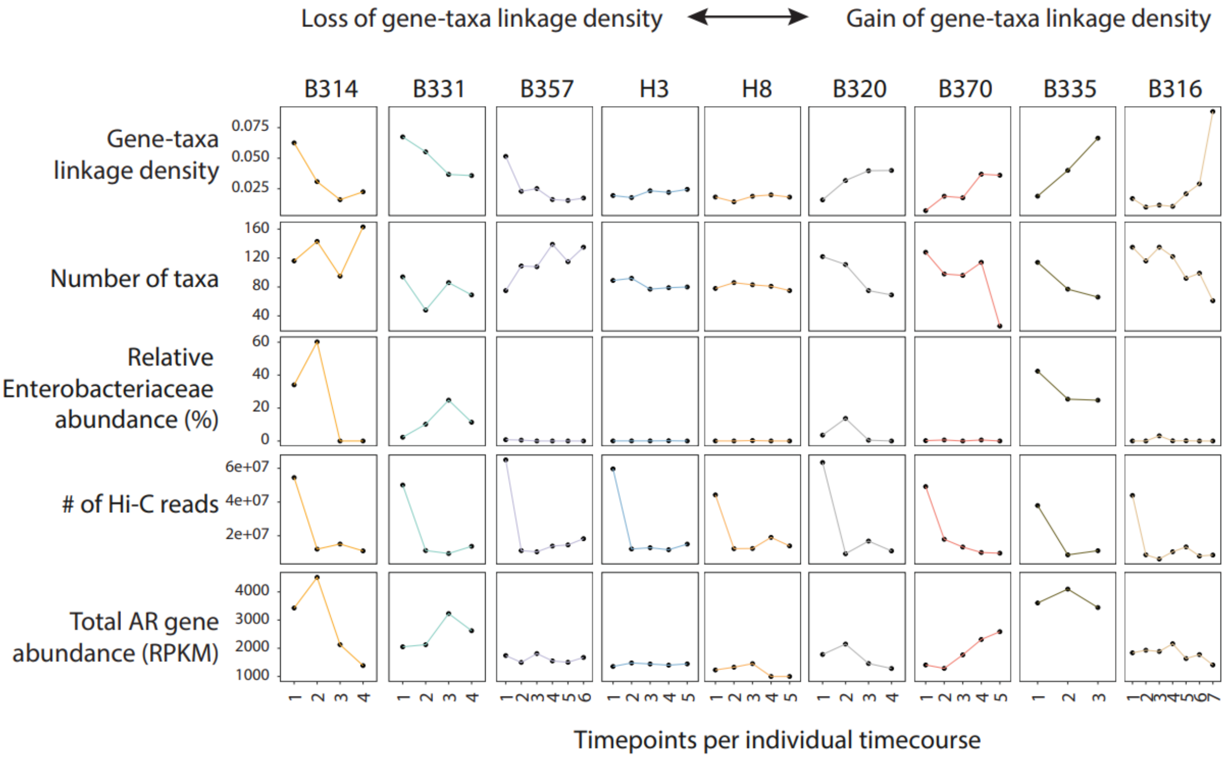


Figure 2.19. Each patient's timecourse is shown according to their gene-taxa linkage density, as defined by both Hi-C and metagenomic assembly; the number of taxa in that sample's metagenome (calculated by Metaphlan), the number of total Hi-C reads; and the abundance of AR genes (RPKM). Individuals were ordered according to the trend of their gene-taxa linkage density over their timecourse.

Emergence of antibiotic resistance in pathogens and commensals

Next, we more closely examined those timecourses with putative HGT events for which we had the highest confidence. To distinguish between the migration of new bacterial strains and HGT, we only considered HGT between strains present at the start of the timecourse. Potential donor strains were required to have Hi-C-verified connections with specific mobile or AR genes in the first patient sample. Individuals' initial Hi-C samples were sequenced 3-4-fold deeper than the remainder of their timecourses to ensure that gene-taxa associations were adequately sampled (Figure 2.10) and that putative recipient strains did not harbor those specific genes of interest at the start. We enforced this by requiring a complete absence of gene-recipient taxa connections inferred by Hi-C or metagenomic assembly, including connections with taxa that could only be annotated at higher taxonomic levels. Finally, we considered HGT as occurring between these donor strains and recipient strains with Hi-C-verified associations with the transferred genes in later timepoints. Providing additional support, 12.2% of the putative transfer events (19 of 155) were supported by Hi-C across multiple timepoints and 32.9% (51 out of 155) were supported by Hi-C links across multiple contigs in both the donor and recipient genomes. Most of the transfers (60%) were between members of the same phylum. Ultimately, evidence of HGT was found in all individuals in our study (Figure 2.18C).

Within these relatively short timecourses, we observed the expansion of the gut commensal reservoir of resistance genes. Although we did not observed the transfer of AR genes conferring resistance specifically to the antibiotics used in this cohort, namely levofloxacin, pipercillin-tazobactam, or trimethoprim-sulfamethoxazole, we did observe transfer of multi-drug resistance cassettes with beta-lactam- and fluoroquinolone-resistance genes, covering two of the corresponding antibiotic classes. Notably, within a few days post-transplant, we see transfer of a plasmid encoding mdtEF, a multidrug efflux pump conferring resistance to fluoroquinolones, and their transcriptional regulators, CRP and gadW, from an *Escherichia coli* strain in patient B331 to a strain most similar to *Bacteroides sp. A1C1*. Despite their ubiquitous antibiotic prophylaxis, only a minority (19.4%) of transfer events involved annotated AR genes in the neutropenic patients.

Additionally, we observe the emergence of novel AR genes in enteric pathogens, originating either from gut commensals or other enteric pathogens, including Enterobacteriaceae. Enterobacteriaceae species are among the most common causes of infection and sepsis in these patients and Enterobacteriaceae from the gut have been shown previously to harbor excessive numbers of AR genes[36] and serve to promote HGT of AR genes[37]. We see the exchange of AR gene-containing plasmids between members of the Enterobacteriaceae, namely *Klebsiella pneumoniae* and *Citrobacter brakii* in patient B335, and between *E. coli* and *Klebsiella* species in B314, and one instance of *K. pneumoniae* in patient B335 acquiring DNA harboring a plasmid-based efflux pump from a commensal, *Blautia hansenii*. We also note the overall transfer of mobile elements between these pathogenic species and other opportunistic pathogens, such as *Streptoccocus parasanguinis, S. salivarius,* and *E. faecium*, exposing the potential for HGT to alter the AR profiles of these bacteria over short periods of time.

Remaining challenges linking bacteria with their mobile genetic elements

These examples highlight the dynamic nature of HGT within the gut ecosystem, especially in the context of gut inflammation, immune dysregulation and antibiotic use. Nevertheless, our method has several limitations. First, we can only assign bacterial hosts for those MGEs and host genomes that we are able to assemble and annotate. Although 95.9% ± 2.8% of our metagenomic reads contribute to assembled contigs and 80.4% ± 10.4% (median 82.4%) of Hi-C reads align to our assemblies, we were only able to annotate 47.8% of our draft assemblies at either the genus- or species-level. Second, the assembly of MGEs can be confounded by their high rates of recombination creating multiple genomic arrangements and transfer, and, thus, redundancy within and across genomes. To mitigate the potential for false positive interactions, we examined only those mobile gene-containing contigs with multiple Hi-C reads directly linking them to taxonomically annotated genome assemblies. We cannot however rule out the possibility that our sensitivity is actually higher, and that our inability to detect linkages at specific time-points reflects true strain-level variation within the microbiome, or undetected real-time mobilization of genetic elements. Third, for those HGT events that we observed, we cannot always confirm the transfer of an entire contig and its associated genes. This issue underscores several observed HGT events, involving plasmids comprising prophage and transposable elements. This is mitigated by the requirement for more Hi-C read linkages and the overall proximity between Hi-C read linkages and the inferred transferred genes (Figure 2.9). Future studies should leverage long reads in hybrid assemblers to better capture co-occurring AR genes and large MGEs[38]. We expect to overcome these limitations with additional technical improvements to the bacterial Hi-C protocol.

Discussion

Here, we observe extensive transfer of mobile and AR genes within a single individual's gut microbiome across distant phylogenetic backgrounds and over relatively short timespans. The transfer networks within each individual's gut microbiome are unique and are likely explained by personal ecological niches that govern local contact rates between organisms. Few of the total AR gene-taxon associations are observed across individuals, which may suggest limited dispersal rates and/or strong selective pressures that prevail within each individual's gut. Although the molecular dynamics of HGT in the gut microbiome are not well-understood, our data from healthy subjects point to a basal level of transmission, even in the absence of inflammation or antibiotic use. Many of the observed transfers appeared transient, which may be due to the limited detection of our method, or by the neutral or deleterious nature of most HGT events[39,40].

The ramifications of HGT in this neutropenic patient population are acute. Our results show increased pathogen load and elevated gene-taxa network densities in neutropenic patients as compared with healthy individuals, suggesting an increased risk of emergence of MDR pathogens in this at-risk patient population. How to translate these findings into the prevention of the emergence of MDR pathogens is paramount. This technology highlights the potential for screening the burden of AR genes and the carriage of enteric pathogens to guide empirical antibiotic therapy. These findings also expose the limitations of taxa-specific therapies to remove AR genes from the gut microbiome[41,42,43], in favor of mechanisms to limit HGT more generally. Overall, these results emphasize a view of the population-wide dissemination of AR genes that includes diverse members of the gut microbiome**.**

Methods

Sample collection

Fresh stool was collected from informed and consenting individuals in accordance with IRB protocols for Weill Cornell Medical College (#1504016114) and Cornell University (#1609006586). Neutropenic patients were all admitted to the Bone Marrow Transplantation Unit at New York Presbyterian Hospital/Weill Cornell Medicine between December 2016 and July 2017. Healthy samples were collected similarly in 2019. Approximately 0.25 g replicates of each time-point were either frozen 'as is' (for metagenomic sequencing) or homogenized in phosphate-buffered saline (PBS) + 20% glycerol before freezing (used for Hi-C sequencing).

Metagenomic sequencing

Frozen stool was thawed on ice and DNA was extracted using the PowerSoil DNA Isolation Kit (Qiagen) with additional Proteinase K treatment and freeze/thaw cycles recommended by the manufacturer for difficult-to-lyse cells. Extractions were further purified using 1.8 volumes of Agencourt AMPure XP bead solution (Beckman Coulter). DNA was diluted to 0.2 ng/uL in nuclease-free water and processed for sequencing using the Nextera XT DNA Library Prep Kit (Illumina).

Proximity ligation

Stool stored in PBS + 20% glycerol was thawed on ice for 15 minutes and homogenized in 5 mL PBS containing 4% v/v formaldehyde. Sample were crosslinked at room temperature with continuous inversion for 30 minutes, then incubated on ice for 30 minutes. Unreacted formaldehyde was quenched by adding glycine to a final concentration of 0.15 M and incubating for 10 minutes on ice. Crosslinked cell mixtures were pelleted (10,000 g, 4° C, 5 min.), the supernatant was removed, and pellets were flash-frozen on dry ice/ethanol and stored at -80° C.

Frozen crosslinked stool cell pellets were thawed on ice then resuspended in 450 µL TES (10 mM Tris, 1 mM EDTA, 100 mM NaCl, pH 7.5) and transferred to 2 mL screw-cap tubes. 50 µL freshly prepared Lysozyme solution (20 mg/mL in TES, Amresco lyophilized powder, 23500 U/mg) was added to each resuspended pellet and incubated at room temperature for 15 minutes with continuous inversion. Sodium dodecyl sulfate (SDS) was added to a final concentration of 0.5% w/v and samples were incubated at room temperature for 10 minutes with continuous inversion. Samples were pelleted and the volume was reduced to 400 µL. 50 µL 10X Lysis Buffer (100 mM Tris pH 7.5, 100 mM NaCl, 1% IGEPAL CA-630 v/v) was added to each sample, followed by 50 µL freshly prepared 10X protease inhibitor (Roche cOmplete mini

EDTA-free tablets). Cells were resuspended by pipetting and incubated on ice for 15 minutes. Manual lysis of cells was carried out by adding 400 μL 0.5 mm sterile glass beads to each tube and vortexing at maximum Hz for 30 seconds, followed by 30 seconds incubation on ice. Vortexing and ice incubation was repeated for 10 cycles. Bead-beaten samples were allowed to settle upright on ice for 15 minutes, then the liquid supernatant (~250 μL) was transferred to a new 1.5 mL tube. Sample volume was equilibrated to 500 μL with cold 2X NEBuffer 1.1 and incubated at 50° C for 10 minutes. After incubation, 30 μL 10% Triton X-100 v/v was added to each tube, mixed by inversion. Cross-linked DNA fragments were digested overnight with 50 U Sau3AI. Digested DNA complexes were pelleted (20,000 g, 4° C, 5 min.) , gently washed with cold 1X NEBuffer 2, and resuspended in 200 μL NEBuffer 2.

Digested DNA was heated to 50° C for 5 minutes to melt paired sticky ends then put into a 200 μL Klenow fragment (exo-, NEB) fill-in reaction containing 36 μM biotin-14-dCTP (Thermo Fisher) and equimolar amounts of dATP, dTTP, and dGTP. Reactions were carried out for 2 hours at room temperature and the polymerase was quenched by adding EDTA to a final concentration of 10 mM. The full volume of each fill-in reaction was put into a dilute blunt-end ligation reaction (640 U T4 DNA Ligase, NEB) and allowed to incubate overnight at 15° C. Protein and crosslink digestion was carried out by adding 50 μL freshly prepared 20 mg/mL Proteinase K (VWR, freeze-dried powder suspended in 10 mM Tris, 1 mM $MgCl_2$, 50% glycerol, pH 7.5) and incubating at 65° C for 6 hours. This digestion was repeated once. Protein was removed by phenol:chloroform extraction and ligated DNA was precipitated from the aqueous fraction with one volume 5M ammonium acetate and 4 volumes cold absolute ethanol. Clean DNA was quantified, and at least 1 μg but no more than 5 μg DNA was put into an end-resection reaction (5 U T4 DNA Polymerase, NEB) to remove biotin from unligated ends.

Exonuclease activity of the polymerase was quenched with 5 mM EDTA and free biotinylated nucleotides were removed via 1.8X Ampure XP bead cleanup. Biotinylated DNA was immobilized on M280 streptavidin beads using the Invitrogen kilobaseBINDER Kit. Bead-bound DNA was quantified and prepared for sequencing using Illumina's Nextera XT kit. Multiplexed libraries were size-selected with Ampure XP beads, quantified, and pooled for sequencing on an Illumina NextSeq 2x150 paired-end platform.

Mock community methods

Bacillus subtilis containing pDR244, *Pseudomonas putida* containing pKJK5, and *Escherichia coli* containing RP4 were cultured in LB under antibiotic selection to maintain plasmids (spectinomycin, tetracycline, and kanamycin, respectively). Overnight cultures were washed with PBS, resuspended in PBS + 20% glycerol v/v, and frozen as aliquots, with one aliquot of each retained for titer determination on selective agar media. To create the mock community, 5×10^8 colony forming units from each frozen stock was thawed and combined, and immediately carried through formaldehyde crosslinking as described for stool. Mock community Hi-C sequences were mapped with HiC-Pro against reference genomes and plasmids using default settings. Valid pairs, *i.e*. those that map to different restriction fragments, were compartmentalized into groups based on whether or not they connected the genome-genome, genome-plasmid, or plasmid-plasmid and coded according to the expected plasmid-host relationship.

Quality filtering and assembly

Metagenomic and Hi-C sequences were quality filtered using Prinseq[45] v0.20.2 to derepelicate, Bmtagger[46] (v2/21/14) to remove human reads, and Trimmomatic[47] v0.36 to remove adapters and quality filter reads (using settings: Leading:3, Trailing:3, Slidingwindow

4:15, Minlen: 50). Metagenomic reads were assembled using SPAdes[48] v3.13.2 with '-meta' setting with a minimum contig size of 1,000bp. Genes on these contigs were called using Prodigal[49] v2.6.3. PhageFinder[50] v2.1 was used to identify large prophage regions and were excised from the first to the last phage gene called and considered separate contigs, unless the surrounding regions were less than 1,000bp, in which case they were also included as the excised phage.

Metagenomic binning

Contigs were binned using several tools (Maxbin[51], MetaBat[52], and Concoct[53]), culminating with a metagenomic binning aggregation strategy, DAS Tool[54], we assessed genome contamination using CheckM and removed bins with contamination >10%, resulting in quality metagenomic bins although in many cases partial bins. To prevent overcalling of partial bins, downstream analyses aggregate at the taxon rather than individually calling unique bins.

Taxonomic Identification

Kraken was applied to each metagenomic bin and annotated each contig individually using its algorithm. Then we assigned each bin the lowest taxonomic level at which more then 50% of the bin was assigned by Kraken with contigs weighted by length (bp). Contigs assigned Eukaryotic taxonomies were removed from further analysis.

Antibiotic resistance gene annotation

All contigs were annotated with CARD's (Comprehensive Antibiotic Resistance Database) Resistance Gene Identifier (RGI)[55] 3.2.1 against the CARD[56] database and with HMMer[57] against the Resfams[58] database with the gathering cutoff. AR genes were clustered using CD-HIT-EST[59] (identity:0.99; word size:8; length difference cutoff: 0.9) after they were sorted by length. Antibiotic resistance mechanisms are defined in Supplementary Data 5. We focused on AR genes that are commonly harbored by the most problematic MDR bacteria[60] and that confer resistance to antibiotics that are most frequently relied on in neutropenic patients:

Table 1. AR genes of high importance

Drug	**Resistance determinant**
Oxacillin[61]	*mecA, mecC*
Penicillin[62,63,64]	*pbp2b, pbp2x*
Ampicillin[65]	***bla***$_{\mathbf{TEM}}$, ***bla***$_{\mathbf{SHV}}$
Cephalosporin[66,67]	***bla***$_{\mathbf{TEM}}$, ***bla***$_{\mathbf{SHV}}$***,*** ***bla***$_{\mathbf{CTX\text{-}M}}$, ***bla***$_{\mathbf{CMY}}$***,*** ***bla***$_{\mathbf{MIR}}$, *bla*$_{\mathrm{MOX}}$, *bla*$_{\mathrm{LAT}}$, *bla*$_{\mathrm{FOX}}$, ***bla***$_{\mathbf{DHA}}$***,*** ***bla***$_{\mathbf{ACT}}$, *bla*$_{\mathrm{CFE}}$
Carbapenem[68]	*bla*$_{\mathrm{KPC}}$, ***bla***$_{\mathbf{NDM}}$, *bla*$_{\mathrm{VIM}}$, *bla*$_{\mathrm{IMP}}$, *bla*$_{\mathrm{OXA\text{-}48}}$, *bla*$_{\mathrm{OXA\text{-}23}}$, *oprD*
Fluoroquinolones[69]	***gyrA, gyrB, parC, parE***, *qnrA*, *qnrS* (Note: we considered any ***qnr*** gene.)
Aminoglycosides[70]	***aac(3’), aac(6’), aad***

Genes in bold above represent those genes that were identified in our cohort’s microbiomes.

Mobile genetic element annotation

All contigs and excised prophage contigs were assessed for the presence of mobile genes using several programs. Contigs were mapped using BLASTN to PlasmidFinder[71] database (best hit, minimum 80% identity and 60% coverage), NCBI’s genomic plasmids downloaded (05/10/2017) (best hit, minimum 1000bp, minimum 80% identity), and IMMEdb[72] (best hit, min 1,000 bp and 80% identity). Contigs were also identified as plasmids using PlasFlow[73] with threshold of 0.95. Genes were mapped using BLASTP to ACLAME[74] database v0.4 (besthit, min 80% identity and 60% coverage) and PHASTER[75] prophage/virus database (v8/3/17) (best hit, min 80% identity and 60% coverage). Genes were mapped using HMMER[76] v3.1b2 to Pfam[77] and known plasmid, phage, and transposons were identified[78,79]. A search of common mobile gene terms against Pfam descriptions was carried out. Terms included for transposon: transpos, insertion element, is element, IS[0-9]; phage: phage, tail protein, tegument, capsid, relaxase, tail fibre, tail assembly, tail sheath, tail tube; plasmid: conjug, Trb, type IV, Tra[A-Z], mob, Vir[A-Z][0-9], t4ss, resolvase, plasmid; other: integrase. All Pfam IDs and descriptions are listed in the Supplementary Data 6. Contigs were also annotated for insertion sequence (IS) elements using

ISEScan[80] v1.5.4. Contigs with taxonomies assigned to the ‘Virus’ domain were considered phage. Contigs with any mobile annotation were annotated as MGEs.

Sequence mapping

Paired-end metagenomic sequences were mapped to the metagenomic contigs using BWA-MEM[81] v0.7.13 requiring primary only alignments and filtered at 90% identity. Paired-end metagenomic sequences were also mapped separately to the AR and mobile gene clusters and filtered at 99% identity. Contig and individual AR and mobile gene RPKM values were calculated using mapped metagenomic reads (total reads mapped to the contigs with >80% contig coverage, divided by the length of the contigs per kilobase and the total read count in that sample per million). Hi-C reads were mapped with HiC-Pro using default parameterswhich internally uses Bowtie2. HicPro requires valid pairs to map to different restriction fragments and allows only unique mapping of reads.

Cleanliness comparisons

Reads mapping between two different contigs were included in the analysis if neither contig carried a mobile gene. Taxonomic associations were determined from residency in a metagenomic bin annotated to at least the taxonomic level of interest.

Mobile and antibiotic resistance gene associations

All mobile and AR gene-containing contigs, including excised phage, were associated with taxa if they were linked to a Hi-C clustered genomic contig with at least two Hi-C read pairs or if they were clustered into an annotated genomic bin. Hi-C linkages between MGEs and their genomic bins are more robust if Hi-C reads map more closely (*i.e.* smaller linear distance (bp)) with the genes that are annotated as mobile. To assess this, we calculated the genetic distance between the mobile or AR gene and the nearest Hi-C read linking any contig with a particular taxonomy. Often this resulted in multiple linkages between the mobile contig and taxonomic contigs clustered with the same taxa. We therefore assessed the strongest data linking the two, the minimum genetic distance, considering the other reads as further support for this gene-taxon assignment.

Comparison of horizontal gene transfer between individual taxa

First, we compared HGT observed between species (as shown in Figures 2.9 and 2.15B,C) defined above, through Hi-C read pair linkages. To create an HGT network, we examined the number of unique (defined as 99% sequence identity) AR or mobile gene linked to genomic bins for each particular taxa. Consequently we could identify taxa-taxa connections based on these identified gene sharing events.

We assessed the rate of HGT per 100 species-species comparisons at different taxonomic levels within and between patients, as a comparison with Smillie *et al.* (2011)[82]. For comparisons between species, we compared each species within a single genus to one another. For every other taxonomic level, we compared species that differed according to that taxonomic level (*i.e*. for comparisons between families, species of one family, *e.g.* the Enterobacteriaceae, were compared exclusively with species in other taxonomic families). When comparing two species, we considered HGT events as those taxa sharing at least one gene of interest (AR or mobile gene) at >99% identity. We compared HGT within each patient or performed pairwise comparisons between the 9 individuals. For each taxonomic level, we compared within vs. between patients using a Mann-Whitney U-test.

Antibiotic resistance gene and phage machinery gene host specificity

Genes of interest associated with taxa through Hi-C alone (*i.e*. not including taxa originally assigned to a contig that contained that AR gene or phage machinery gene) were compared to taxonomies identified by comparison using BLASTN (e-value < 1e-100) to PATRIC[83] genome database (downloaded October 1, 2018) for the AR genes or compared to NCBI nt database (downloaded June 4, 2018) for the phage machinery genes and placed into one of several categories defined in Figure 2.2. This was assessed at different taxonomic levels.

Network density

Network density was calculated by dividing the number of observed connections between mobile or AR genes and binned organisms in each sample out of the theoretical maximum number of connections (number of AR genes or mobile genes multiplied by number of distinct organisms). Number of total species in a population were identified from MetaPhlan.

Measuring novel horizontal gene transfer during individuals' timecourses

Within an individuals' timecourse, we identified novel HGT events by comparing the first timepoint to subsequent timepoints and requiring that the gene-taxa connection met several criteria. HGT events were only considered between donor organisms strongly linked with a mobile or mobile AR gene at the start of the timecourse and recipient organisms present, albeit unlinked to the mobile or mobile AR gene, at the start of the timecourse. We sequenced the initial timepoint 3-4 times more deeply than the remainder of the timecourse to be able to distinguish between migration of strains and HGT. Mobile or AR gene-containing contigs were required to be linked via at least 2 Hi-C reads (mean = 28.6) to genome assemblies that were taxonomically annotated at the level of genus or species. We required an absence of association between the mobile or AR gene of interest and potential recipient taxa. In other words, one Hi-C read was sufficient to disqualify a putative HGT event, as was any taxonomic marker on that contig associating it with a congruent recipient taxon, or any association with a genome assembly with any congruent higher-order taxonomy. We required recipient taxa to have at least

2 Hi-C reads (mean = 14.8) associating each mobile or AR gene-containing contig with the recipient genome assembly. We tallied the number of HGT events that were supported by more than one timepoint, both strictly by the same genes and also by the same taxa; as well as those that were supported across multiple contigs.

Data and code availability

Metagenomic and Hi-C sequences, filtered for quality and human-reads are available on NCBI's Short Read Archive (PRJNA649316). Our mock metagenomic sample is at SAMN15663484. Code relies heavily on published packages and several databases, including CARD's (Comprehensive Antibiotic Resistance Database) Resistance Gene Identifier (RGI), Resfams, IMMEdb, PlasmidFinder, ACLAME, Pfam, PHASTER, the PATRIC genome database and the NCBI genome plasmid database. Code for this project is available on Github at https://github.com/acvill/microbiome.

References

1. Huddleston, J. R. Horizontal gene transfer in the human gastrointestinal tract: potential spread of antibiotic resistance genes. *Infect Drug Resist*. **7**, 167-176. (2014).
2. Satlin, M. J. & Walsh, T. J. Multidrug-resistant Enterobacteriaceae, Pseudomonas aeruginosa, and vancomycin-resistant Enterococcus: Three major threats to hematopoietic stem cell transplant recipients. *Transpl Infect Dis* **19,** (2017).
3. Sommer, M. O. A., Dantas, G., Church, G. M. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science*. **325**, 1128-1131 (2009).
4. Gibson, M. K., Forsberg, K. J. & Dantas, G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J* **9,** 207–216 (2015).
5. Kaminski, J. *et al.* High-Specificity Targeted Functional Profiling in Microbial Communities with ShortBRED. *PLoS Comput. Biol.* **11,** e1004557 (2015).
6. Yaffe, E. & Relman, D. A. Tracking microbial evolution in the human gut using Hi-C reveals extensive horizontal gene transfer, persistence and adaptation. Nat Microbiol (2019) doi:10.1038/s41564-019-0625-0.
7. Marbouty, M. *et al.* Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. *Elife* **3,** e03318 (2014).
8. Burton, J. N., Liachko, I., Dunham, M. J. & Shendure, J. Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. *G3 (Bethesda)* **4,** 1339–1346 (2014).
9. Beitel, C. W. *et al.* Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ* **2,** e415 (2014).

10. Marbouty, M. *et al.* Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. *Elife* **3,** e03318 (2014).
11. Stewart, R. D., *et al.* Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nature Comm* **9**, 870 (2018).
12. Bickhart, D. M. *et al.* Assignment of virus and antimicrobial resistance genes to microbial hosts in a complex microbial community by combined long-read assembly and proximity ligation. *Genome Biology* **20**, 153 (2019).
13. Stalder, T., Press, M. O., Sullivan, S., Liachko, I. & Top, E. M. Linking the resistome and plasmidome to the microbiome. *The ISME Journal* **13**, 2437–2446 (2019).
14. Marbouty, M., Baudry, L., Cournac, A. & Koszul, R. Scaffolding bacterial genomes and probing host-virus interactions in gut microbiome by proximity ligation (chromosome capture) assay. *Sci Adv* **3,** e1602105 (2017).
15. Satlin, M. J. & Walsh, T. J. Multidrug-resistant Enterobacteriaceae, Pseudomonas aeruginosa, and vancomycin-resistant Enterococcus: Three major threats to hematopoietic stem cell transplant recipients. *Transpl Infect Dis* **19,** (2017)
16. Pop, M. Genome assembly reborn: recent computational challenges. *Brief Bioinform*. **10**, 354-366 (2009).
17. Krawczyk, P. S., Lipinski, L., Dziembowski, A. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res*. **46**, e35 (2018). Apr 6; 46(6): e35.
18. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32,** 605–607 (2016).
19. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol.* 3, 836-843. (2018)
20. Pop, M. Genome assembly reborn: recent computational challenges. *Brief Bioinform*. **10**, 354-366 (2009).
21. Ross, A., Ward, S. & Hyman, P. More Is Better: Selecting for Broad Host Range Bacteriophages. *Front. Microbiol.* **7,** (2016).
22. Yu, J., Lim, J.-A., Kwak, S.-J., Park, J.-H. & Chang, H.-J. Comparative genomic analysis of novel bacteriophages infecting Vibrio parahaemolyticus isolated from western and southern coastal areas of Korea. *Arch. Virol.* **163**, 1337–1343 (2018).
23. Doulatov, S. *et al.* Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements. *Nature* **431**, 476–481 (2004).
24. Ross, A., Ward, S. & Hyman, P. More Is Better: Selecting for Broad Host Range Bacteriophages. *Front Microbiol* **7**, 1352 (2016).
25. Crémazy, F. G. *et al.* Determination of the 3D genome organization of bacteria using Hi-C. *Methods Mol Biol*. **1837**, 3-18 (2018).
26. Marbouty, M. *et al.* Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. *Elife* **3**, e03318 (2014).
27. Brito, I. L. *et al.* Mobile genes in the human microbiome are structured from global to individual scales. *Nature* **535,** 435–439 (2016).
28. Yassour, M. *et al.* Natural history of the infant gut microbiome and impact of antibiotic treatments on strain-level diversity and stability. *Sci Transl Med* **8**, 343ra81 (2016).
29. Smillie, C. S. *et al.* Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480,** 241–244 (2011).
30. Kaakoush, N. O. Insights into the Role of Erysipelotrichaceae in the Human Host. *Front Cell Infect Microbiol*. **5**, 84 (2015).
31. Rossi, O. *et al.* Faecalibacterium prausnitzii A2-165 has a high capacity to induce IL-10 in human and murine dendritic cells and modulates T cell responses. *Scientific Reports* **6**, 18507 (2016).
32. Zhu, C. *et al.* Roseburia intestinalis inhibits interleukin-17 excretion and promotes regulatory T cells differentiation in colitis. *Mol Med Rep*. **17**, 7567–7574 (2018).

33. Modi, S. R., Lee, H. H., Spina, C. S. & Collins, J. J. Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature* **499**, 219–222 (2013).
34. Diard, M. *et al.* Inflammation boosts bacteriophage transfer between Salmonella spp. *Science* **355**, 1211–1215 (2017).
35. Bakkeren, E. *et al.* Salmonella persisters promote the spread of antibiotic resistance plasmids in the gut. *Nature* **573**, 276–280 (2019).
36. Sommer, M. O. A., Dantas, G. & Church, G. M. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science* **325,** 1128–1131 (2009).
37. Bakkeren, E. *et al.* Salmonella persisters promote the spread of antibiotic resistance plasmids in the gut. *Nature* **573**, 276–280 (2019).
38. Bertrand, D. *et al.* Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat. Biotechnol.* **37**, 937–944 (2019).
39. Knöppel, A., Lind, P. A., Lustig, U., Näsvall, J. & Andersson, D. I. Minor fitness costs in an experimental model of horizontal gene transfer in bacteria. *Mol. Biol. Evol.* **31**, 1220–1227 (2014).
40. McCarthy, A. J. *et al.* Extensive horizontal gene transfer during Staphylococcus aureus co-colonization in vivo. *Genome Biol Evol* **6**, 2697–2708 (2014).
41. Citorik, R. J., Mimee, M., & Lu, T. K. Sequence-specific antimicrobials using efficiently delivered RNA-guided nucleases. *Nature Biotechnology*. **32**, 1141-1145 (2014).
42. Vercoe, R. B. *et al.* Cytotoxic chromosomal targeting by CRISPR/Cas systems can reshape bacterial genomes and expel or remodel pathogenicity islands. *PLoS Genet*. **9**, e1003454 (2013).
43. Yosef, I., Manor, M., Kiro, R., & Qimron U. Temperate and lytic bacteriophages programmed to sensitize and kill antibiotic-resistant bacteria. *Proc Natl Acad Sci U S A*. 112, 7267-72 (2015).
44. Press, M. O. *et al.* Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions. Preprint at bioXriv. https://doi.org/10.1101/198713 (2017)
45. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863–864 (2011).
46. Rotmistrovsky, K. & Agarwala, R. BMTagger: Best Match Tagger for removing human reads from metagenomics datasets. Unpublished (2011).
47. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120 (2014).
48. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res*. 27, 824–834 (2017).
49. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11, 119 (2010).
50. Fouts, D. E. Phage Finder: Automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res* 34, 5839–5851 (2006).
51. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32, 605–607 (2016).
52. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 7, e7359 (2019).
53. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat Methods* 11, 1144–1146 (2014).
54. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol*. 3, 836-843. (2018)
55. Jia, B. *et al.* CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res* 45, D566–D573 (2017).
56. McArthur, A. G. *et al.* The comprehensive antibiotic resistance database. *Antimicrob. Agents Chemother*. 57, 3348–3357 (2013).

57. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res*. 41, e121 (2013).
58. Gibson, M. K., Forsberg, K. J. & Dantas, G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J* 9, 207–216 (2015).
59. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659 (2006).
60. Centers for Disease Control. Antibiotic Resistance Threats in the United States. Atlanta, GA (2013)
61. Lakhundi, S. & Zhang, K. Methicillin-Resistant Staphylococcus aureus: Molecular Characterization, Evolution, and Epidemiology. *Clin Microbiol Rev*. 31, e00020-18 (2018).
62. Magill, S. S. *et al*. Prevalence of antimicrobial use in US acute care hospitals, May-September 2011. *JAMA* 312, 1438-1446 (2014).
63. Dowson, C. G. *et al*. Penicillin-resistant viridans streptococci have obtained altered penicillin binding protein genes from penicillin-resistant strains of Streptococcus pneumoniae. *Proc Natl Acad Sci U S A*. 87, 5858-5862 (1990).
64. van der Linden, M. *et al*. Insight into the Diversity of Penicillin-Binding Protein 2x Alleles and Mutations in Viridans Streptococci. Antimicrob Agents Chemother. 61, e02646-16 (2017).
65. Paterson, D. L. & Bonomo, R. A. Extended-Spectrum β-Lactamases: a Clinical Update. *Clinical Microbiology Reviews.* 18, 657-686. (2005).
66. Paterson, D. L. & Bonomo, R. A. Extended-spectrum beta-lactamases: a clinical update. *Clin Microbiol Rev*. 18, 657-686. (2005).
67. Strahilevitz, J., Jacoby, G. A., Hooper, D. C., Robicsek, A. Plasmid-mediated quinolone resistance: a multifaceted threat. *Clin Microbiol Rev*. 22, 664-689 (2009).
68. Queenan, A. M., & Bush, K. Carbapenemases: the Versatile β-Lactamases. *Clinical Microbiology Reviews*. 20, 440-458 (2007).
69. Hooper D., C., & Jacoby, G., A.Topoisomerase Inhibitors: Fluoroquinolone Mechanisms of Action and Resistance. *Cold Spring Harb Perspect Med*. 6, a025320 (2016).
70. Doi, Y., Wachino, J., I., Arakawa, Y. Aminoglycoside Resistance: The Emergence of Acquired 16S Ribosomal RNA Methyltransferases. *Infect Dis Clin North Am*. 30, 523-537 (2016).
71. Carattoli, A. *et al*. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother*. 58, 3895–3903 (2014).
72. Jiang, X., Hall, A. B., Xavier, R. J., & Alm, E. J. Comprehensive analysis of mobile genetic elements in the gut microbiome reveals phylum-level niche-adaptive gene pools. *PLoS One*. Doi: 10.1371/journal.pone.0223680 (2019)
73. Krawczyk, P. S., Lipinski, L. & Dziembowski, A. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res*. 46, e35 (2018).
74. Leplae, R., Lima-Mendez, G. & Toussaint, A. ACLAME: a CLAssification of Mobile genetic Elements, update 2010. *Nucleic Acids Res*. 38, D57-61 (2010).
75. Arndt, D. *et al*. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res*. 44, W16-21 (2016).
76. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 41, e121 (2013).
77. Finn, R. D. *et al*. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res*. 44, D279-285 (2016).
78. Sczyrba, A. *et al*. Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nature Methods* 14, 1063–1071 (2017).
79. Schlüter, A., Krause, L., Szczepanowski, R., Goesmann, A. & Pühler, A. Genetic diversity and composition of a plasmid metagenome from a wastewater treatment plant. *Journal of Biotechnology* 136, 65–76 (2008).
80. Xie, Z. & Tang, H. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics* 33, 3340–3347 (2017).

81. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, 25:1754-60. (2009).
82. Smillie, C. S. *et al.* Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480, 241–244 (2011)
83. Wattam, A. R. *et al.* PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res*. 42, D581-591 (2014).

CHAPTER 3: Run-On Sequencing Reveals Transcriptional Dynamics Within The Human Microbiome

This chapter is adapted from a manuscript under review at *Nature Microbiology*:

> AC Vill, EJ Rice, I De Vlaminck, CG Danko, IL Brito. Run-on sequencing reveals transcriptional dynamics within the human microbiome.

A.C.V., I.D.V., C.G.D., and I.L.B. conceived of the project. A.C.V. designed and performed the experiments, processed the metagenomic and RNAseq libraries, wrote the code, performed the analyses, and created the figures. E.J.R processed the PRO-seq libraries. A.C.V. and I.L.B. wrote the manuscript.

Abstract

Precise regulation of transcription initiation and elongation enables bacteria to control cellular responses to environmental stimuli. RNAseq is the most common tool for measuring the transcriptional output of bacteria, comprising predominantly mature transcripts. To gain further insight into transcriptional dynamics, it is necessary to discriminate actively transcribed loci from those represented in the total RNA pool. One solution is to capture RNA polymerase (RNAP) in the act of transcription, but current methods are restricted to culturable and genetically tractable organisms. Here, we apply precision run-on sequencing (PRO-seq) to profile nascent transcription, a method amenable to diverse species. We find that PRO-seq is well-suited to profile small, structured, or post-transcriptionally modified RNAs, which are often excluded from RNAseq libraries. When PRO-seq is applied to the human microbiome, we identify taxon-specific RNAP pause motifs. We also uncover concurrent transcription and cleavage of guide RNAs and tRNA fragments at active CRISPR and tRNA loci. We demonstrate

the specific utility of PRO-seq as a tool for exploring transcriptional dynamics in diverse microbial communities.

Introduction

Bacterial transcriptional circuitry underlies cellular stress responses, host-pathogen immune interactions, group-level dynamics, and other responses to environmental stimuli. Within the gut microbiome, these transcriptional responses may reveal pathways involved in pathogenesis or define the resilience of communities under different selective pressures. Metagenomic sequencing has been used to infer the potential functions of microbiome members, though only a fraction of genes in a cell are expressed at any given time. RNAseq has therefore been used to provide a more accurate depiction of cellular function. However, RNAseq, as performed on microbiomes, gives limited information about transcriptional dynamics across genes, requires depletion of ribosomal RNA, which may introduce species- and sequence-specific biases, and may fail to capture small, structured, or post-transcriptionally modified RNAs.

RNAseq indiscriminately sequences the pool of mature and accessible RNA molecules. In comparison, the nascent transcriptome comprises only RNA molecules that are being actively transcribed by RNA polymerase (RNAP). While total RNA sequencing has great utility in measuring steady-state levels of messenger RNA, the nascent transcriptome represents the state of a cell agnostic to the different degradation rates of RNA species. In model eukaryotes, nascent transcriptomics has aided the study of RNAP kinetics and revealed species of transient noncoding RNAs important for transcriptional regulation (reviewed in [1]).

In bacteria, nascent transcriptomics has shed light on the pausing and elongation dynamics of RNAP. However, these observations have been largely limited to genetically tractable model organisms due to significant methodological constraints. NET-seq involves the immunoprecipitation of RNAP, and thus requires either clade-specific RNAP antibodies or genetic manipulation to add epitope tags; to date, it has only been applied to *Escherichia coli* and *Bacillus subtilis* [2,3]. Other methods rely on discrimination of mature and immature RNAs by enzymatic recognition of 5' nucleotide chemistry. Differential RNA-seq (dRNA-seq) has been applied to diverse bacterial species and employs 5'-P-dependent exonuclease to degrade monophosphorylated mature transcripts, leaving immature triphosphorylated transcripts to be sequenced [4–6]. Likewise, Cappable-seq has been applied to *E. coli* and a mouse cecal microbiome and relies on a 5'-PPP capping enzyme to incorporate biotin into nascent transcripts in order to map transcription start-sites (TSS) [7]. While these methods are well-equipped to identify TSSs by mapping 5' transcript ends, they do not provide information about the position and procession of RNAP.

Precision run-on sequencing (PRO-seq) has been developed to uncover transient transcriptional signals in eukaryotes [8–10]. PRO-seq involves capturing RNA bound by engaged and actively transcribing RNAP (Figure 3.1A). In PRO-seq, cells are first permeabilized to deplete endogenous nucleotide triphosphates (NTPs), halting transcription. Then, lysates are subject to a 'run-on' reaction, which introduces biotinylated NTPs to reinitiate transcription and tag the 3' ends of nascent transcripts. Nascent RNA molecules are then enriched using streptavidin-coated beads and sequenced. Apart from eukaryotic RNAPII, transcription elongation by run-on reaction has been demonstrated for T7 RNAP [11] and mitochondrial POLRMT [12,13], suggesting that PRO-seq may be amenable to RNA polymerases across the tree of life. Here, we establish PRO-seq as a method for prokaryotes using an *E. coli* heat shock model and in human gut microbiomes, to measure nascent transcription across diverse species simultaneously.

Results

Paired PRO-seq and RNA-seq discriminate promoters by sigma factors in E. coli

The response to heat shock in *E. coli* is controlled, in part, at the level of transcription. We performed an experiment comparing RNAseq and PRO-seq in *E. coli* MG1655 cells subject to 7 minutes of heat-shock at 50 °C, hypothesizing that we could identify differences in cellular responses at genes controlled by specific sigma factors involved in the heat shock response. Pairwise scatterplots of technical triplicates show that PRO-seq is replicable in bacteria (Figure 3.1B), and, as expected, rank-ordering of transcripts suggests that PRO-seq and RNAseq signals are correlated ($\rho = 0.875$ for control; $\rho = 0.76$ for heat-shock, Spearman's rank correlation). Bacterial RNAseq requires ribosomal RNA (rRNA) depletion to reduce rRNA representation in sequencing; across *E. coli* treatments, rRNA depletion reduces rRNA from 73.7 ± 2.7% to 0.013 ± 0.004% of the library. In contrast, *E. coli* PRO-seq libraries are 1.39 ± 0.31% rRNA reads, demonstrating that PRO-seq is agnostic to bias from highly stable RNA species (Figure 3.1C). Removing the need for rRNA depletion has the benefit of reducing the potential bias introduced by such handling steps [14].

Examining the RNAseq data, there was no difference between the read depth profiles proximal to transcription start sites under control of σ70 promoters in the different treatments, consistent with the role of σ70 as the major regulator of housekeeping genes (Figure 3.1D). This was also apparent in the PRO-seq data, where the position of RNA polymerase is denoted by the 3' ends of nascent transcripts. During heat shock, the PRO-seq profiles across the same loci were comparatively reduced as transcription continues into gene bodies. This may be explained by aborted transcription of housekeeping genes in favor of genes needed to mount a response to thermal stress. Accordingly, at operons regulated by σ32, the master regulator of the heat shock response, we saw upregulation in both the RNAseq and PRO-seq datasets upon heat shock. The σ24 envelope stress response is only active in response to extreme heat stress. Transcription proximal to σ24 promoters is solely captured by PRO-seq during heat shock. These data suggest that PRO-seq enables the observation of active loading of RNA polymerase at σ24-controlled loci preceding the accumulation of mature transcripts.

We were also able to identify pause site motifs in *E. coli* using PRO-seq (Figure 3.1E). We defined PRO-seq peaks as any genomic position centered in a 50 bp window with a minimum 3’ read end depth of 10 and a Z-score of at least 5. RNAP pause sites were found at both 5’ untranslated regions and within gene bodies, suggesting that PRO-seq can be used to uncover promoter-adjacent regulatory pausing as well as elemental pausing. The sigma factor repertoires of peak-containing regulatory regions are concordant with the treatments: heat-shocked *E. coli* operons regulated by σ32 and σ24 are enriched in the merged heat shock dataset relative to the control (Figure 3.1F).

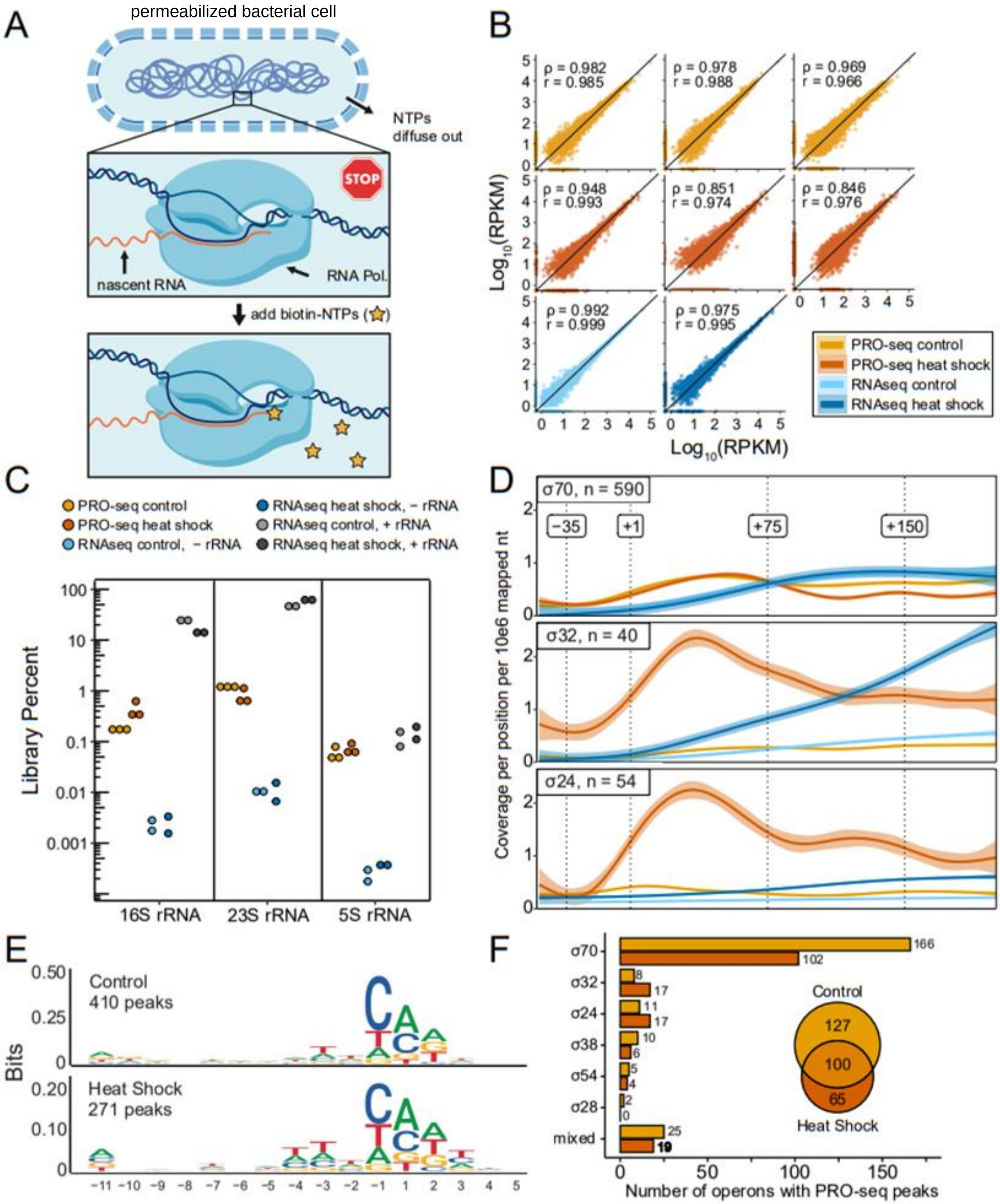

A
permeabilized bacterial cell
NTPs diffuse out
STOP
nascent RNA
RNA Pol.
add biotin-NTPs (☆)
B
ρ = 0.982 r = 0.985
ρ = 0.978 r = 0.988
ρ = 0.969 r = 0.966
ρ = 0.948 r = 0.993
ρ = 0.851 r = 0.974
ρ = 0.846 r = 0.976
ρ = 0.992 r = 0.999
ρ = 0.975 r = 0.995
Log10(RPKM)
PRO-seq control
PRO-seq heat shock
RNAseq control
RNAseq heat shock
C
PRO-seq control
PRO-seq heat shock
RNAseq control, − rRNA
RNAseq heat shock, − rRNA
RNAseq control, + rRNA
RNAseq heat shock, + rRNA
Library Percent
100
10
1
0.1
0.01
0.001
16S rRNA
23S rRNA
5S rRNA
D
σ70, n = 590
σ32, n = 40
σ24, n = 54
−35
+1
+75
+150
Coverage per position per 10e6 mapped nt
E
Control 410 peaks
Heat Shock 271 peaks
Bits
0.50
0.25
0
0.20
0.10
0
−11 −10 −9 −8 −7 −6 −5 −4 −3 −2 −1 1 2 3 4 5
F
σ70
σ32
σ24
σ38
σ54
σ28
mixed
166
102
8
17
11
17
10
6
5
4
2
0
25
19
Control
127
100
65
Heat Shock
0
50
100
150
Number of operons with PRO-seq peaks

Figure 3.1.
(A) Outline of bacterial PRO-seq. Cells are permeabilized to liberate NTPs and halt RNA polymerization. Addition of biotinylated NTPs allows RNA polymerase to incorporate a single biotin-NTP to the 3' end of the nascent RNA strand.
(B) Correlation of reads aligning to genes in *E. coli* in replicate samples in rRNA-depleted RNAseq (n=2) and PRO-seq (n=3) libraries in control and heat shock-treated *E. coli* cells. Spearman's rank correlation coefficients (ρ) and Pearson's correlation coefficients (r) are inset.
(C) Percent of reads aligning to E. coli 16S, 23S and 5S rRNA genes in RNAseq libraries without rRNA depletion, RNAseq libraries with rRNA depletion, and PRO-seq libraries made from control and heat shock-treated samples.
(D) Normalized and smoothed mean read depth profiles proximal to *E. coli* transcription start sites (TSS, position +1) under control of promoters regulated by σ70, σ32, and σ24, as annotated by RegulonDB v10.9. Replicate libraries were combined for each library type + treatment pair: rRNA-depleted RNAseq control (light blue), PRO-seq control (light orange), rRNA-depleted RNAseq heat shock (dark blue), and PRO-seq heat shock (dark orange). For RNAseq libraries, composite profiles represent full reads, whereas PRO-seq profiles only represent read 3' ends. For operons under the control of multiple promoters, plots are centered at the TSS closest to the start codon of the first gene, and operons regulated by multiple sigma factors are excluded. Bounds represent normal confidence intervals.
(E) Logos for sequences surrounding PRO-seq read 3' end peaks coincident with regulatory regions, which are defined for each operon as the sequence starting from the left-most TSS and ending with the first base of the start codon of the first gene. The range of nucleotides in physical association with *E. coli* RNAP is plotted (-11 to +5), where position -1 represents the RNAP pause site and position 1 represents the next nucleotide added.
(F) For peaks and regulatory regions described in (E), bar plots show the distribution of sigma factors regulating promoters within regulatory regions containing one or more peaks. "Mixed" regulatory regions contain promoters under control of two or more different sigma factors. The inset Venn diagram shows the overlap between peak-containing regulatory regions for control and heatshock libraries, replicates merged.

PRO-seq is suitable for diverse species of human-associated microbiota

We next investigated the utility of PRO-seq to capture nascent transcripts from diverse microbial communities. We performed PRO-seq and RNAseq, with replicates, on gut microbiome samples from two healthy individuals. The first step in performing PRO-seq is permeabilization, which results in the rapid depletion of NTPs from cells, thus halting transcription. We were concerned that the permeabilization protocol used on *E. coli* would be insufficient to permeabilize microbiome-derived cells, as harsher lysis methods are typically required to minimize extraction biases [15,16]. As heat and proteinase treatment are incompatible with PRO-seq, we opted for bead-beating in a nonionic detergent buffer, preserving halted RNAP-RNA complexes (which can be very stable [17]) for subsequent run-on reactions. Overall, we found strong concurrence between replicate samples (ρ = 0.954 and ρ = 0.938 for the two microbiome samples, Spearman's rank correlation) (Figure 3.2A). We subset reads according to the metagenomic assembled genomes to which they aligned, and found an enrichment of certain strains in the PRO-seq libraries compared to the RNAseq libraries (Figure 3.2B). With the exception of *Ruminococcus bromii*, Firmicutes were less well represented in the PRO-seq libraries than RNAseq libraries. This may be attributed to more efficient lysis of Gram negative Bacteroidetes. Alternatively, Bacteroidetes are highly abundant in both of these samples, which may reflect their overall higher growth rates and possibly more active transcription.

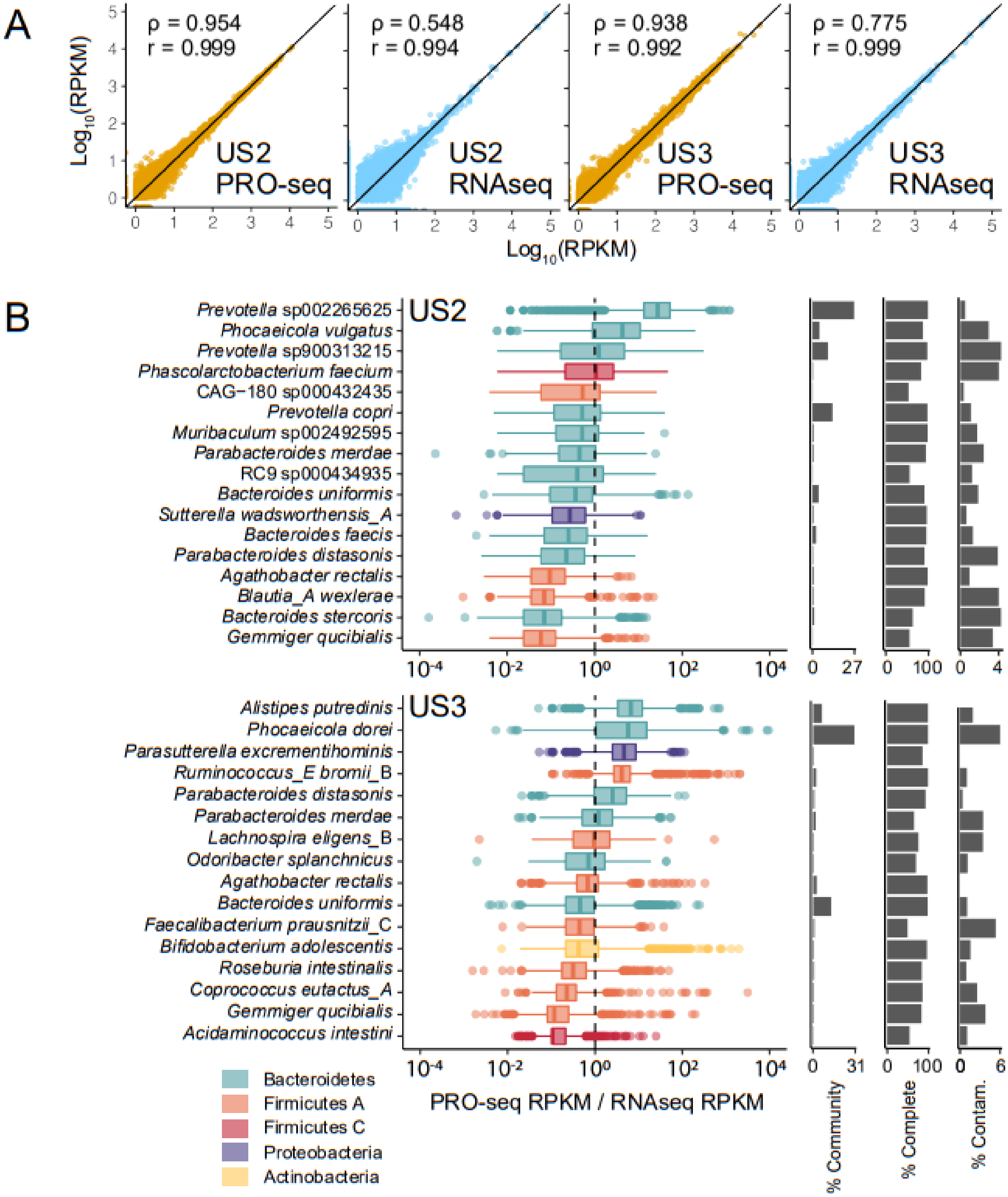


Figure 3.2. **(A)** Correlation of reads aligning to metagenomic features for replicate samples in rRNA-depleted RNAseq (n = 2) and PRO-seq (n = 2) libraries. Spearman's rank correlation coefficients (ρ) and Pearson's correlation coefficients (r) are inset. **(B)** For metagenomic bins that are least 90% complete with less than 5% contamination, box plots show the distribution of PRO-seq RPKM divided by RNAseq RPKM for each feature; replicate libraries are merged. Dotted lines demarcate equal coverage in both sequencing types. For each bin, relative abundance, percent completeness, and percent contamination are provided (right).

PRO-seq captures concurrent transcription and cleavage of CRISPR RNAs

We next turned our focus towards specific genomic loci that tend to be difficult to capture by RNAseq. Non-coding RNAs may be structured or sequestered in protein complexes, affecting their representation in metatranscriptomic experiments [18]. CRISPR arrays are comprised of repeated elements and unique spacers which are transcribed and cleaved to create functional guide RNAs. RNAseq reads that align to CRISPR loci are typically sparse [19]. Whereas CRISPR arrays are less represented in our RNAseq data as well, we see active transcription across these loci in the PRO-seq data. Furthermore, at CRISPR loci with high PRO-seq coverage, we observe a distinct periodic pattern with a pile-up of PRO-seq read 5' ends at consistent positions within repeats (Figures 3.3A, 3.4A & B). When examining further, we found these pile-ups occur at predicted sites of endonuclease hydrolysis, corresponding to the 3' ends of the predicted repeat stem loops (Figure 3.3B). It is currently unclear whether transcription of the full pre-crRNA precedes endonuclease processing or if pre-crRNAs are co-transcriptionally cleaved. Our data suggest that the latter is the case, as the capture of individual crRNAs in our PRO-seq libraries implies those transcripts are bound by RNAP. In support of this finding, on a contig for which we were able to assemble a CRISPR array and its associated Cas proteins, we find active expression of the upstream Cas5d endonuclease (Figure 3.3C). Given that in most CRISPR systems, the newest spacers are incorporated at the end of the array closest to the leader [20], this model of co-transcriptional cleavage is consistent with the need to rapidly assemble CRISPR-Cas complexes to respond to incoming phage or other mobile genetic elements.

PRO-seq profiles indicate additional transcriptional dynamics at CRISPR loci. For instance, we detect anti-sense transcription for a subset of the spacers closest to the leader (Figure 3.3A), a phenomenon that has been previously observed but whose functional significance is poorly understood [19,21–23]. In some of the detected CRISPR arrays, pile-ups of

PRO-seq read 5’ ends are coincident with spacers, not repeats, indicative of pre-crRNA processing in some systems employing RNase III [24]. This implies that PRO-seq can capture transcription across CRISPR arrays that undergo diverse modes of maturation. We also observe regular 3’ end transcript pile-ups at specific nucleotides within the CRISPR array (Figures 3.3A & 3.4), which may point to regulation at these loci at the level of RNAP procession.

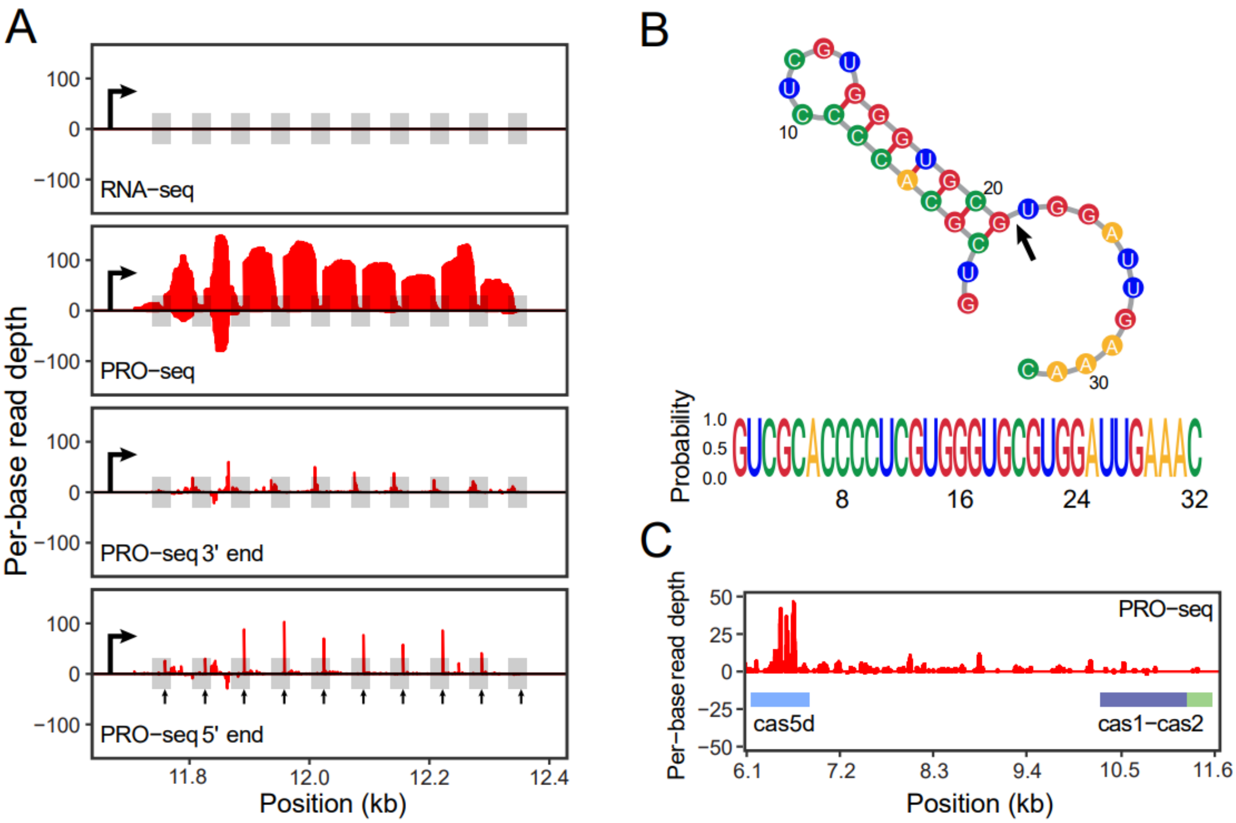


Figure 3.3. **(A)** Coverage of PRO-seq and RNA-seq reads across a CRISPR array in a US2 *Prevotella* sp. contig. Shaded boxes represent repeats. The large black arrow in each panel represents the leader sequence containing a putative promoter. Small black arrows in the "PRO-seq 5' end" panel correspond to the predicted site of crRNA cleavage proximal at the base of the repeat stem loop. **(B)** Predicted crRNA repeat secondary structure. The black arrow points to the phosphodiester bond that is likely cleaved by Cas5d during pre-crRNA processing, which marks the same position in the repeat as the small arrows in (A). The sequence logo shows perfect conservation of the repeat sequence for this array. **(C)** PRO-seq captures nascent transcription of cas5d upstream of and contiguous with the CRISPR array.

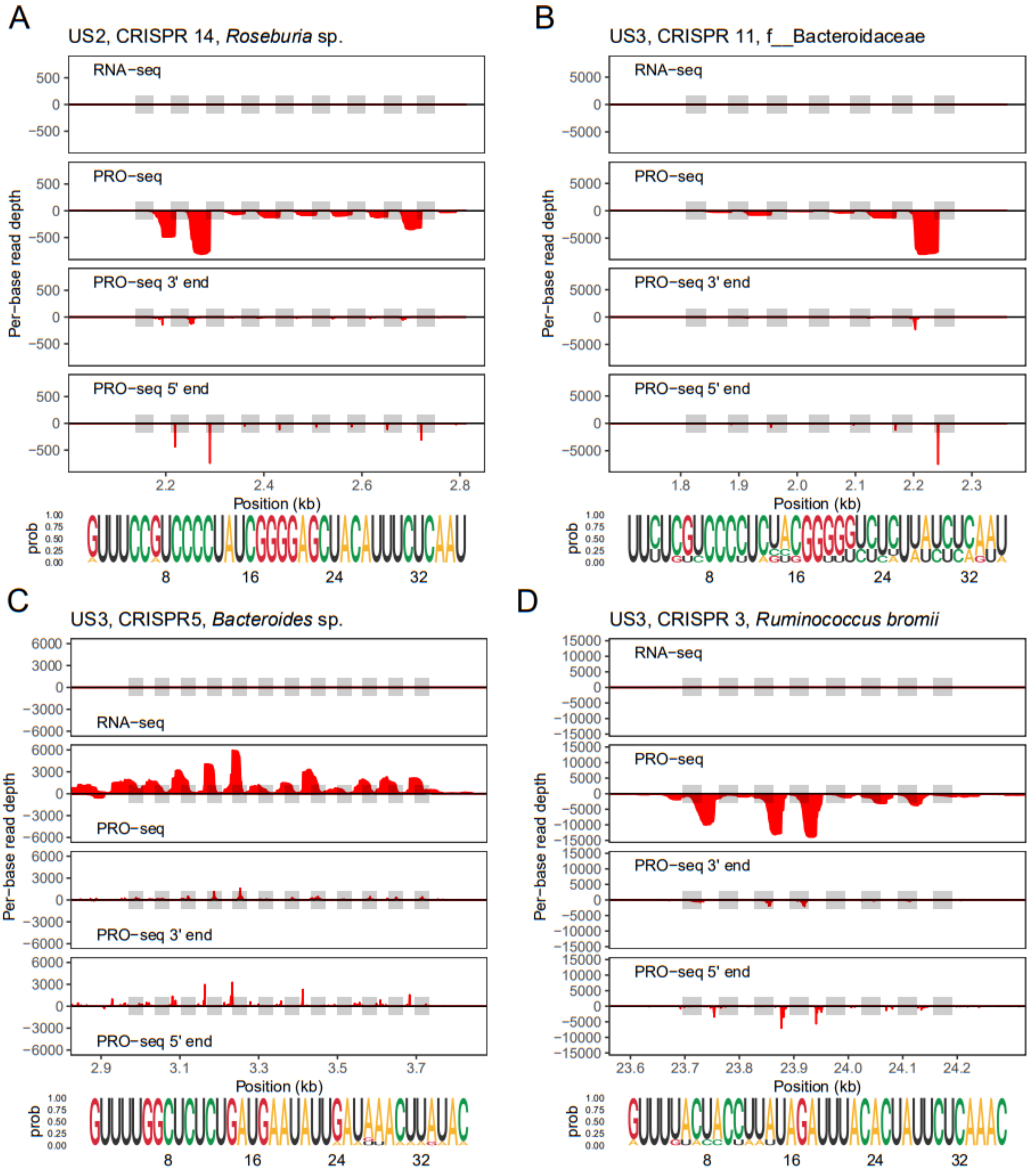


Figure 3.4. Strand-specific RNAseq and PRO-seq read depths, in addition to PRO-seq reads' 3'- and 5'-ends, are plotted for several well-covered CRISPR loci. Shaded boxes represent repeats. Sequence logos below each plot show repeat conservation. As in Figure 3.3, **(A)** and **(B)** show PRO-seq read 5' end pile-ups at the same position across repeats. **(C)** and **(D)** show PRO-seq read 5' end pile-ups within spacers.

Concurrent transcription and cleavage also occurs at tRNA loci

Non-coding RNAs (ncRNAs) are often decorated with post-transcriptional modifications that render them difficult to amplify using reverse transcriptase [25,26]. In particular, tRNA derivatives formed by cleavage and base-specific modifications are interesting because they serve functions beyond their canonical role in translation [27,28], with implications for pathogenesis [29,30] and bacterial physiology [31,32]. Quantifying microbiome tRNA abundances often requires mass spectrometry or tailored protocols to remove these modifications prior to sequencing [33,34]. We compared active transcription of tRNA loci in PRO-seq libraries to mature transcripts in RNAseq libraries. We initially focused on three *Prevotella* species found in high abundance in one of the samples for which we could annotate numerous tRNA isoforms. We noticed that a greater proportion of PRO-seq reads could be attributed to these loci than RNAseq reads (0.21 ± 0.07% vs. 0.013 ± 0.016%) and that a larger number of isoforms per tRNA were observed (Figure 3.5), highlighting the utility of PRO-seq to capture differences in ncRNA transcription between closely related bacterial strains.

Among metagenomic tRNA loci, we noticed pile-ups of PRO-seq read 5'-ends within tRNA gene bodies (Figures 3.6A,B & 3.7), a phenomenon also observed within the cultured *E. coli* heat-shock samples (Figure 3.8). We hypothesized that this may be due to processing of tRNAs into tRNA fragments, which act as signaling molecules in many bacterial species [30]. In one example of a tRNA gene cluster in *Ruminococcus bromii*, we noticed PRO-seq read 5' end pile-ups in Arg, His and Lys tRNA genes, corresponding to predicted tRNA cleavage sites within each anticodon loop (Figure 3.6C). Transcription of this locus was absent in the RNAseq data, despite comparable transcription detected across both RNAseq and PRO-seq libraries at protein-coding genes on the same contig (Figure 3.6D). This example, among others present in a diverse set of species (Figure 3.7), suggests that, similarly to CRISPR loci, tRNA processing is temporally coupled with transcription.

There are two alternative hypotheses concerning the interpretation of the pile-up of PRO-seq read 5’ ends within tRNA anticodon loops. (1) In the PRO-seq protocol, nascent RNAs are fragmented by alkaline hydrolysis prior to 3’ adapter ligation. ssRNA is more susceptible to chemical hydrolysis than dsRNA[35], so unprotected bases within tRNA loops may be overrepresented as sites for hydrolysis products for a given tRNA isoform. If this is the case, we would expect to see similar peaks in the 5’ end pile-up at the T- and D-arms of nascent tRNAs. However, peaks within T- and D-arms are rare in the PRO-seq traces relative to peaks coincident with anticodon loops, suggesting that preferential hydrolysis of non-base-paired RNA cannot fully explain the patterns we observe. (2) A crucial step in all RNA sequencing protocols is reverse transcription, by which DNA is created from an RNA template for library construction and sequencing. Reverse transcriptase (RT) is sensitive to both the structural conformation and chemical modifications of the template RNA strand[36,37], and anticodon loops are common sites for methylation in bacteria[38,39]. Therefore, tRNA anticodon loops may be a common site for RT stalling, leading to false inference of stall sites as the true 5’ ends of nascent transcripts. However, 5’ adapter ligation is carried out *before* reverse transcription, so cDNAs made from aborted RT products will lack a 5’ adapter and the concomitant PCR handle. It is therefore unlikely that such truncated cDNAs would be represented in the sequencing library, and RT stalling is therefore insufficient to explain the patterns we observe.

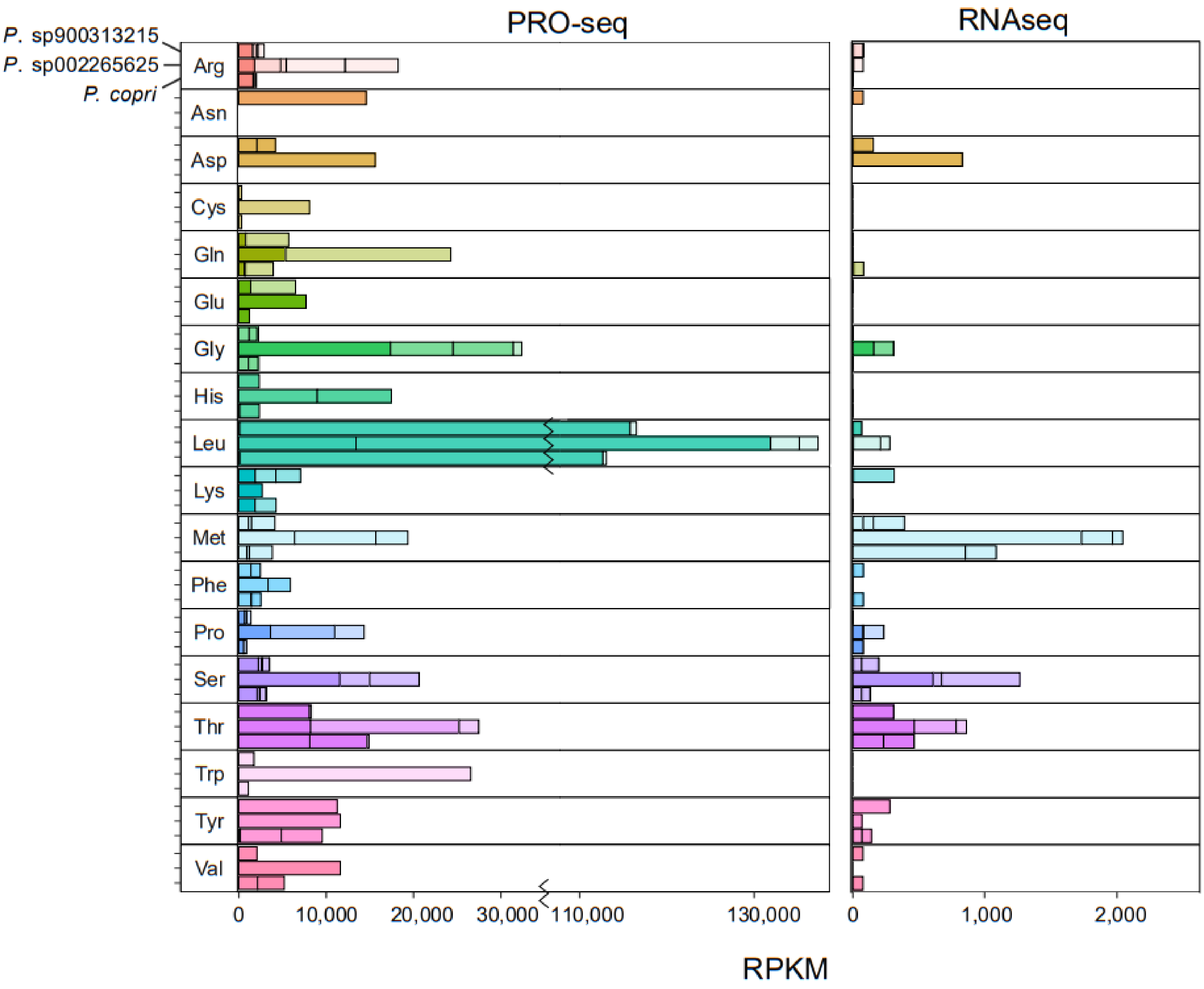


Figure 3.5. tRNA genes were identified in three highly complete US2 bins: *Prevotella* sp900313215, *Prevotella* sp002265625 and *Prevotella copri*. Different colors in the stacked bar plots represent different tRNA isoforms.

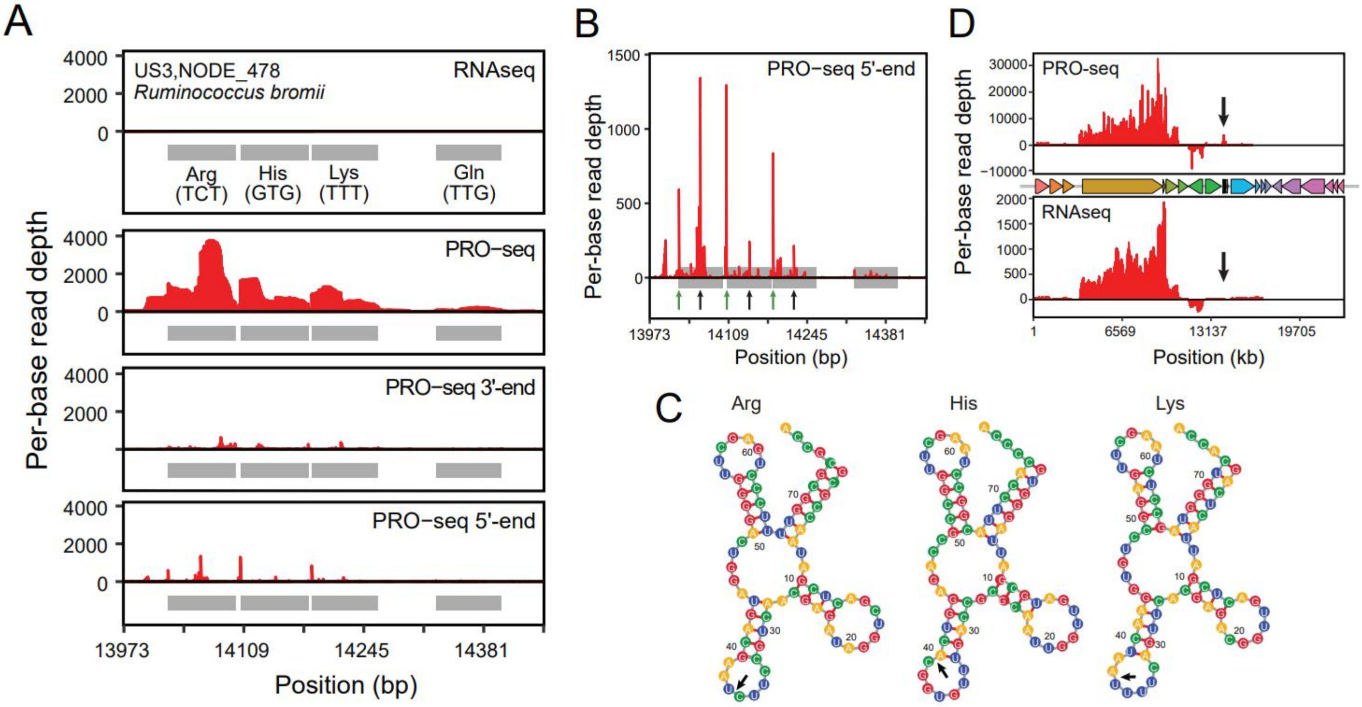
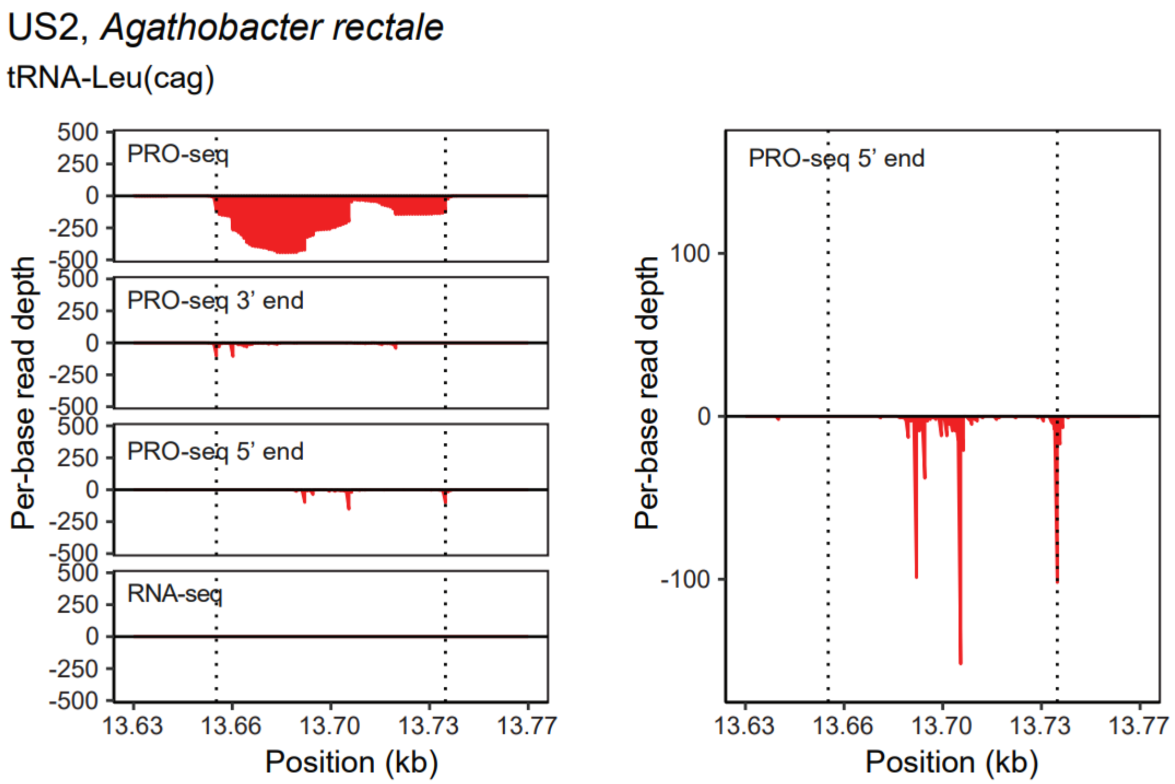
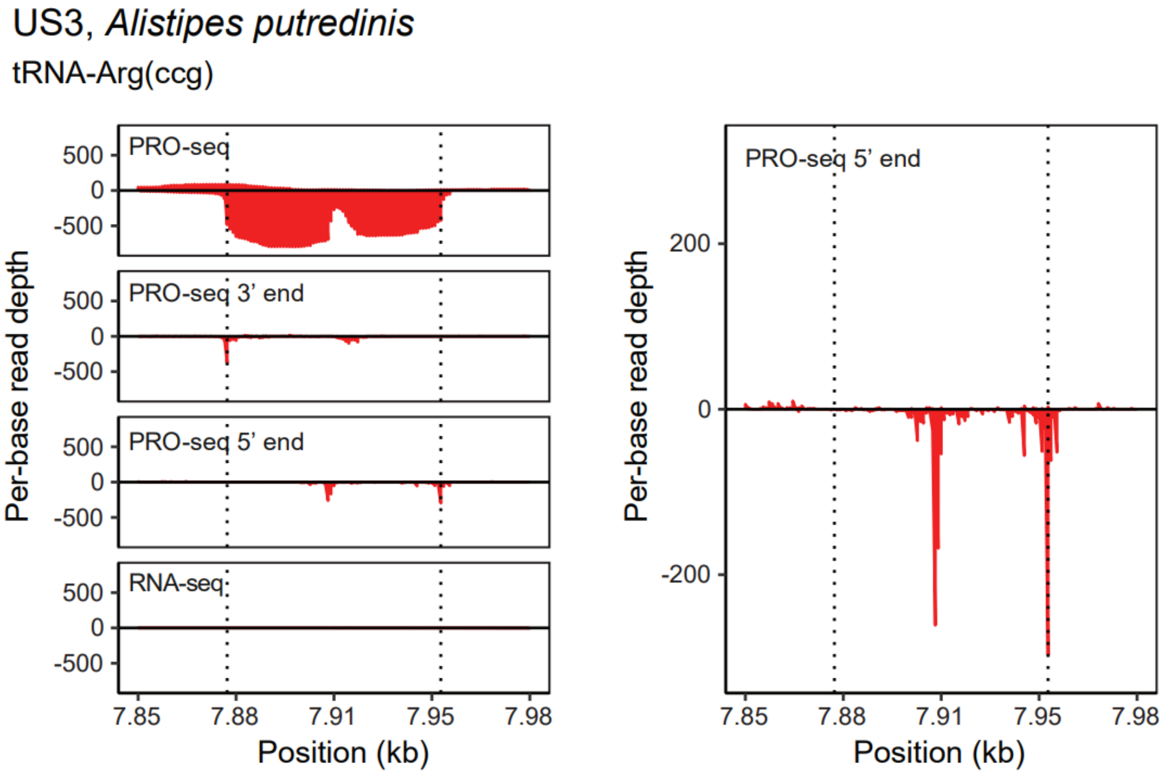


Figure 3.6. **(A)** Coverage of PRO-seq and RNA-seq reads across a tRNA gene cluster in a US3 *Ruminococcus bromii* contig. Shaded boxes represent tRNA genes. Small black arrows in the “PRO-seq 5’ end” panel correspond to the base of the predicted repeat stem loop that serves as the site of crRNA cleavage. **(B)** Coverage of PRO-seq 5’ ends for the tRNA array shown in (A). Green arrows show the starts of the tRNA genes. Black arrows show the predicted cleavage sites within anticodon loops.
(C) Coverage of PRO-seq and RNAseq reads over the entire contig. The positions of gene bodies are shown (middle). Black arrows point to the site of the tRNA array shown in (A). **(D)** Predicted structures and cleavage sites (black arrows) of tRNA genes shown in (A).



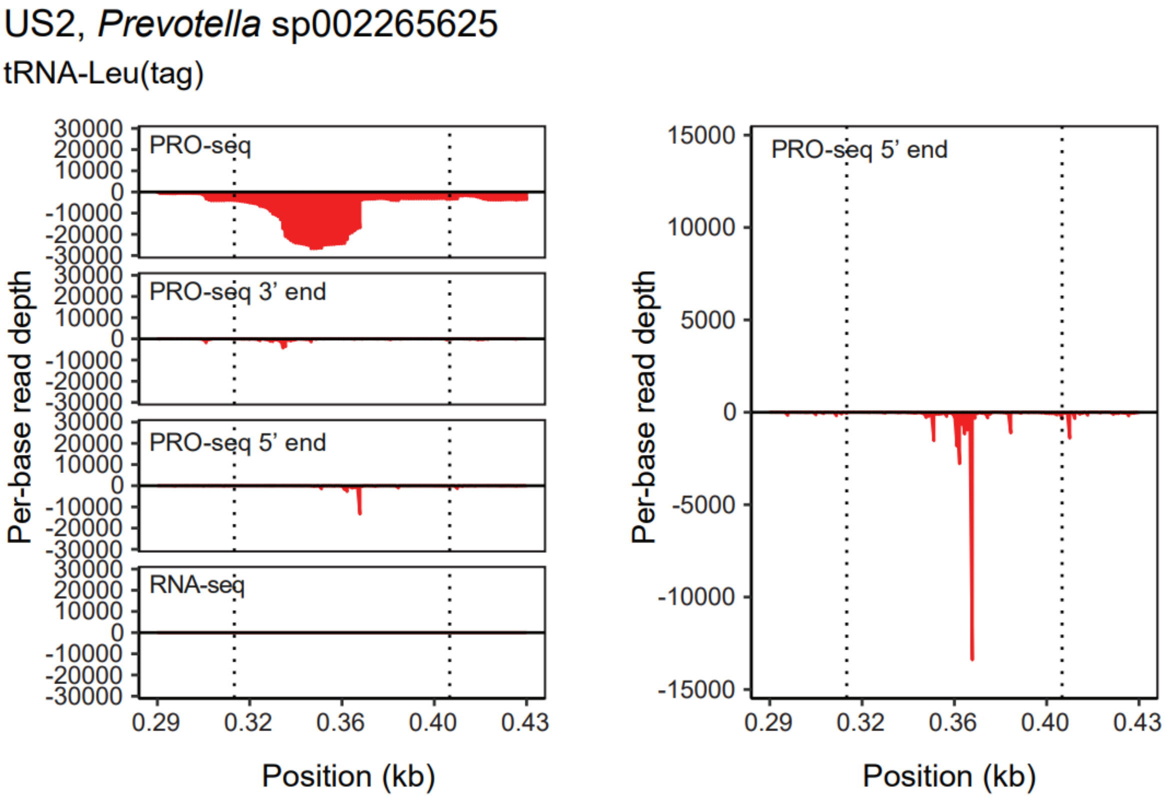
US2, *Agathobacter rectale*

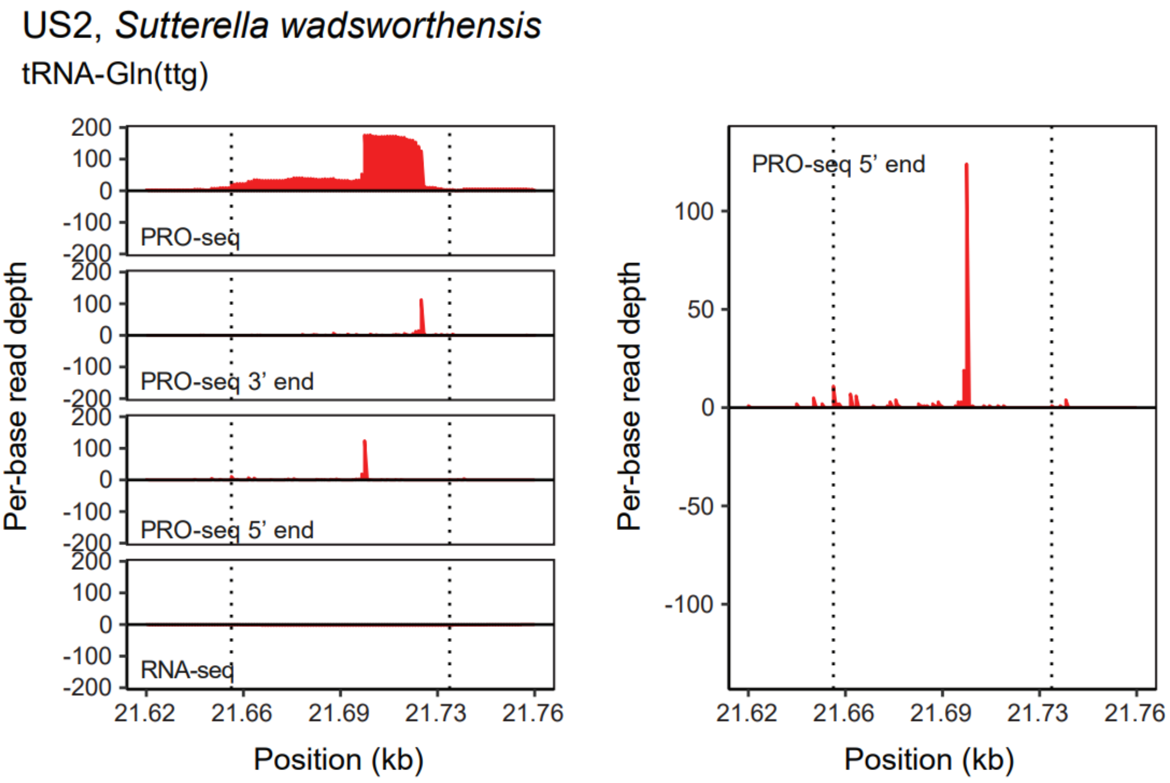
tRNA-Leu(cag)



US3, *Alistipes putredinis*

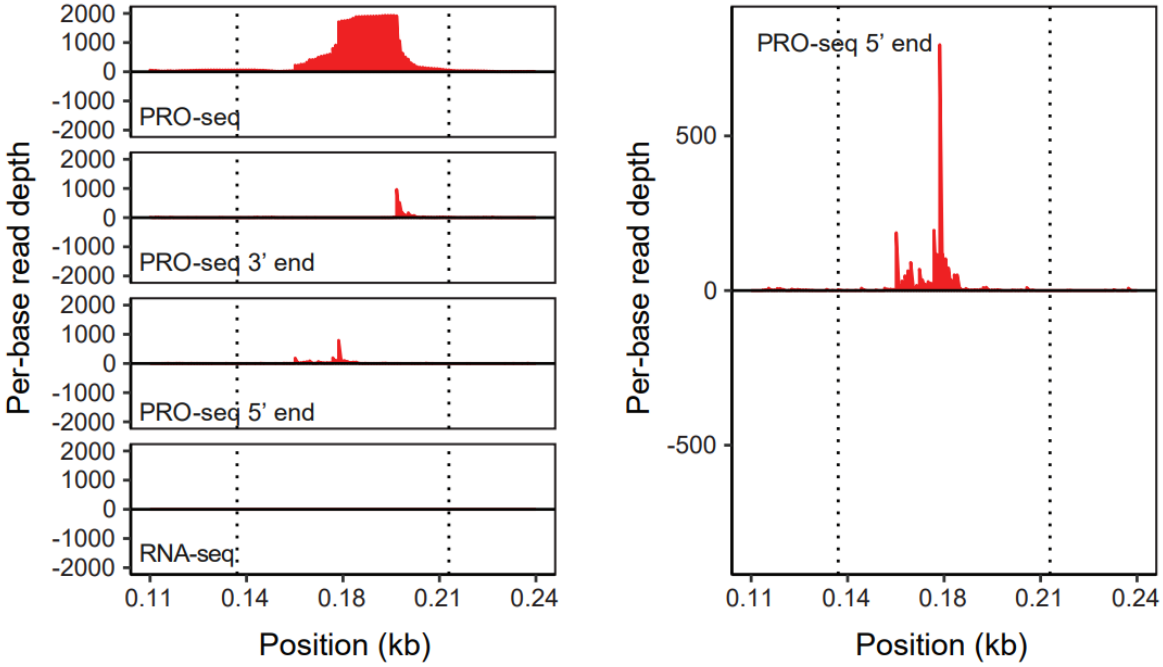
tRNA-Arg(ccg)


US2, *Prevotella* sp002265625
tRNA-Leu(tag)
PRO-seq
PRO-seq 3' end
PRO-seq 5' end
RNA-seq
30000
20000
10000
0
-10000
-20000
-30000
Per-base read depth
0.29
0.32
0.36
0.40
0.43
Position (kb)
PRO-seq 5' end
15000
10000
5000
0
-5000
-10000
-15000
Per-base read depth
Position (kb)


US2, *Sutterella wadsworthensis*
tRNA-Gln(ttg)
200
100
0
-100
-200
PRO-seq
PRO-seq 3' end
PRO-seq 5' end
RNA-seq
Per-base read depth
21.62
21.66
21.69
21.73
21.76
Position (kb)
PRO-seq 5' end
100
50
0
-50
-100
Per-base read depth
Position (kb)

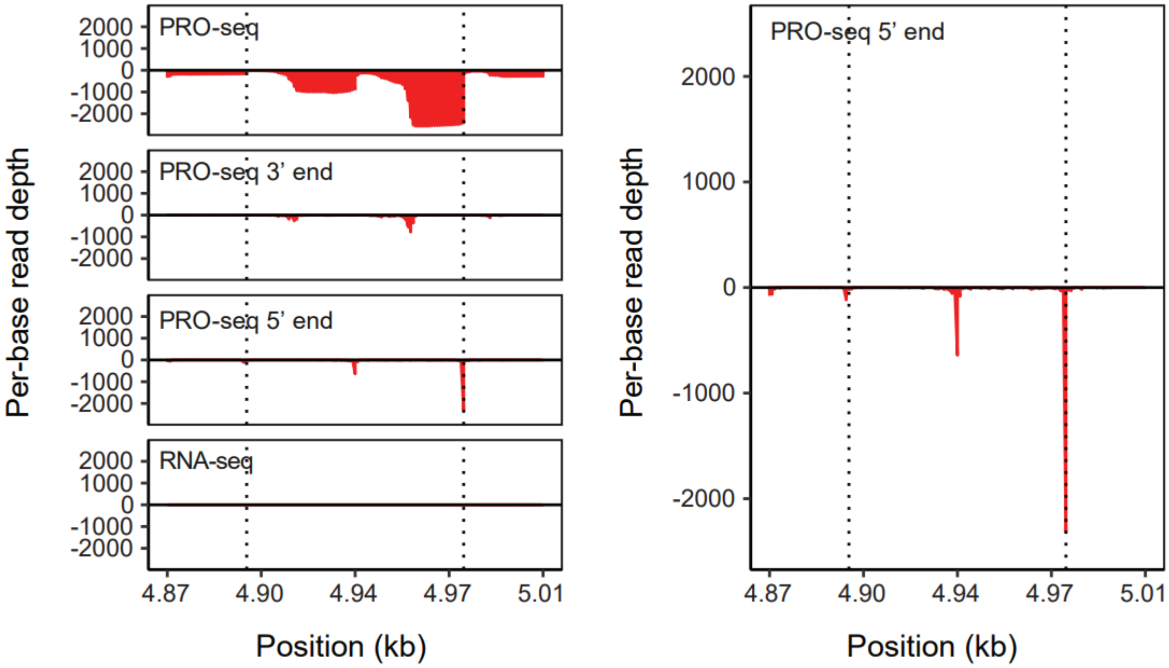
US2, *Prevotella* sp900313215

tRNA-Trp(cca)



US3, *Bifidobacterium adolescentis*

tRNA-Tyr(gta)



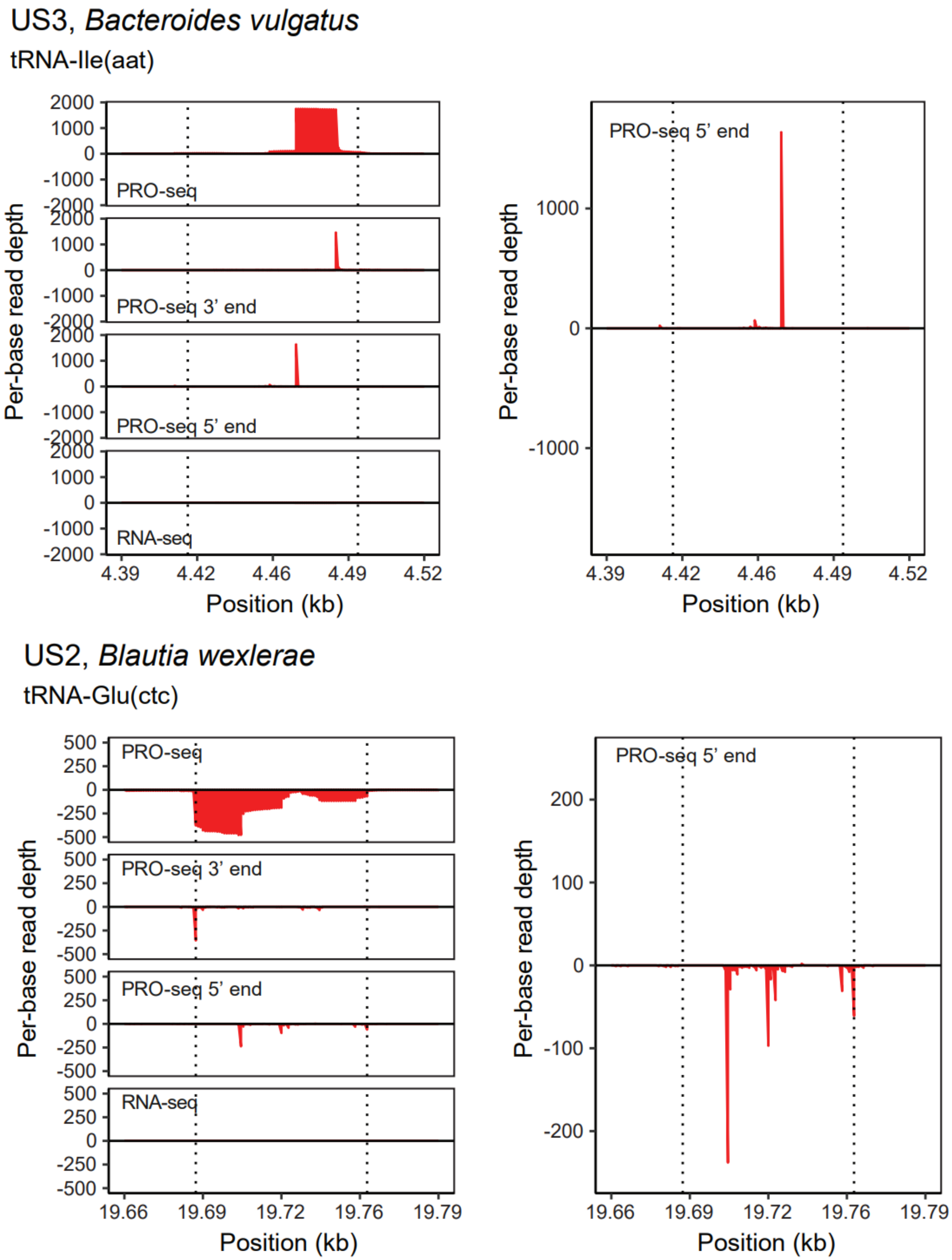
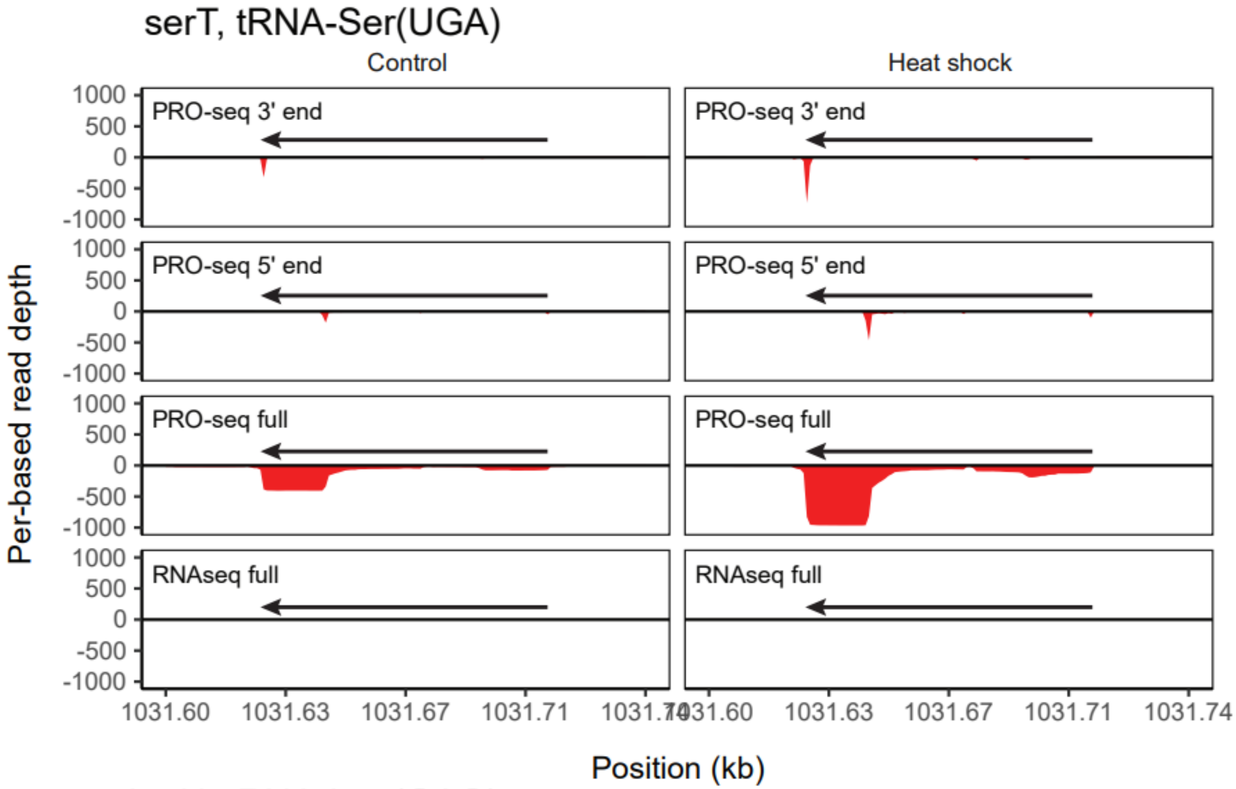
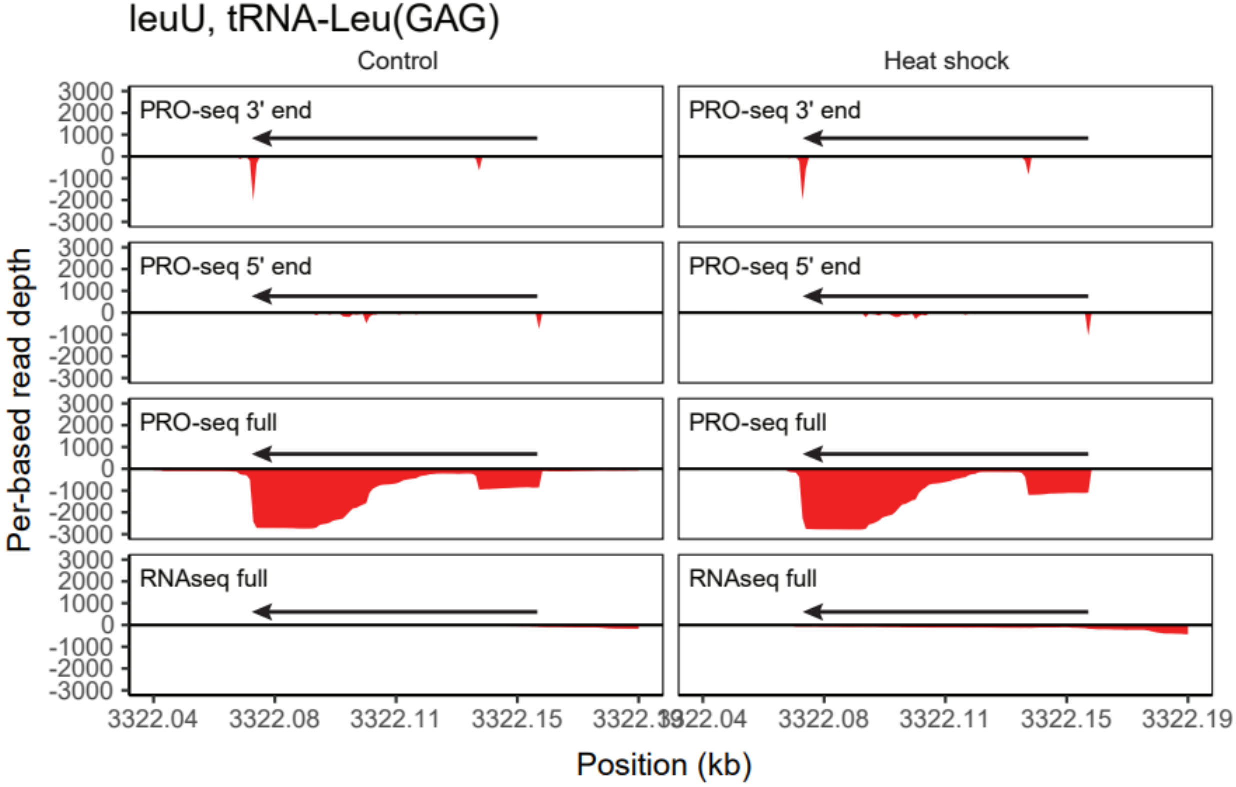
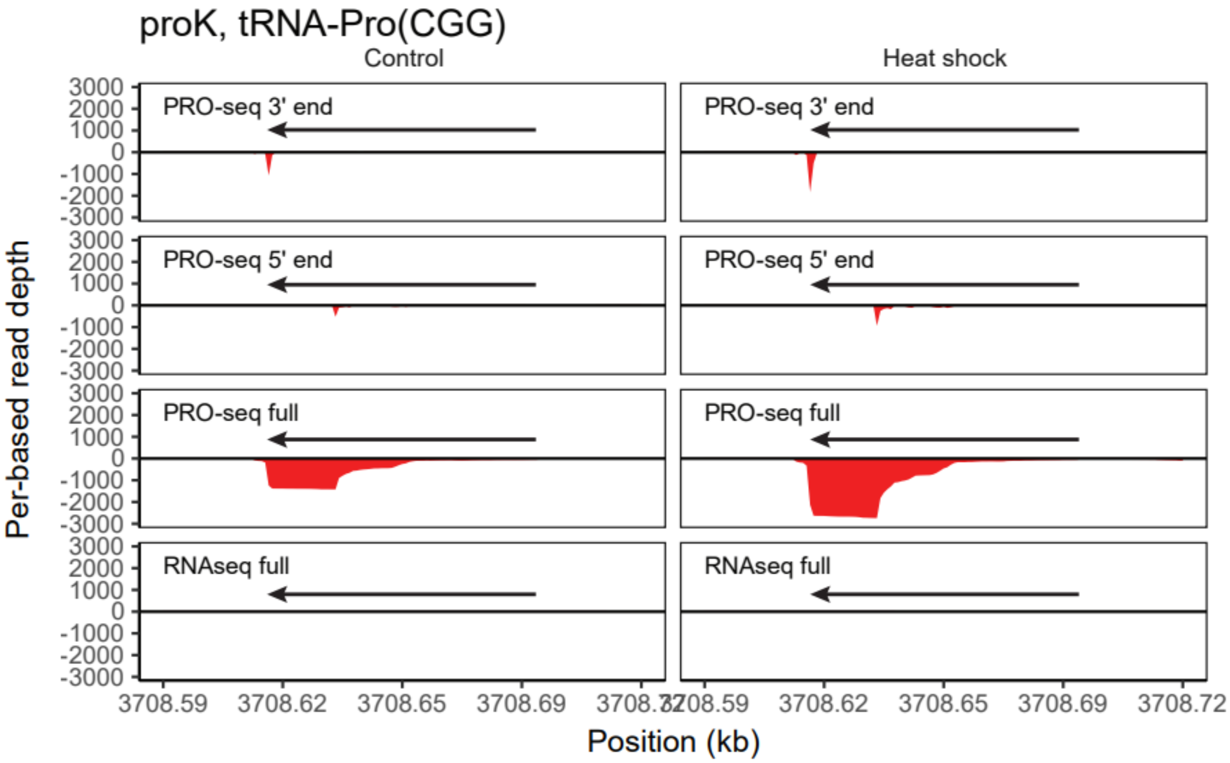
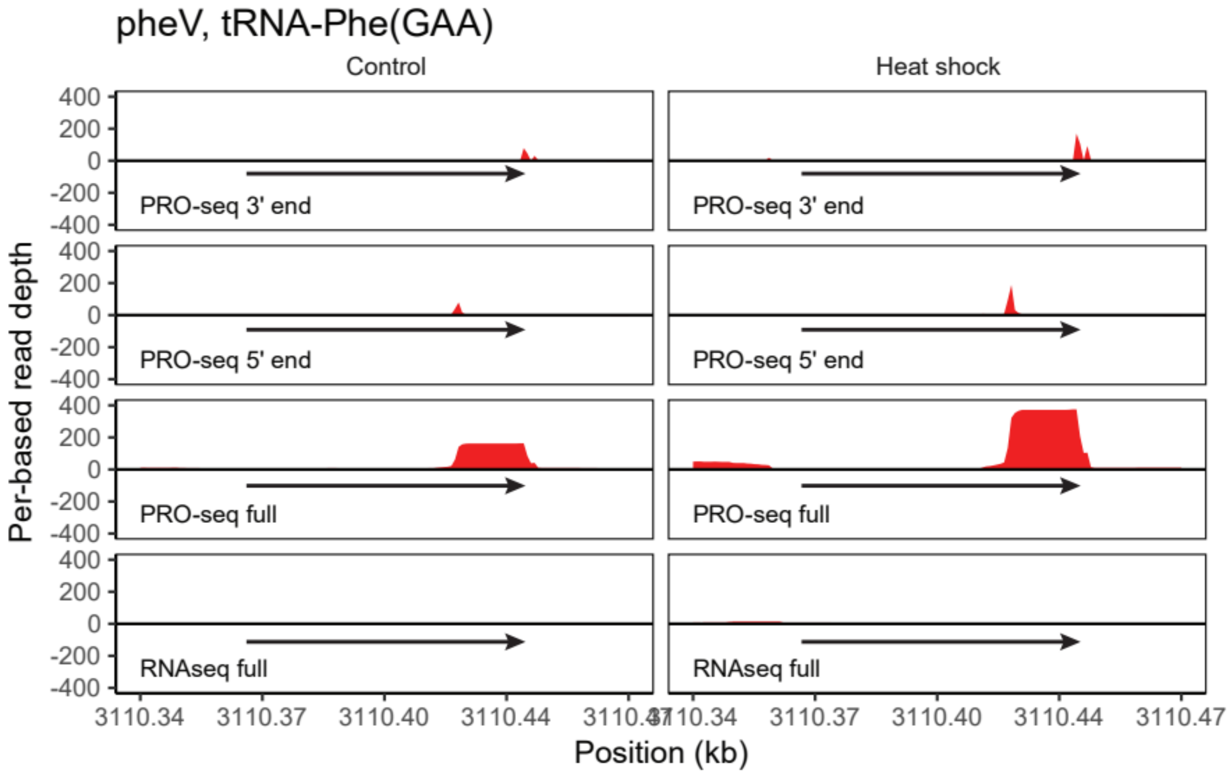


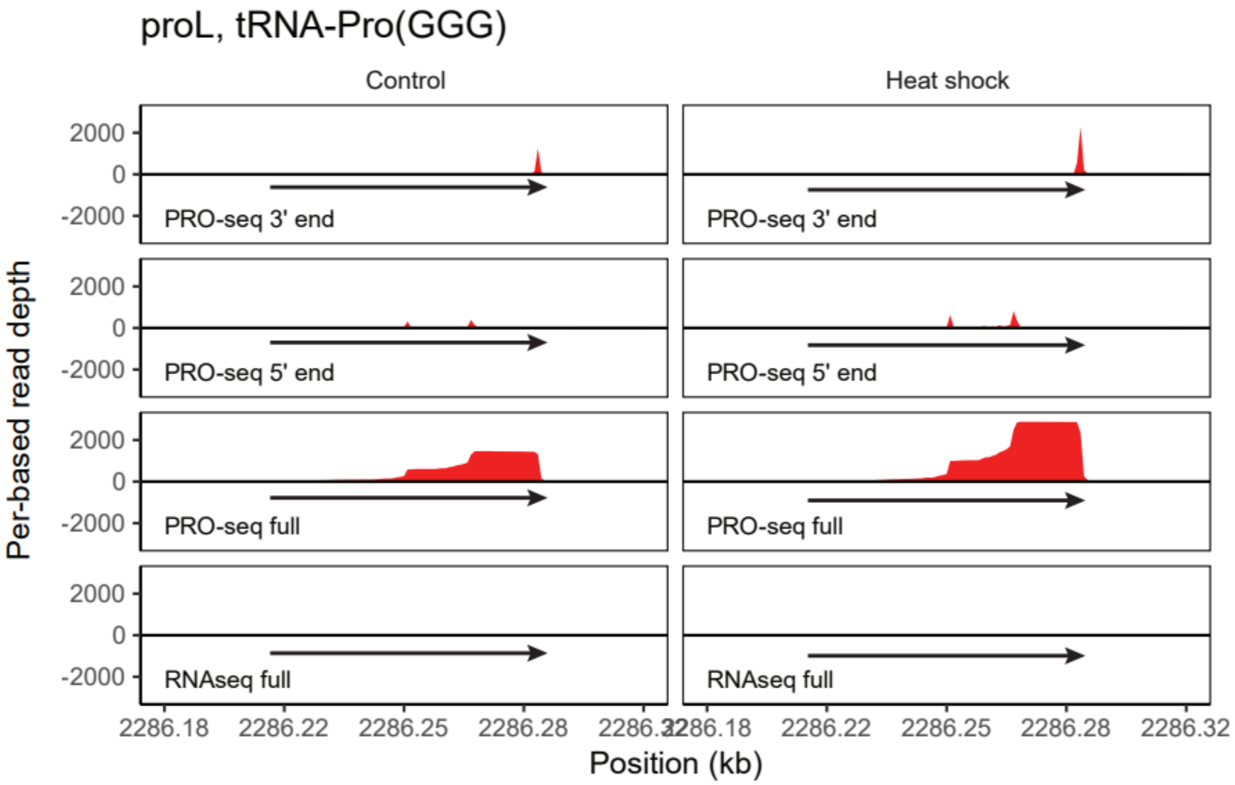
Figure 3.7. Representative tRNA genes, listed according to the sample, species annotation, and anticodon, are depicted from the two human microbiome samples. PRO-seq coverage, pile-up of PRO-seq 3'and 5' read ends, and RNAseq coverage are shown for each tRNA gene (left). A zoomed-in PRO-seq read 5' end pile-up is shown for each tRNA gene (right).

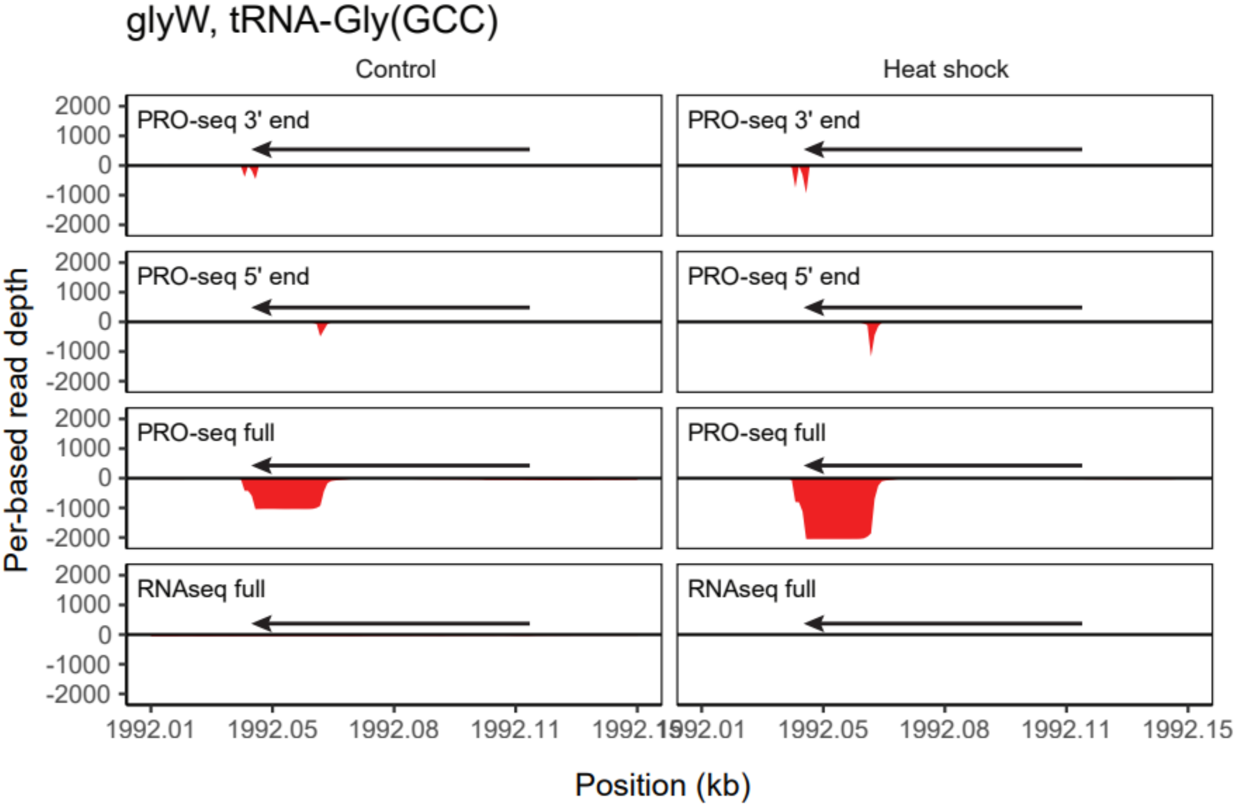

serT, tRNA-Ser(UGA)
Control
Heat shock
PRO-seq 3' end
PRO-seq 5' end
PRO-seq full
RNAseq full
1000
500
0
-500
-1000
Per-based read depth
1031.60
1031.63
1031.67
1031.71
1031.74
Position (kb)


leuU, tRNA-Leu(GAG)
Control
Heat shock
PRO-seq 3' end
PRO-seq 5' end
PRO-seq full
RNAseq full
3000
2000
1000
0
-1000
-2000
-3000
Per-based read depth
3322.04
3322.08
3322.11
3322.15
3322.19
Position (kb)


proK, tRNA-Pro(CGG)
Control
Heat shock
Per-based read depth
3000
2000
1000
0
-1000
-2000
-3000
PRO-seq 3' end
PRO-seq 5' end
PRO-seq full
RNAseq full
3708.59
3708.62
3708.65
3708.69
3708.72
Position (kb)


pheV, tRNA-Phe(GAA)
Control
Heat shock
Per-based read depth
400
200
0
-200
-400
PRO-seq 3' end
PRO-seq 5' end
PRO-seq full
RNAseq full
3110.34
3110.37
3110.40
3110.44
3110.47
Position (kb)


proL, tRNA-Pro(GGG)
Control
Heat shock
Per-based read depth
2000
0
-2000
PRO-seq 3' end
PRO-seq 5' end
PRO-seq full
RNAseq full
2286.18
2286.22
2286.25
2286.28
2286.32
Position (kb)


glyW, tRNA-Gly(GCC)
Control
Heat shock
Per-based read depth
2000
1000
0
-1000
-2000
PRO-seq 3' end
PRO-seq 5' end
PRO-seq full
RNAseq full
1992.01
1992.05
1992.08
1992.11
1992.15
Position (kb)

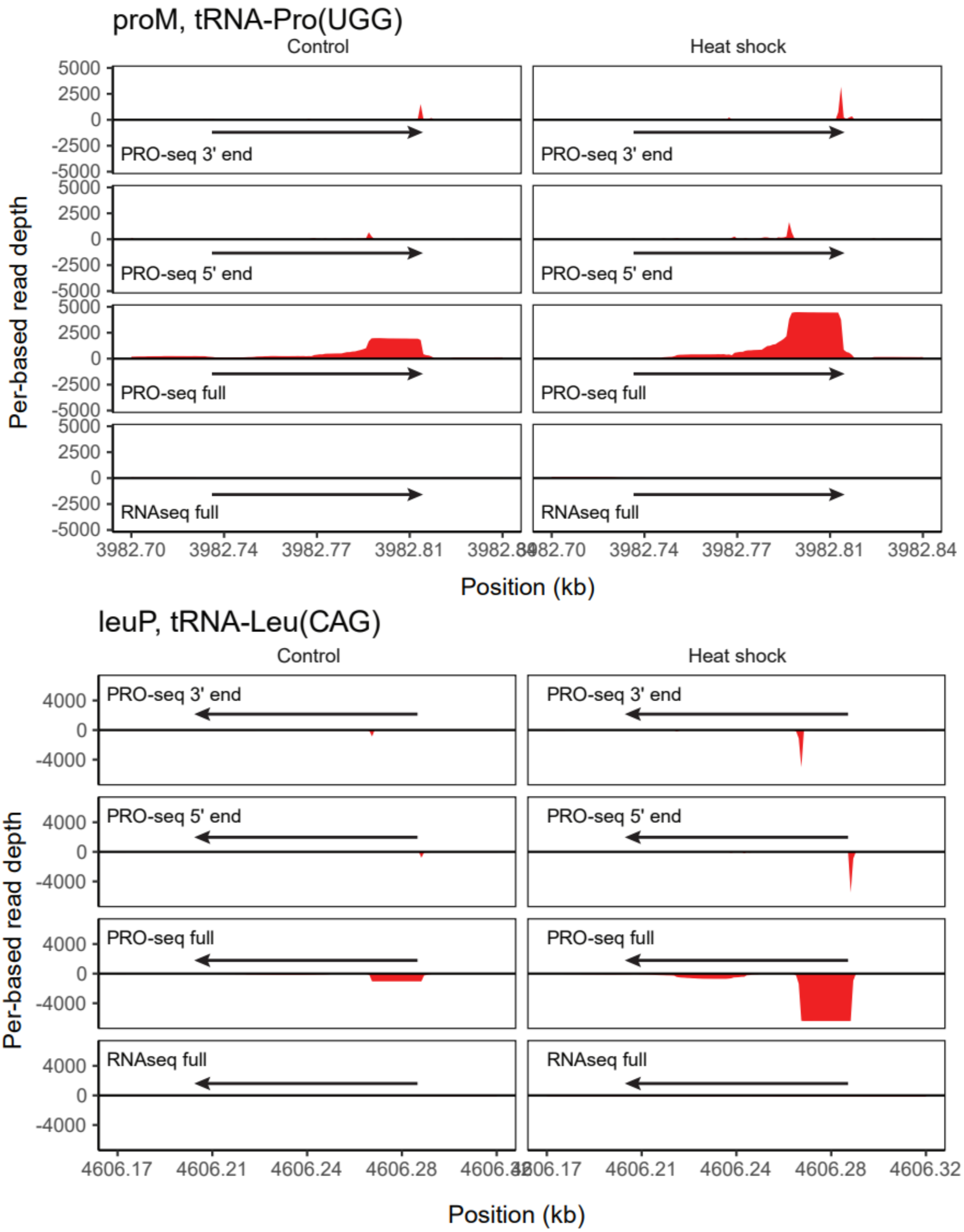


Figure 3.8. Representative *E. coli* tRNA genes, listed by isoform, are shown for control (left) and heat shock (right) conditions. PRO-seq coverage, pile-up of PRO-seq 3'and 5' read ends, and RNAseq coverage are shown for each tRNA gene.

RNAP pause-site motifs annotated in diverse species

The procession of RNAP across a gene body can be interrupted by pauses at specific sequences or secondary structures. These pauses are involved in synchronizing transcription and translation, coordinating the recruitment of regulatory factors, and the dissociation of elongation complexes [40,41]. Transcription pause sites have previously been shown to differ between *E. coli* and *B. subtilis* [2,42], suggesting that they may vary across the members of the gut microbiome. To test this, we called PRO-seq 3' end peaks across gene bodies in well-covered and near-complete metagenomic assembled genomes. We found concordant consensus pause sites between members of the Bacteroidetes phylum, a *Parabacteroides* and a *Prevotella* species, across two individuals (Figure 3.9). This TA-rich consensus site, was similar to that found in the Firmicutes species *Agathobacter rectale*, found in both individuals' microbiomes. On the contrary, the pause site motif identified in *Sutterella wadsworthensis*, a Proteobacteria, was more closely aligned with the consensus pause site identified in our earlier experiments with *E. coli*, a different Proteobacteria. These observations suggest that PRO-seq is applicable to a wider range of comparative transcription dynamics questions across diverse species.

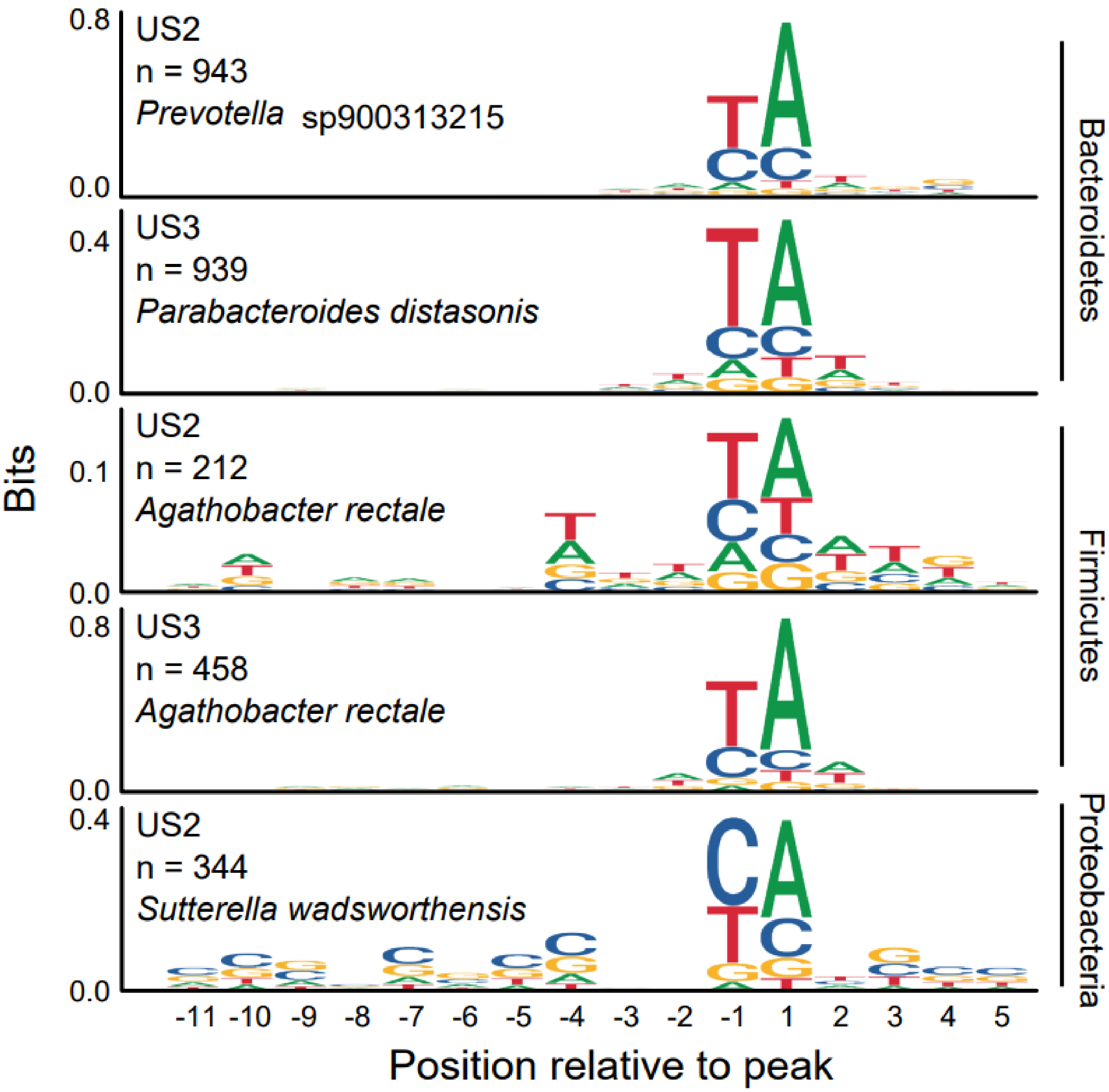
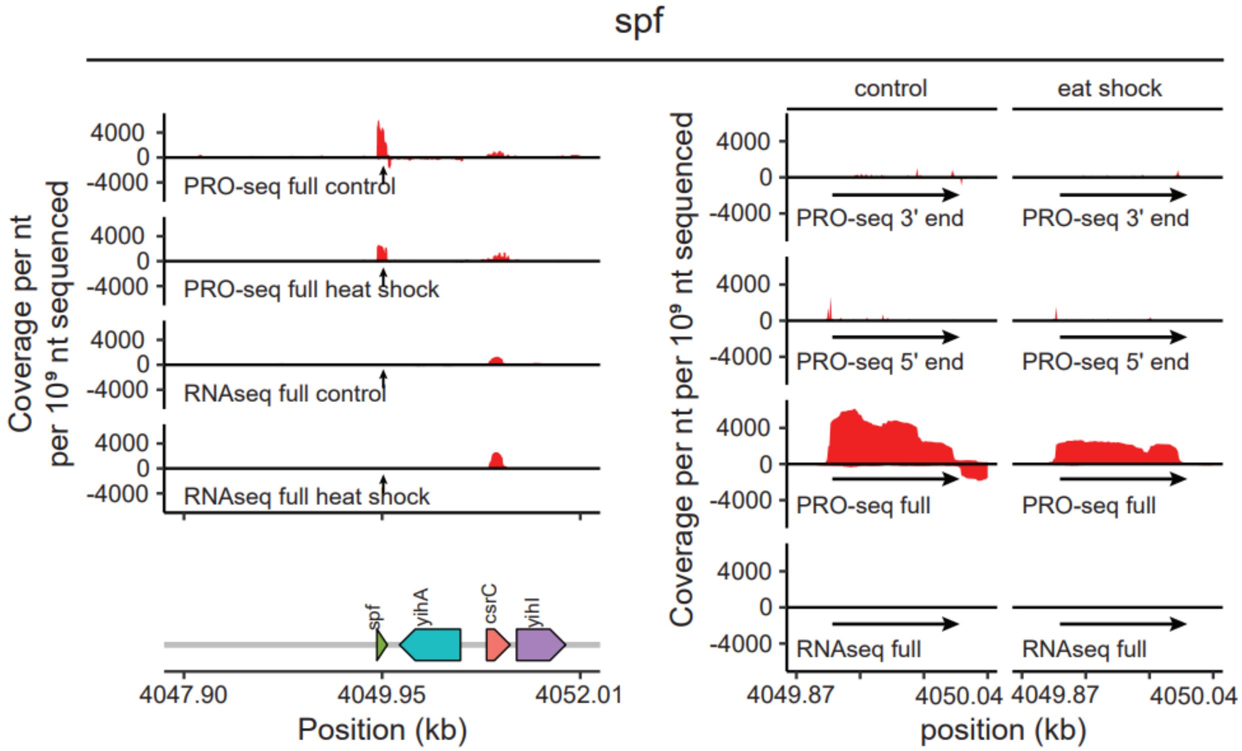


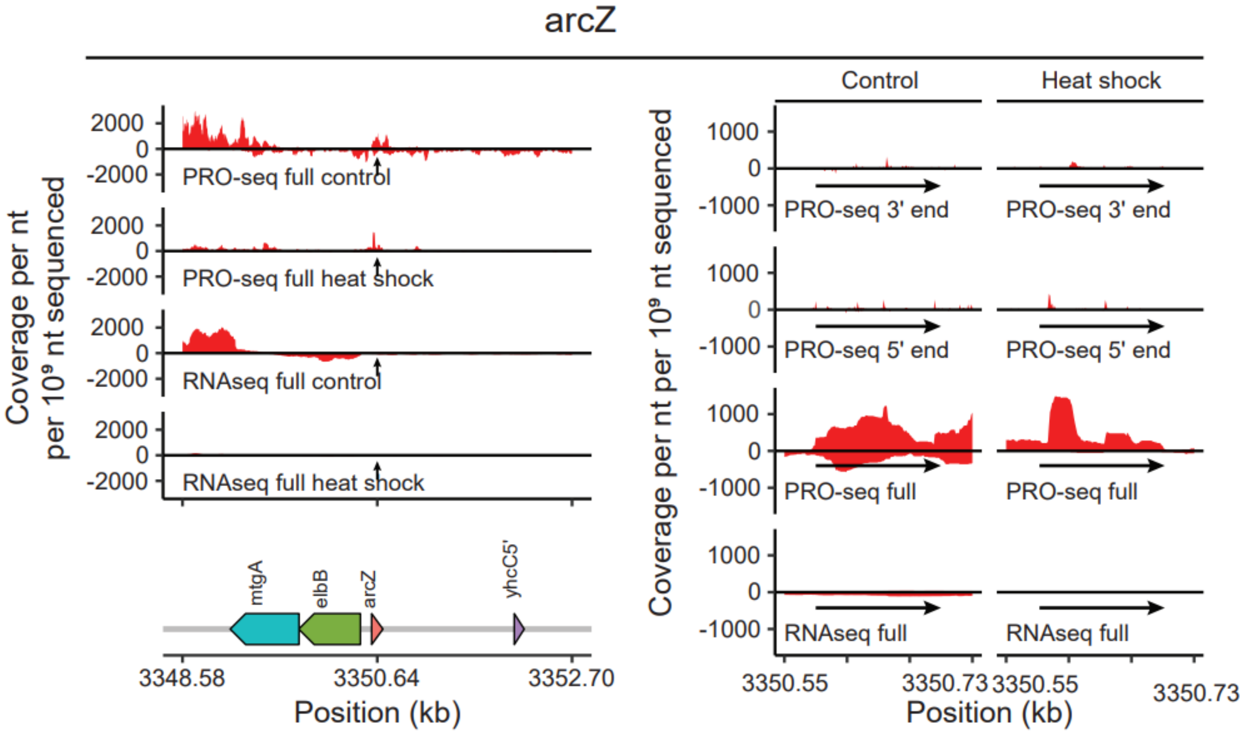
Figure 3.9. Logos for sequences surrounding PRO-seq read 3’ end peaks annotated for two Bacteroidetes, two Firmicutes and one Proteobacteria across two microbiome samples. Position -1 represents the RNAP pause site and position 1 represents the next nucleotide added.

Discussion

Bacterial nascent transcriptomics provides insights into co-transcriptional RNA processing and RNAP activity and localization. Rather than antibodies or epitopes fused to RNAP, PRO-seq leverages the universal function of RNAP to profile active transcription. We demonstrate its applicability to diverse species within microbiomes, showing that PRO-seq is capable of capturing nascent transcriptional dynamics without the need for cell culture. Our experiments using PRO-seq on heat-shocked *E. coli* highlight the potential for PRO-seq to decipher regulatory circuits operating under other environmental perturbations. Our observation of transcription of RNA species unique to the PRO-seq libraries in both cultured *E. coli* and gut microbiome samples (Figures 3.3, 3.4, 3.5, 3.6, 3.7, 3.8, 3.10 & 3.11) illustrates the utility of PRO-seq in identifying regulatory non-coding RNAs and co-transcriptionally processed RNA products. For bacteria in which non-coding RNAs have not yet been documented, PRO-seq offers a means to broadly survey these RNA molecules. In *E. coli*, where non-coding RNAs are well-annotated, we find that they are enriched in PRO-seq libraries compared to RNAseq libraries (Figures 3.10 & 3.11). PRO-seq data from metagenomes can therefore provide better guidance on the outputs of these genes than RNAseq data alone. Similarly, given the observation that PRO-seq can be used to broadly profile the transcription of tRNA isoforms and CRISPR loci across diverse species, we expect this method to shed light on the expression and processing of these molecules across conditions.

PRO-seq can also be combined with existing transcriptomic tools to examine transcriptional dynamics at a much finer scale than achievable with RNAseq alone. PRO-seq reveals RNAP positioning with base-pair resolution and also captures immature RNA cleavage products. We note that Cappable-seq, the only other nascent transcription method applied to microbiomes, is not suited to the identification of co-transcriptionally processed RNAs due to the need for intact 5’-PPP. However, PRO-seq can also be combined with Cappable-seq for paired analysis of transcription start sites and RNAP localization. PRO-seq may be further paired with NET-seq, albeit in genetically tractable organisms due to its reliance on the immunoprecipitation of RNAP, to discriminate nascent transcripts that are being actively polymerized from those in backtracked states [43]. Altogether, PRO-seq demonstrates that a larger fraction of bacterial genomes is actively transcribed than represented by traditional RNA sequencing, and that nascent transcription of microbiomes has potential, in concert with other -omics methods, to uncover co-transcriptional dynamics that provide functional insight into the gut microbial community.


spf
Coverage per nt
per 10⁹ nt sequenced
4000
0
-4000
PRO-seq full control
PRO-seq full heat shock
RNAseq full control
RNAseq full heat shock
spf
yihA
csrC
yihI
4047.90
4049.95
4052.01
Position (kb)
control
eat shock
Coverage per nt per 10⁹ nt sequenced
PRO-seq 3' end
PRO-seq 5' end
PRO-seq full
RNAseq full
4049.87
4050.04
position (kb)


arcZ
Coverage per nt
per 10⁹ nt sequenced
2000
0
-2000
PRO-seq full control
PRO-seq full heat shock
RNAseq full control
RNAseq full heat shock
mtgA
elbB
arcZ
yhcC5'
3348.58
3350.64
3352.70
Position (kb)
Control
Heat shock
Coverage per nt per 10⁹ nt sequenced
1000
0
-1000
PRO-seq 3' end
PRO-seq 5' end
PRO-seq full
RNAseq full
3350.55
3350.73
Position (kb)

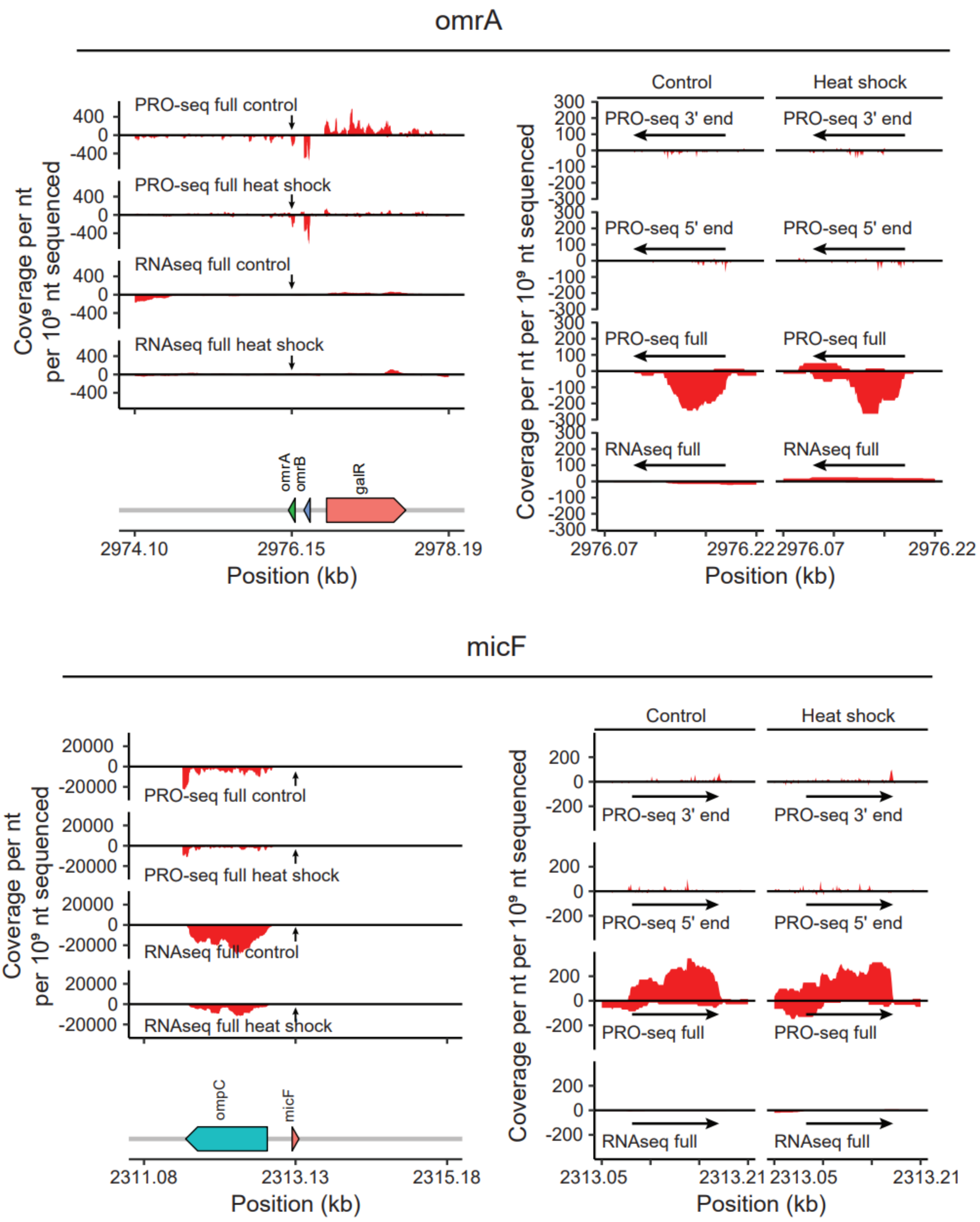


Figure 3.10. Selected *E. coli* small non-coding RNA (sRNA) loci shown with coverage (per nucleotide per 10^9 sequenced reads) surrounding each locus (left) in RNAseq libraries and PROseq libraries from under control and heat shock conditions. On the right, RNAseq coverage, composite PRO-seq read coverage, 5’ end and 3’ end coverage are shown for the specific portion of the locus encoding the sRNA.

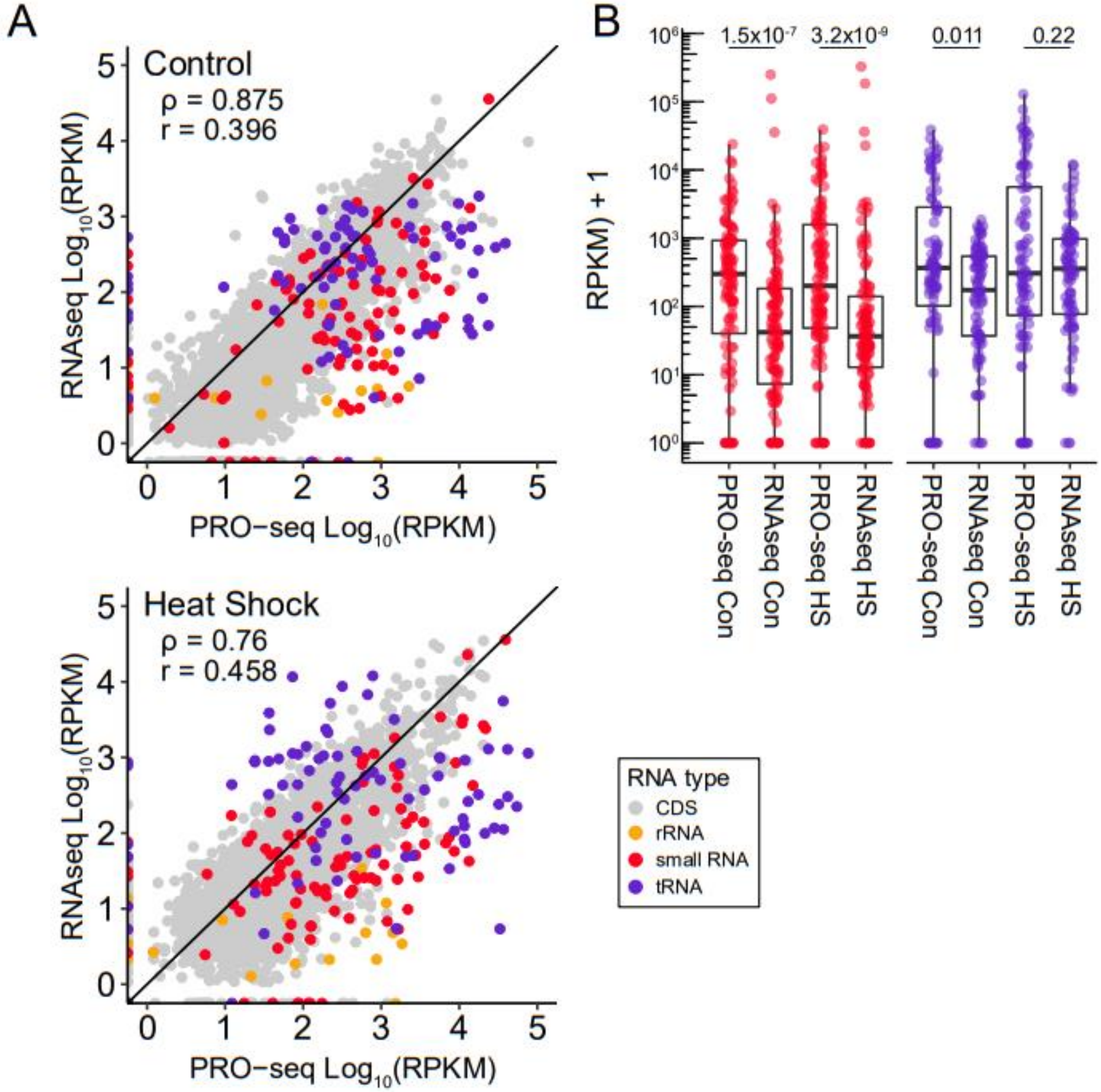


Figure 3.11. **(A)** Log-log RPKM plots comparing merged PRO-seq and RNAseq libraries for control and heat-shock conditions. Genes are colored by RNA type. Spearman's rank correlation coefficients (*ρ*) and Pearson's correlation coefficients (*r*) are inset. **(B)** Box plots show the RPKM distribution for small non-coding RNAs and tRNAs across control and heat-shock conditions. Black lines represent medians. P-values from Wilcoxon signed-rank tests are reported for each RNA type + treatment pair.

Materials and methods

E. coli heat-shock experiment

An overnight culture of *E. coli* MG1655 was subcultured in 50 mL LB and grown at 37°C to OD_{600} = 0.95. The culture was then split into 2 × 25 mL, with one half subjected to continued incubation at 37°C and the other half subjected to heat shock at 50 °C for 7 minutes, as described elsewhere [44]. Cultures were then split into 5 mL aliquots and pelleted by centrifugation at 3000 × g. At this point, pellets were either flash-frozen and stored at –80 °C for RNAseq or carried through permeabilization for PRO-seq.

Human gut microbiome sample collection

Freshly voided stool samples were collected and homogenized in an equal volume of cold, O_2-depleted phosphate-buffered saline, pH 7.2. Stool slurries were centrifuged to remove insoluble material (500 × g, 4 °C, 10 min.), then 12 mL of the liquid supernatant was layered over 3 mL 50% Nycodenz (Accurate Chemical) and centrifuged to concentrate cells (5000 × g, 4 °C, 20 min.). The cell-rich layer above the Nycodenz was collected and stored on ice; this was repeated until all stool homogenate was processed. 2 mL aliquots of each sample were stored at –80 °C for metagenome preparation and RNAseq. Four 600 µL aliquots of each sample were kept on ice until PRO-seq permeabilization. All individuals gave informed consent and all samples were collected under protocol #1609006585 approved by the Cornell University Institutional Review Board.

RNAseq sample preparation and sequencing

For each *E. coli* treatment, RNA was extracted from replicate samples using the RNeasy Mini Kit (Qiagen), including the optional β-mercaptoethanol treatment specified in the manufacturer’s protocol. On-column DNase I treatment was carried out using components of the

DNase Max Kit (Qiagen). RNA was eluted in at least 50 µL nuclease-free water and quantified with the Qubit RNA HS Assay Kit (Thermo Fisher). RNA was combined with 0.1× volume 3M sodium acetate, 3× volumes cold absolute ethanol, and 1 µL GlycoBlue Coprecipitant (Thermo Fisher) and allowed to precipitate at –80 °C for 30 minutes. RNA was pelleted by centrifugation (20,000 × g, 4 °C, 15 min.), washed with cold 70% ethanol, air-dried, and resuspended in nuclease-free water at a concentration of 91 ng/µL. One µg (11 µL) of each sample was subject to rRNA depletion using 2 µL NEBNext Bacterial rRNA Depletion Solution and 2 µL NEBNext Probe Hybridization Buffer (New England Biolabs). Sequencing libraries were prepared from both rRNA-depleted and whole RNA aliquots using the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (New England Biolabs) following the manufacturer's protocol for library preparation from intact RNA. RNAclean XP beads were substituted for AMPure XP beads supplemented with 10 U/mL SUPERase-In RNase Inhibitor (Thermo Fisher). Library concentrations were quantified by Qubit dsDNA HS Assay Kit (Thermo Fisher), and library size distributions were visualized by polyacrylamide gel electrophoresis.

For each Nycodenz-purified stool cell sample, RNA was extracted using the RNeasy PowerMicrobiome Kit (Qiagen), following the manufacturer's protocol to increase the representation of small RNAs. Total RNA was eluted from columns with 100 µL nuclease-free water and quantified using the Qubit RNA BR Assay Kit (Thermo Fisher). Duplicate 1 µg aliquots were subject to RNA fragmentation and rRNA depletion using the QIAseq FastSelect – 5S/16S/23S Kit (Qiagen), assuming a RNA integrity number ≥ 8 for all samples. Sequencing libraries were prepared from rRNA-depleted samples as described for *E. coli*. Sequencing platforms and number of reads for each replicate are listed in Table 2.

Table 2. List of samples sequenced in this project

Sample name	**Replicate**	**Sequencing platform**	**Number of clean paired-end reads**
E. coli – control – RNAseq	1	NextSeq, 2 × 150	1,713,282
	2	NextSeq, 2 × 150	1,871,542
E. coli – control – RNAseq with rRNA depletion	1	NextSeq, 2 × 150	1,834,682
	2	NextSeq, 2 × 150	1,456,787
E. coli – control – PRO-seq	1	NextSeq, 2 × 75	3,982,364
	2	NextSeq, 2 × 75	1,615,517
	3	NextSeq, 2 × 75	6,026,982
E. coli – heat shock – RNAseq	1	NextSeq, 2 × 150	1,533,500
	2	NextSeq, 2 × 150	1,790,382
E. coli – heat shock – RNAseq with rRNA depletion	1	NextSeq, 2 × 150	1,290,188
	2	NextSeq, 2 × 150	1,442,903
E. coli – heat shock – PRO-seq	1	NextSeq, 2 × 75	4,857,572
	2	NextSeq, 2 × 75	5,730,267
	3	NextSeq, 2 × 75	2,703,945
US2 – RNAseq with rRNA depletion	1	NextSeq, 2 × 150	23,115,457
	2	NextSeq, 2 × 150	19,572,778
US2 – PRO-seq	1	HiSeq X, 2 × 150	120,441,062
	2	HiSeq X, 2 × 150	89,204,805
US3 – RNAseq with rRNA depletion	1	NextSeq, 2 × 150	21,299,806
	2	NextSeq, 2 × 150	21,617,152
US3 – PRO-seq	1	HiSeq X, 2 × 150	125,455,387
	2	HiSeq X, 2 × 150	117,270,406
US2 metagenome	n/a	NextSeq, 2 × 150	15,572,009
US3 metagenome	n/a	NextSeq, 2 × 150	16,747,211

PRO-seq sample preparation and sequencing

For each *E. coli* treatment, the cell pellets described above were resuspended in 1.5 mL cold cell permeabilization buffer (10 mM Tris-HCl, pH 7.4, 300 mM sucrose, 10 mM KCl, 5 mM $MgCl_2$, 1 mM EGTA, 0.05% v/v Tween-20, 0.1% v/v IGEPAL CA-630, 0.1% v/v Triton X-100, 0.5 mM DTT, 1× Roche cOmplete Protease Inhibitor Cocktail (Sigma-Aldrich), and 20 U/mL SUPERase-In RNase Inhibitor (Thermo Fisher); modified from Mahat *et al.* [8]) and incubated on ice for 5 minutes. Pelleting, resuspension in permeabilization buffer, and incubation was repeated for a total of 3 permeabilization washes. Cell lysates were then pelleted by centrifugation (10,000 × g, 4 °C, 5 min.), resuspended in 250 µL storage buffer (10 mM Tris-

HCl, pH 8.0, 25% v/v glycerol, 5 mM $MgCl_2$, 0.1 mM EDTA, and 5 mM DTT), split into 5 × 50 μL aliquots, flash-frozen on dry ice / ethanol, and stored at –80 °C until run-on. Final cell concentrations inferred from plating pre-permeabilization cell suspensions were 2.5×10^{10} and 5.0×10^{10} CFU/mL for control and heat-shocked samples, respectively.

To improve the permeabilization of Gram-positive organisms, 1000 U of Ready-Lyse Lysozyme Solution (Lucigen) was added to each 600 μL Nycodenz-purified stool cell aliquot and incubated for 10 minutes on ice. Then, cell suspensions were transferred to 2 mL screw-cap tubes and combined with 400 μL sterile 0.5 mm glass beads and 1 mL cold cell permeabilization buffer. Cells were pulverized by vortexing for 3 cycles of 2 minutes at max Hz followed by 2 minutes on ice. Lysates were stored upright on ice for 10 minutes to allow beads to settle, then 1 mL supernatant from each tube was transferred to a 1.5 mL tube and centrifuged to collect cell contents (10,000 × g, 4 °C, 5 min.). Pellets were washed once with 1 mL cold storage buffer, pelleted again, and resuspended in 200 μL cold storage buffer. Lysates were flash-frozen and stored as described for *E. coli*.

For all samples, PRO-seq was carried out following the “4-Biotin run-on” variant of the protocol described in Mahat *et al* [8]. Briefly, permeabilized cells were thawed on ice and run-on reactions were carried out at 37 °C using a master mix containing Biotin-11-ATP, Biotin-11-CTP, Biotin-11-GTP, and Biotin-11-UTP. Total RNA was extracted by TRIzol and ethanol precipitation, and RNA was fragmented by NaOH hydrolysis. 3’ adapters were ligated, then biotinylated transcripts were enriched and washed with hydrophilic streptavidin magnetic beads. 5’ de-capping and phosphorylation were carried out with nascent transcripts bound to the beads, then RNA was eluted from the beads by TRIzol extraction and ethanol precipitation. 5’ adapters with unique molecular identifiers were ligated to nascent transcripts, and excess adapters were removed by again capturing biotinylated RNA on streptavidin beads, washing the beads, and re-extracting RNA with TRIzol and ethanol precipitation. Nascent RNA was reverse transcribed, and cDNA was quantified by qPCR to determine the appropriate number of cycles for PCR amplification. Library amplification was carried out using custom PCR primers to incorporate Illumina adapter sequences and i7 barcodes. PCRs were cleaned up with Exonuclease I and Shrimp Alkaline Phosphatase. DNA concentration was quantified with the Qubit dsDNA HS Assay Kit, and library quality was assessed by polyacrylamide gel electrophoresis. Sequencing platforms and number of reads for each replicate are listed in Supplemental Table 1.

Metagenomic library preparation, sequencing, assembly, binning, and annotation

To prepare metagenomes against which to map transcriptomics reads, DNA was isolated from 250 μL of Nycodenz-purified stool cells using the DNeasy PowerSoil Kit (Qiagen). DNA was eluted in 100 μL warm 0.1× TE, quantified by Qubit dsDNA BR Assay, and diluted to 0.2 ng/μL for input to the Nextera XT DNA Library Preparation Kit (Illumina). Sequencing libraries were prepared from 1 ng fecal DNA following the manufacturer’s protocol, and libraries were cleaned up using 1.5× volumes of AMPure XP beads. Library concentration was quantified by Qubit dsDNA HS Assay, and fragment size distribution was visualized by 8% polyacrylamide gel electrophoresis.

Metagenomes were sequenced as referenced in Table 2. Raw reads were processed with PRINSEQ lite v0.20.4 [45] and trimmomatic v0.36 [46] to remove duplicates and sequencing adapters. Reads mapping to the human genome were discarded using BMTagger [47]. Clean reads were assembled using SPAdes v3.14.0 [48] (paired-end mode and --meta option) and reads were aligned to contigs using BWA-MEM v0.7.17 [49,50]. Contigs were binned using CONCOCT v1.1.0 [51], metaBAT v2.12.1 [52], and MaxBin v2.2.4 [53], then bins from different programs were resolved into metagenome-assembled genomes (MAGs) using DAS Tool v1.1.2 [54] with DIAMOND v2.0.4 [55] for single copy gene identification. The completeness and contamination of MAGs was assessed with CheckM v1.1.2 [56] and taxonomic classifications were assigned to MAGs using GTDB-Tk v1.0.2 [57]. MAG features were annotated using prokka v1.14.5 [58] (--metagenome, --rfam), which uses Prodigal [59], ARAGORN [60], barrnap [61], and Infernal [62] for identification of protein-coding sequences, tRNAs, rRNAs, and ncRNAs, respectively.

Transcriptomics data processing and analysis

PRO-seq reads were processed with proseq2.0.bsh (https://github.com/Danko-Lab/proseq2.0) to trim by quality, remove adapter sequences, and remove duplicates by their unique molecular identifiers (UMIs). RNAseq reads were similarly processed, but without UMI deduplication. Cleaned paired-end reads were aligned to their respective references using BWA: metatranscriptome reads were aligned to the assemblies described above; *E. coli* reads were aligned to the GenBank Reference Sequence for *E. coli* K12, version NC_000913.3 [63,64]. BAM files were filtered with SAMtools v1.11 [65] to include only paired reads in proper pairs with a minimum MAPQ score of 30 (-f 3 -q 30) and exclude all unmapped or non-primary alignments (-F 2316). Reads were assigned to features using the featureCounts function from subread v2.0.2 [66]. *E. coli* protein-coding genes and regulatory loci were identified using the regutools R package [67] and RegulonDB v10.9 [68]. The genomecov function from BEDTools v2.29.2 [69] was used to report strand-specific PRO-seq and RNAseq depth at each position in the metagenome (-ibam -d -pc -strand). For PRO-seq, metagenomic depth profiles from 3’ and 5’ fragment ends were additionally reported as follows: since the P5 Illumina adapter is ligated to the 3’ end of the nascent transcript, the 5’ end of the first read in each pair gives the 3’ end of the nascent transcript on the opposite strand (samtools view -f 64 -b $bam | bedtools genomecov -5 -d -

strand - > plus_3p.txt); likewise, the 5' end of the second read in each pair gives the 5' end of the nascent transcript on the same strand, since proper pairs align to opposite strands (samtools view -f 128 -b $bam | bedtools genomecov -5 -d -strand + > plus_5p.txt).

Read depth profiles at regions of interest were visualized with ggplot2 [70] using custom R code available at https://github.com/britolab/PRO-seq. CRISPR repeats were detected using MinCED [71], which is derived from CRISPR Recognition Tool [72]. CRISPR RNA and tRNA secondary structures were predicted with the ViennaRNA secondary structure server [73] and visualized with forna [74]. Pearson's correlation coefficients (r) and Spearman's rank correlation coefficients (ρ) were calculated for correlation plots using the stats package from base R [75]. Wilcoxon signed-rank tests were performed using the ggpubr package v0.4.0 [76].

Peaks were called from 3' end depth data by first filtering all positions for a minimum depth of 10 reads. Then, the mean coverage over a ±25 nt interval surrounding each position was calculated, and Z scores were determined for each peak centered in its interval. Positions with Z scores of at least 5 were kept, and sequences surrounding those peaks were pulled to create sequence logos with the ggseqlogo package [77].

Data and code availability

Scripts are available at: https://github.com/britolab/PRO-seq. Sequencing data has been uploaded to NCBI's Sequence Read Archive and is associated with BioProjects PRJNA800038 and PRJNA800070.

References

1. Wissink, E. M., Vihervaara, A., Tippens, N. D. & Lis, J. T. Nascent RNA analyses: tracking transcription and its regulation. *Nat. Rev. Genet.* **20**, 705–723 (2019).
2. Larson, M. H. *et al.* A pause sequence enriched at translation start sites drives transcription dynamics in vivo. *Science (80-.).* **344**, 1042–1047 (2014).
3. Imashimizu, M. *et al.* Visualizing translocation dynamics and nascent transcript errors in paused RNA polymerases in vivo. *Genome Biol.* **16**, 1–17 (2015).

4. Sharma, C. M. *et al.* The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature* **464**, 250–255 (2010).
5. Thomason, M. K. *et al.* Global transcriptional start site mapping using differential RNA sequencing reveals novel antisense RNAs in Escherichia coli. *J. Bacteriol.* **197**, 18–28 (2015).
6. Sharma, C. M. & Vogel, J. Differential RNA-seq: The approach behind and the biological insight gained. *Curr. Opin. Microbiol.* **19**, 97–105 (2014).
7. Ettwiller, L., Buswell, J., Yigit, E. & Schildkraut, I. A novel enrichment strategy reveals unprecedented number of novel transcription start sites at single base resolution in a model prokaryote and the gut microbiome. *BMC Genomics* **17**, 1–14 (2016).
8. Mahat, D. B. *et al.* Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat. Protoc.* **11**, 1455–1476 (2016).
9. Blumberg, A. *et al.* Characterizing RNA stability genome-wide through combined analysis of PRO-seq and RNA-seq data. *BMC Biol.* 1–17 (2021). doi:10.1101/690644
10. Patel, R. K., West, J. D., Jiang, Y., Fogarty, E. A. & Grimson, A. Robust partitioning of microRNA targets from downstream regulatory changes. *Nucleic Acids Res.* **48**, 9724–9746 (2020).
11. Mentesana, P. E., Chin-Bow, S. T., Sousa, R. & McAllister, W. T. Characterization of halted T7 RNA polymerase elongation complexes reveals multiple factors that contribute to stability. *J. Mol. Biol.* **302**, 1049–1062 (2000).
12. Blumberg, A., Rice, E. J., Kundaje, A., Danko, C. G. & Mishmar, D. Initiation of mtDNA transcription is followed by pausing, and diverges across human cell types and during evolution. *Genome Res.* **27**, 362–373 (2017).
13. Zhang, J., Cavallaro, M. & Hebenstreit, D. Timing RNA polymerase pausing with TV-PRO-seq. *Cell Reports Methods* **1**, 100083 (2021).
14. Alberti, A. *et al.* Comparison of library preparation methods reveals their impact on interpretation of metatranscriptomic data. *BMC Genomics* **15**, 1–13 (2014).
15. Wesolowska-Andersen, A. *et al.* Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. *Microbiome* **2**, 19 (2014).
16. Teng, F. *et al.* Impact of DNA extraction method and targeted 16S-rRNA hypervariable region on oral microbiota profiling. *Sci. Rep.* **8**, 1–12 (2018).
17. Liu, X. & Martin, C. T. Transcription elongation complex stability: The topological lock. *J. Biol. Chem.* **284**, 36262–36270 (2009).
18. Croucher, N. J. & Thomson, N. R. Studying bacterial transcriptomes using RNA-seq. *Curr. Opin. Microbiol.* **13**, 619–624 (2010).
19. Yuzhen, Y. E. & Quan, Z. Characterization of CRISPR RNA transcription by exploiting stranded metatranscriptomic data. *Rna* **22**, 945–956 (2016).
20. Barrangou, R. *et al.* CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science (80-.).* **315**, 1709–1712 (2007).
21. Juranek, S. *et al.* A genome-wide view of the expression and processing patterns of Thermus thermophilus HB8 CRISPR RNAs. *Rna* **18**, 783–794 (2012).
22. Lillestøl, R. K. *et al.* CRISPR families of the crenarchaeal genus Sulfolobus: Bidirectional transcription and dynamic properties. *Mol. Microbiol.* **72**, 259–272 (2009).
23. Richter, H. *et al.* Characterization of CRISPR RNA processing in Clostridium thermocellum and Methanococcus maripaludis. *Nucleic Acids Res.* **40**, 9887–9896 (2012).
24. Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602–607 (2011).
25. Xu, H., Yao, J., Wu, D. C. & Lambowitz, A. M. Improved TGIRT-seq methods for comprehensive transcriptome profiling with decreased adapter dimer formation and bias correction. *Sci. Rep.* **9**, 1–17 (2019).
26. Boivin, V. *et al.* Reducing the structure bias of RNA-Seq reveals a large number of non-annotated non-coding RNA. *Nucleic Acids Res.* **48**, 2271–2286 (2020).

27. Marbaniang, C. N. & Vogel, J. Emerging roles of RNA modifications in bacteria. *Curr. Opin. Microbiol.* **30**, 50–57 (2016).

28. de Crécy-Lagard, V. & Jaroch, M. Functions of Bacterial tRNA Modifications: From Ubiquity to Diversity. *Trends Microbiol.* **29**, 41–53 (2021).

29. Antoine, L. *et al.* Rna modifications in pathogenic bacteria: Impact on host adaptation and virulence. *Genes (Basel).* **12**, (2021).

30. Li, Z. & Stanton, B. A. Transfer RNA-Derived Fragments, the Underappreciated Regulatory Small RNAs in Microbial Pathogenesis. *Front. Microbiol.* **12**, (2021).

31. Haiser, H. J., Karginov, F. V., Hannon, G. J. & Elliot, M. A. Developmentally regulated cleavage of tRNAs in the bacterium Streptomyces coelicolor. *Nucleic Acids Res.* **36**, 732–741 (2008).

32. Houserova, D. & Yulong Huang, Mohan V. Kasukurthi2, Brianna C. Watters1, 3, Fiza F. Khan1, Raj V. Mehta1, Neil Y. Chaudhary1, Justin T. Roberts1, 4, Jeffrey D. DeMeis1, Trevor K. Hobbs1, Kanesha R. Ghee1, 3, Cameron H. McInnis1, 3, Nolan P. Johns1, 3. Salmonella Outer Membrane Vesicles contain tRNA Fragments (tRFs) that Inhibit Bacteriophage P22 infection. *bioRxiv* (2021).

33. Schwartz, M. H. *et al.* Microbiome characterization by high-throughput transfer RNA sequencing and modification analysis. *Nat. Commun.* **9**, (2018).

34. Kimura, S., Srisuknimit, V. & Waldor, M. K. Probing the diversity and regulation of tRNA modifications. *Curr. Opin. Microbiol.* **57**, 41–48 (2020).

35. Zhang, K., Hodge, J., Chatterjee, A., Moon, T. S. & Parker, K. M. Duplex structure of double-stranded RNA provides stability against hydrolysis relative to single-stranded RNA. *Environ. Sci. Technol.* **55**, 8045–8053 (2021).

36. Ovcharenko, A. & Rentmeister, A. Emerging approaches for detection of methylation sites in RNA. *Open Biol.* **8**, (2018).

37. Hauenschild, R. *et al.* The reverse transcription signature of N-1-methyladenosine in RNA-Seq is sequence dependent. *Nucleic Acids Res.* **43**, 9950–9964 (2015).

38. Boccaletto, P. *et al.* MODOMICS: A database of RNA modification pathways. 2017 update. *Nucleic Acids Res.* **46**, D303–D307 (2018).

39. Björk, G. R., Wikström, P. M. & Byström, A. S. Prevention of translational frameshifting by the modified nucleoside 1-methylguanosine. *Science (80-.).* **244**, 986–989 (1989).

40. Belogurov, G. A. & Artsimovitch, I. Regulation of Transcript Elongation. *Annu. Rev. Microbiol.* **69**, 49–69 (2015).

41. Henderson, K. L. *et al.* Mechanism of transcription initiation and promoter escape by E. coli RNA polymerase. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E3032–E3040 (2017).

42. Vvedenskaya, I. O. *et al.* Interactions between RNA polymerase and the "core recognition element" counteract pausing Irina. *Science (80-.).* **344**, 1285–1289 (2014).

43. Sun, Z., Yakhnin, A. V., FitzGerald, P. C., Mclntosh, C. E. & Kashlev, M. Nascent RNA sequencing identifies a widespread sigma70-dependent pausing regulated by Gre factors in bacteria. *Nat. Commun.* **12**, 1–14 (2021).

44. Chuang, S. E. & Blattner, F. R. Characterization of twenty-six new heat shock genes of Escherichia coli. *J. Bacteriol.* **175**, 5242–5252 (1993).

45. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).

46. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

47. Rotmistrovsky, K. & Agarwala, R. BMTagger: best match tagger for removing human reads from metagenomics datasets.

48. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. MetaSPAdes: A new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).

49. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

50. Li, H. *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. (2013).
51. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
52. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
53. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
54. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).
55. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
56. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from. *Cold Spring Harb. Lab. Press Method* 1–31 (2015). doi:10.1101/gr.186072.114
57. Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* **36**, 1925–1927 (2020).
58. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
59. Hyatt, D. *et al.* Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, (2010).
60. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* **32**, 11–16 (2004).
61. Seemann, T. barrnap 0.9: rapid ribosomal RNA prediction.
62. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
63. Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* **44**, D67–D72 (2016).
64. Freddolino, P. L., Amini, S. & Tavazoie, S. Newly identified genetic variations in common Escherichia coli MG1655 stock cultures. *J. Bacteriol.* **194**, 303–306 (2012).
65. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
66. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
67. Chávez, J. *et al.* Programmatic access to bacterial regulatory networks with regutools. *Bioinformatics* **36**, 4532–4534 (2020).
68. Santos-Zavaleta, A. *et al.* RegulonDB v 10.5: Tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. *Nucleic Acids Res.* **47**, D212–D220 (2019).
69. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
70. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis.* (Springer-Verlag New York., 2016).
71. Minced: Mining CRISPRs in Environmental Datasets.
72. Bland, C. *et al.* CRISPR Recognition Tool (CRT): A tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**, 1–8 (2007).
73. Hofacker, I. L. Vienna RNA secondary structure server. *Nucleic Acids Res.* **31**, 3429–3431 (2003).
74. Kerpedjiev, P., Hammer, S. & Hofacker, I. L. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics* **31**, 3377–3379 (2015).
75. R Core Team. R: a language and environment for statistical computing. (2014).
76. Kassambara, A. ggpubr: 'ggplot2' Based Publication Ready Plots. (2020).
77. Wagih, O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **33**, 3645–3647 (2017).

CHAPTER 4: Concluding remarks, and the future of meta'omics

In this dissertation, I've touched on two major aspects of the biology of microbiomes: genetic composition and transcription. Shotgun metagenomics involves the assembly of short reads along two dimensions: *contiguity* and *coverage*. Hi-C adds a third dimension, *proximity*, by which reads can be linked to contigs, though there are additional data types that can be leveraged to improve metagenomic assembly and the association of mobile elements with bacterial hosts. Long-read sequencing, besides greatly increasing the contiguity of assemblies[1], allows exploration of DNA modifications through detection of differentially methylated motifs[2]. These motifs can be leveraged to associate mobile elements with their hosts[3], given that chromosomal DNA and plasmids are exposed to the same methyltransferases. Equally promising for microbiome science is single-cell metagenomics, which has so far been applied to uncover sub-strain genomic variation that is typically collapsed into single MAGs during bulk sequencing and assembly[4,5].

On the gene expression front, much work still needs to be done to assess interpersonal and strain-level variability in transcription within the human microbiome. Though transcription regulation in prokaryotes is simpler than in eukaryotes by a number-of-components metric, there is an immense diversity of regulatory mechanisms within human-associated bacteria, some of which control expression in specific host niches[6,7] or respond to antibiotic-induced stress[8]. As with metagenomics, metatranscriptomics can greatly benefit from single-cell techniques. Currently, microSPLiT[9] and PETRI-seq[10] have both been described for prokaryotic cells, each circumventing the need to segregate single microbes by the clever application of combinatorial indexing. Though either technique has yet to be applied to uncultured cells, the road has been paved for a revolution in microbiome RNA sequencing. Another method worth mentioning is

MetaRibo-Seq[11], which involves the isolation of intact ribosomes from microbiome-derived bacteria and the sequencing of mRNAs engaged in translation. Where PRO-seq sheds light on the birth of transcripts, MetaRibo-Seq reveals their destiny, and, in doing so, spans the gap between nascent transcriptomics and metaproteomics[12,13].

Microbiomes are messy, convoluted things. Bioinformaticians are working diligently to address the compositional nature of microbiomes and make the best use of existing data, though computational innovation must be paired with creativity at the bench to glean true biological insights. An immense amount of work has already been done to characterize the interplay between commensal bacteria and their hosts, though there is much left to discover, and it will require the synthesis of data at every level – DNA, RNA, and protein – before we can approach a comprehensive understanding of the microbes that inhabit us.

References

1. Bickhart,D.M., Kolmogorov,M., Tseng,E., Portik,D.M., Korobeynikov,A., Tolstoganov,I., Uritskiy,G., Liachko,I., Sullivan,S.T., Shin,S.B., *et al.* (2022) Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. *Nat. Biotechnol.*, 10.1038/s41587-021-01130-z.
2. Tourancheau,A., Mead,E.A., Zhang,X.S. and Fang,G. (2021) Discovering multiple types of DNA methylation from bacteria and microbiome using nanopore sequencing. *Nat. Methods*, **18**, 491–498.
3. Beaulaurier,J., Zhu,S., Deikus,G., Mogno,I., Zhang,X.S., Davis-Richardson,A., Canepa,R., Triplett,E.W., Faith,J.J., Sebra,R., *et al.* (2018) Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nat. Biotechnol.*, **36**, 61–69.
4. Chijiiwa,R., Hosokawa,M., Kogawa,M., Nishikawa,Y., Ide,K., Sakanashi,C., Takahashi,K. and Takeyama,H. (2020) Single-cell genomics of uncultured bacteria reveals dietary fiber responders in the mouse gut microbiota. *Microbiome*, **8**, 1–14.
5. Arikawa,K., Ide,K., Kogawa,M., Saeki,T., Yoda,T., Endoh,T., Matsuhashi,A., Takeyama,H. and Hosokawa,M. (2021) Recovery of strain-resolved genomes from human microbiome through an integration framework of single-cell genomics and metagenomics. *Microbiome*, **9**, 1–16.
6. Rabinovich,L., Sigal,N., Borovok,I., Nir-Paz,R. and Herskovits,A.A. (2012) Prophage excision activates listeria competence genes that promote phagosomal escape and virulence. *Cell*, **150**, 792–802.

7. Pasechnek,A., Rabinovich,L., Stadnyuk,O., Azulay,G., Mioduser,J., Argov,T., Borovok,I., Sigal,N. and Herskovits,A.A. (2020) Active Lysogeny in Listeria Monocytogenes Is a Bacteria-Phage Adaptive Response in the Mammalian Environment. *Cell Rep.*, **32**, 107956.
8. Jiang,X., Hall,A.B., Arthur,T.D., Plichta,D.R., Covington,C.T., Poyet,M., Crothers,J., Moses,P.L., Tolonen,A.C., Vlamakis,H., *et al.* (2019) Invertible promoters mediate bacterial phase variation, antibiotic resistance, and host adaptation in the gut. *Science (80-.).*, **187**, 181–187.
9. Kuchina,A., Brettner,L.M., Paleologu,L., Roco,C.M., Rosenberg,A.B., Carignano,A., Kibler,R., Hirano,M., DePaolo,R.W. and Seelig,G. (2021) Microbial single-cell RNA sequencing by split-pool barcoding. *Science (80-.).*, **371**.
10. Blattman,S.B., Jiang,W., Oikonomou,P. and Tavazoie,S. (2020) Prokaryotic single-cell RNA sequencing by in situ combinatorial indexing. *Nat. Microbiol.*, **5**, 1192–1201.
11. Fremin,B.J. and Bhatt,A.S. (2018) Metagenome-wide measurement of protein synthesis in the human fecal microbiota using MetaRibo-Seq. *bioRxiv*, 10.1101/482430.
12. Kleiner,M. (2019) Metaproteomics: Much More than Measuring Gene Expression in Microbial Communities. *mSystems*, **4**.
13. Van Den Bossche,T., Arntzen,M., Becher,D., Benndorf,D., Eijsink,V.G.H., Henry,C., Jagtap,P.D., Jehmlich,N., Juste,C., Kunath,B.J., *et al.* (2021) The Metaproteomics Initiative: a coordinated approach for propelling the functional characterization of microbiomes. *Microbiome*, **9**, 1–4.