TRANSMISSION OF THE HUMAN MICROBIOME: FROM INFANTS TO INFECTIONS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF GENETICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Benjamin Aram Siranosian
November 2021

Abstract

The human body is colonized by trillions of microorganisms, the vast majority of which reside in the large intestine. This collection of bacteria, fungi, archaea and viruses is collectively called the gut microbiome. Large enough to be considered its own organ, the gut microbiome has vast impacts on every day human function, including digestion [64], the immune system [180], even the brain and a host's mood [109]. New research also suggests the gut microbiome can modulate an individual's response to anti-cancer immunotherapy [15, 35].

The anaerobic, nutrient rich environment of the mammalian intestine provides bacterial populations with everything needed to grow, proliferate and differentiate. While certain bacterial species that inhabit the gut microbiome may live within the food we eat, many species are uniquely adapted to the human intestine and do not live in other environments. The following conclusions logically follow:

1. Bacteria uniquely adapted to the human gut must be transmitted between individuals.
2. The human body has the ability to acquire new microbes from the environment and other individuals. Proper functioning of the gut microbiome is required for health, therefore this process is beneficial to the human body.
3. The gut microbiome is stable over time in most individuals, therefore selection takes place for which microbes engraft in the gut long-term.

However, these conclusions raise many more questions about the acquisition and transmission of gut microbes:

1. At what age do humans begin to acquire and develop a microbiome? At what age does acquisition of new microbes slow down or stop?
2. Where do the new microbes colonizing an infant come from?
3. How are healthy, commensal microbes selected from pathogens?
4. Does transmission between individuals also spread fungi, archaea and phages?

5. Does transmission occur at the community level, or is it a single species at a time?

Throughout my graduate work in Dr. Ami Bhatt's laboratory at Stanford University, I have attempted to shed light on a small portion of the microbiome transmission problem. Using advanced metagenomic sequencing techniques, I have tracked bacteria and phages down to the strain level to understand if and how they pass between individuals. This research has been limited to transmission between humans in non-experimental settings - no specific interventions or animal models were used. In these observational studies, I desired to capture acquisition and transmission events as they happened naturally by using the precise archaeological record stored in an individual's microbiome. I hope that my modest contributions will advance the field, improve understanding of how humans acquire and transmit members of their gut microbiota and provide stepping stones for future research to expand upon my findings.

This thesis consists of 6 chapters:

1. An introduction explaining the overarching themes of acquisition and transmission in the human microbiome, and specifics that are not covered in the introduction of individual chapters.

2. "Intestinal microbiota domination under extreme selective pressures characterized by metagenomic read cloud sequencing and assembly," my first manuscript using advanced metagenomic techniques to measure the microbiome of Hematopoietic Cell Transplantation (HCT) patients as they experienced treatment with antibiotics.

3. "Acquisition, transmission and strain diversity of human gut-colonizing crAss-like phages," where I measured infant acquisition of crAss-like phages and showed that direct transmission from the mother was likely responsible for colonization in 50% of infants.

4. "Rare transmission of commensal and pathogenic bacteria in the gut microbiome of hospitalized adults," my final effort to measure transmission of bacteria between HCT patients who were roommates in the hospital. Here, we provided high quality, time course resolved genomic evidence for transmission of bacteria between the gut microbiome of adults, a first in the academic literature.

5. "Building bioinformatics workflows for scalability and reproducibility," some helpful tips and methodologies I learned for conducting large-scale computational experiments and ensuring that the workflows are scalable and reproducible.

6. Future directions of my work, speculation on the results, and conclusions.

Acknowledgments

This thesis is not remotely the work of one individual. Besides the official lists of authors in each chapter, I would like to extend a personal note of thanks to the following individuals:

Ami Bhatt, whose mentorship style drew me into the lab initially, and whose endless energy for research was an inspiration. It's been exciting to watch the lab grow and our research directions grow and mature over the years.

Members of my thesis committee: Arend Sidow, Benjamin Good and Hua Tang, who provided mentorship when the problems were hard, encouragement when it was needed and praise when it was deserved.

My past scientific mentors, including Uri Ben-David and Aarvind Subramanian at the Broad Institute, Nicola Neretti, Peter Shank, Sorin Istrail, Steven Criscione and Sarah Taylor at Brown, and Ms. McLoughlin and Mr. Gilmore at Sandwich High School.

Previous members of the Bhatt lab, who shaped the welcoming community and culture of scientific rigor we have today, including: Eli Moss, who provided mentorship on all things lab, career and automotive, Fiona Tamburini, who's endless energy inspired me to work hard on our collaborative projects, Jessica Ribado, who always found the best way to investigate or plot data, Brayon Fremin, who was convinced no problem wouldn't yield to his effort, Matt Durrant, who taught me new ways to think about what matters in microbiome data, Tessa Andermann, who's dedication to patients and the lab's biobank was an inspiration, and Summer Vance, who kept a tight ship and I enjoyed sharing Jasper Ridge patrols with outside of lab.

Current members of the Bhatt lab, who I will definitely miss but I'm excited to watch for their publications, including Dylan Maghini, Alvin "HotMice" Han, Soumaya Zlitni, Aravind Natarajan, Chris Severyn, Aaron Behr, Erin brooks, Sierra Bowden, Ann Lin and others.

Members of my genetics cohort. We've been through a lot the past few years in California, wildfires and a pandemic to name just a few. I've especially enjoyed our weekly DnD crew of Adam, Jennifer, Kelsey, Alex, Vy and Akshay. Plus the friendship of Nicole and Becca.

My housemates Noah Youkilis, Nick Bousse and Cady van Assendelft who I rode bikes, traveled, brewed beer, roasted coffee, and generally enjoyed the company of for most of 2020 and 2021. Members of the Stanford Cycling Team, too many to count, who I've been on casual rides and

diehard races with over the years, and who I've consumed thousands and thousands of calories of pastries with.

Other members of the Stanford community who I collaborated with, looked up to, or learned from, including Matt Olm, Alex Colville, Egan Peltan, Vandon Duong and Matt Buckley.

I'll be working at Cellular Longevity Inc. (Loyal) after I defend, a biotech startup aiming to increase canine lifespan and healthspan. I've already enjoyed working with Celine Halioua, Tom Roseberry, Alex Naka, Jeff Rafter and Michael Lampe, and I'm excited to see what we can accomplish. At Imago Biosciences, I've enjoyed working with Hugh Rienhoff and and Georges Natsoulis, who provided mentorship on the move into biotech.

My parents, Ed and Kathy Siranosian, who nurtured the scientific spirit in me from a young age, even if that meant putting up with me taking apart family electronics or helping me clean the mess of a science experiment gone wrong. I attribute so much of my biological interest to my mother and my quantitative interest to my father. My sisters Jen and Liz have kept me grounded and have reminded me of the importance of family.

Last on this list and first in my mind is Kelsey Grabarek. She has provided endless support during the highs and lows of the PhD endeavour, and she truly understands that a fed grad student is a productive grad student. I'm so lucky she decided to take a chance on a nerdy grad student, and I'm always looking forward to what we'll do next.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 The developing human microbiome and strain acquisition

The rate of change in an individual's microbiome is highest in the first few years of life [125]. Genomic characterization of paired mother and infant stool samples has shown that infants acquire many of their first gut microbes via direct transmission from their mother [195, 13]. Mother-infant transmission is important for seeding certain key microbes like *Bifidobacterium infantis*, which contains the unique ability to digest human milk oligosaccharides found in breast milk [103, 176]. Mother-infant transmission is important for the normal development of an infant's microbiome and healthy infant development as well.

After the rapid period of microbiome modification ceases in adolescence, humans may still rarely acquire new microbial strains from sources like food, the environment, or other individuals. In some cases, a drastic reshaping of the gut microbiome is achieved through Fecal Microbiota Transplantation (FMT), which is typically used as a treatment for *Clostridiodes difficile* induced colitis [162]. Although FMT can be technically considered microbiome transmission, I will exclude it from further considerations because it is not an event that occurs in the daily life of most individuals.

Individuals may also attempt to modify their microbiome through changes in diet. Genomic evidence suggests that diet can lead to microbiome changes that are significant, but temporary [85]. There has also been a rise in the use and availability of probiotic supplements in recent years. These supplements are manufactured to contain high numbers of live bacteria in pre-defined quantities. Probiotic use may result in decreased symptoms for indications like irritable bowel syndrome [201], but the effect sizes in trials are generally smaller than claimed by probiotic manufacturers. Additionally, most probiotic strains do not engraft long-term, and taking a probiotic may even slow the gut microbiome's recovery from antibiotic treatment [167]. Somewhat worryingly, bloodstream

infections have been shown to originate from probiotic supplements in certain cases [196].

Overall, the evidence for how human microbiomes acquire new strains after adolescence is limited. While diet, environmental exposure or probiotic supplementation may have a short term impact, there are not yet genomic studies addressing the question of how adults acquire new microbiome strains that persist for the long term.

1.1.1 What is known about microbiome transmission between adults?

While mother-infant transmission of bacteria and phages is well-established, evidence for microbiome transmission in adults is lacking or less clear. Early evidence using 16S rRNA sequencing showed that cohabiting individuals may share gut bacteria [164], but the resolution provided by this method is insufficient to prove strain identity. More recent experiments from isolated communities in Fiji showed that individuals living together have more similar microbiomes than those outside the group [21]. While this work suggests that transmission between individuals is at play, the lack of time course sampling and lack of assembled genome evidence weakens the conclusions. More convincing evidence exists in model organisms like mice [108], but I maintain that genomically characterized, time course resolved evidence for microbiome transmission in human adults did not exist prior to this work.

1.1.2 Measuring transmission of the microbiome

In the Bhatt Lab's publications, we have analyzed microbiome transmission by advanced metagenomic sequencing techniques, including traditional Illumina short read, 10X Genomics Read Cloud [19], and Nanopore long read sequencing [110]. These methods allow for assembly of complete genomes from the microbes in a stool sample and enable strain-specific investigation into microbial identities.

There are four levels of microbiome measurement that are relevant to investigating transmission:

1. Species level - attainable through short read sequencing and classifying reads against a database of known microbes. If transmission between microbiomes is occurring, the same species needs to be present in both locations.

2. Full genome level - short read and long read sequencing allows us to create representative metagenome assembled genomes (MAGs) from a mixed microbial community. Comparing MAGs gives information about the degree of identity between them. However, MAGs only represent the dominant strain in a sample and are often confounded by assembly and binning errors.

3. Strain level - attainable when you have an accurate representative MAG or isolate genome and deep metagenomic sequencing (typically short read). When measuring strain-level variation

in a metagenomic sample, each sequencing read is thought of as a single observation of a cell in the sample. By aligning sequencing reads back to representative MAGs, you can measure the fraction of the bacterial population that has a particular allele. Challenges with this approach include picking the best reference genome, handling multi-mapping reads, and poorly assembled or missing reference sequences.

4. Complete phased haplotypes: Using a long-read sequencing technique like Oxford Nanopore, it's possible to phase bacterial strain haplotypes. This is the "holy grail" of strain-specific analysis, and would give a complete picture of the stain populations and their relative frequencies.

1.1.3 Determining transmission between two gut microbiome samples

In the abstract sense, strains colonizing two individuals recently after a transmission event should be "the same," as they were derived from the same ancestral population. However, confounding factors, including differing strain populations in the two samples, divergence in both strain populations since the transmission event, and sequencing errors all make determining "the same" more difficult. Identity between two bacterial strains can also be measured in different ways, depending on the abundance, strain diversity and sequencing depth:

1. SNP-based comparison: by aligning sequencing reads from multiple samples to the same reference genome and calling SNPs, strains present in low abundance can be compared. While very informative, SNP-based comparisons can miss structural variations, mobile genetic element insertions and other large-scale events.

2. Assembled genome comparison: by assembling MAGs from two samples and aligning them, the dominant strain in a sample can be compared. If an assembled genome has high identity, a SNP comparison of the same strains will also result in high identity. Identical assembled genomes significantly adds to my confidence that strains in two samples are identical.

3. Isolation, culture and sequencing: Isolating a collection of strains from each sample and sequencing them can reveal additional information beyond metagenomic sequencing. A combination of metagenomics and culture has recently been used in strain-specific investigations [2].

1.1.4 How close do strains need to be for them to be "the same?"

In the case of mother-infant transmission of crAssphage, it was clear when the assembled genomes from mothers and infants were the same strain. There were typically 1-2 SNPs separating the 100kb

genomes from mother-infant pairs, and thousands of SNPs separating genomes from different families. The heatmap in Figure 3.1 makes this clear: the cases of mother infant crAssphage transmission stand out as blocks of identity against the background.

Asking the same question in bacterial genomes for the HCT transmission work proved to be much more difficult. Bacterial genomes are 30-70x the size of the crAssphage genome, there are significant mobile genetic elements and repetitive regions that confound assembly, and strain diversity within the population makes comparisons much more difficult. The relatively simple methods I developed for the crAssphage research failed on more complex bacterial genomes. I was in the process of developing a method compatible with the high levels of bacterial strain diversity when I found a preprint from Matt Olm introducing the software tool called inStrain[122]. InStrain shared many of my ideas about comparing diverse bacterial populations and was implemented in a very computationally efficient manner. With this tool, I could skip the method development and get straight to investigating interesting biology.

InStrain analyzes alignments of sequencing reads from two samples against the same reference genome. The key metric is population average nucleotide identity, or popANI. Under this metric, a SNP is only called when the two samples do not share *any* alleles at a given site in the reference genome. This is different from consensus ANI (traditional SNP calling), where a SNP would be called whenever the consensus position differs between two genomes. SNPs are called in far fewer locations with popANI than conANI, therefore the popANI comparing two samples will always be greater than or equal to the conANI.

If two strains are related by a transmission event, they descended from the same ancestral population. Therefore, the popANI should approach 100%, but allowances have to be made for sequencing errors and divergence since the transmission event. Therefore it's necessary to set a threshold that is sensitive enough to capture all true cases of transmission, but specific enough to reject cases where closely related strains are not identical by descent. Based on evaluations of the same strain in time course samples from a single patient, as well as positive control cases where transmission is expected to occur (mother-infant and FMT patient samples), I settled on a popANI threshold of 99.999%. This means that two strains must have less than one difference out of every one hundred thousand base pairs in the genome to be called the same. These SNP based comparisons are not perfect, and will miss regions of the reference genome that are not covered with sequencing reads (either due to the region missing in the sample, or poor coverage), new genomic insertions, or large chromosomal rearrangements. Assembly-based approaches are required to catch these larger events.

1.1.5 Strains matter in the microbiome

While "species" is the basic unit of microbial classification, "strain" is the basic unit of microbiome assembly. Definitions vary, but species are thought to share 95% ANI [124]. How then, do you

define a strain? The right definition depends on the context. When searching for transmission, two strains should be nearly 100% identical before being called "the same". In an experiment focused on functional capabilities, such as antibiotic resistance, two strains might be called "the same" if they have identical antibiotic resistance profiles, even if the genomes were more divergent. In short, microbiome analysis needs to be made strain-specific to truly understand the complex interactions and capabilities of each unit.

An individual may have multiple different strains of the same species colonizing their gut, each with slightly different functional capabilities. In healthy individuals, genetic diversity within a species is stable on the scale of years [187]. Evolution within a strain population has its limits, and large-scale changes in microbiome composition are usually due to colonization with new strains rather than evolution of existing strains [54, 193].

Strain-specific microbiome analysis has only become common in the last few years, and methods for measuring, quantifying and assembling genomes from individual strains continue to be developed. I posit that many microbiome association experiments that have turned up null findings or failed to replicate may be because these associations were conducted at the species level, while the actual biological effect is occurring at the strain level. As modern methods develop and databases of MAGs from diverse human microbiomes continue to grow, repeating some of these analyses at the strain level may yield new findings.

Chapter 2

Intestinal microbiota domination under extreme selective pressures characterized by metagenomic read cloud sequencing and assembly

The work in this chapter was presented in:

Kang, J.B.*, Siranosian, B.A.*, Moss, E.L., Banaei, N., Andermann, T.M., and Bhatt, A.S. (2019). Intestinal microbiota domination under extreme selective pressures characterized by metagenomic read cloud sequencing and assembly. BMC Bioinformatics 20, 585.

2.1 Abstract

Low diversity of the gut microbiome, often progressing to the point of intestinal domination by a single species, has been linked to poor outcomes in patients undergoing hematopoietic cell transplantation (HCT). Our ability to understand how certain organisms attain intestinal domination over others has been restricted in part by current metagenomic sequencing technologies that are typically unable to reconstruct complete genomes for individual organisms present within a sequenced microbial community. We recently developed a metagenomic read cloud sequencing and assembly approach that generates improved draft genomes for individual organisms compared to conventional short-read sequencing and assembly methods. Herein, we applied metagenomic read cloud sequencing to four stool samples collected longitudinally from an HCT patient preceding treatment and over the course of heavy antibiotic exposure. Characterization of microbiome composition by taxonomic

classification of reads reveals that that upon antibiotic exposure, the subject's gut microbiome experienced a marked decrease in diversity and became dominated by Escherichia coli. While diversity is restored at the final time point, this occurs without recovery of the original species and strain-level composition. Draft genomes for individual organisms within each sample were generated using both read cloud and conventional assembly. Read clouds were found to improve the completeness and contiguity of genome assemblies compared to conventional assembly. Moreover, read clouds enabled the placement of antibiotic resistance genes present in multiple copies both within a single draft genome and across multiple organisms. The occurrence of resistance genes associates with the timing of antibiotics administered to the patient, and comparative genomic analysis of the various intestinal *E. coli* strains across time points as well as the bloodstream isolate showed that the subject's *E. coli* bloodstream infection likely originated from the intestine. The *E. coli* genome from the initial pre-transplant stool sample harbors 46 known antimicrobial resistance genes, while all other species from the pre-transplant sample each contain at most 5 genes, consistent with a model of heavy antibiotic exposure resulting in selective outgrowth of the highly antibiotic-resistant *E. coli*. This study demonstrates the application and utility of metagenomic read cloud sequencing and assembly to study the underlying strain-level genomic factors influencing gut microbiome dynamics under extreme selective pressures in the clinical context of HCT.

2.2 Introduction

Metagenomics involves the sequencing of a whole community of microorganisms directly from an environmental sample, such as soil or the human intestinal tract, often without prior knowledge of which species are present within the sample. In silico reconstruction of complete and contiguous genomes for individual organisms within a sequenced population remains a major challenge in the field of metagenomics. This is a challenging problem when using conventional shotgun short-read sequencing and assembly methods because short reads alone may not be able to determine the correct positions of DNA sequences that are both longer than the sequenced DNA fragment length (usually 50-300 base pairs) and present in multiple copies at different locations in the metagenome. The presence of such repeated regions (e.g. insertion sequences or the bacterial 16S rRNA gene) often result in fragmented assemblies where multiple instances of the repeated sequence are collapsed into a single contig instead of correctly placed in between unique flanking regions in multiple genomic locations.

Read cloud sequencing is a relatively new technique that was initially used in the context of human genomics to phase haplotypes [203]. This method has also been termed "linked-read sequencing." The main difference between read cloud and conventional short-read sequencing is that read cloud sequencing augments the library preparation stage to ultimately generate "read clouds," which are shortread sequences annotated with long-range information in the form of molecular barcodes. This is achieved by physically partitioning long DNA fragments into nanoliter-scale droplets and subsequently tagging all sequencing reads originating from a long fragment with a droplet-specific molecular barcode. Read cloud sequencing offers a favorable combination of long-range information, high base call accuracy, high throughput, and low input DNA mass requirements [203]. The 10x Genomics Chromium platform is a commercially available read cloud library preparation system that automates the pipetting steps necessary to generate the molecular barcodes. Recently, we developed an approach to adapt read cloud sequencing for metagenomic applications. The resultant barcoded data is deconvolved and genome draft assembly is achieved using a combination of existing standard genome assemblers as well as a custom assembly tool called Athena [19]. We have recently applied the approach to sequence ocean sediment samples and the healthy human microbiome, for which it was able to generate contiguous draft genomes for individual organisms from bacterial mixtures [19].

In this study, we investigate a clinical application of metagenomic read cloud sequencing in the context of hematopoietic cell transplantation (HCT), which is a complex medical procedure used in the treatment of hematologic disorders such as leukemia and lymphoma. During HCT, patients initially undergo intensive treatment with chemotherapy and sometimes radiation therapy; this 'conditioning regimen' serves to prepare patients to receive a hematopoietic stem cell graft. Multipotent hematopoietic stem cells derived from bone marrow, peripheral blood, or umbilical cord blood are then infused into the patient to reconstitute all blood cell lines. The procedure can

be curative but comes with high risk for complications, including infection and graft-versus-host disease (GVHD), an inflammatory disease where donor immune cells attack the recipient's healthy tissue. Intestinal microbial dysbiosis preceding and following HCT has been found to be associated with an increased risk for developing bloodstream infections [170]. Previous studies also show that decreased intestinal diversity is associated with development of GVHD and higher overall mortality in HCT [171]. Broad-spectrum antibiotics and other drugs administered during the course of HCT can greatly change the composition of the gut microbiota. In some cases, such microbial dysbiosis leads to domination of the intestine by a few or even a single genus or species, increasing the likelihood of complications like bloodstream infections in these immunocompromised patients [170]. Intestinal domination may happen because certain bacterial strains carry an advantage, such as antibiotic resistance, that enables them to flourish after other antibiotic-sensitive commensal microbes are eliminated. While intestinal domination is relatively common in this patient population, the process by which it occurs is not well-understood.

Herein, we apply the metagenomic read cloud sequencing approach to patient stool samples collected over multiple time points pre- and post-HCT to elucidate microbiome dynamics in response to extreme selective pressures during HCT. We find that antibiotic exposure is associated with intestinal domination by Escherichia coli in our study subject. Read cloud sequencing, but not short read sequencing alone, was able to identify many antibiotic resistance genes within the dominating strain of *E. coli*. Thus, we postulate that the gut domination observed was the consequence of enhanced fitness of this organism in the presence of antibiotics.

2.3 Results

2.3.1 Microbiome composition and diversity across the clinical time course

Stool samples were collected from the patient over five time points spanning 70 days. The samples (denoted A-E) correspond to days - 2, + 19, + 27, + 33, and + 68 relative to transplantation. Figure 1 plots the microbial diversity as measured by the Shannon diversity index as well as the species-level taxonomic composition (from metagenomic classification of conventional short-read data) of the patient's gut microbiome over time in relation to when the patient was administered various antibiotics. Across the time course spanning 70 days, Shannon diversity was found to decrease markedly from time point A through a period of intestinal *E. coli* domination (samples C and D) before completely recovering by time point E. The patient exhibits the *E. coli* gut domination after the time of GVHD onset on day + 19 and before the clinical manifestation of the *E. coli* bloodstream infection on day + 60. We calculated the Bray-Curtis dissimilarity index between pairs of samples and performed Principal Coordinates Analysis (PCoA) to visualize microbiome composition (Fig. 2). Most of the variance in the PCoA plot is captured by the stark difference in *E. coli*-dominated samples (C and D), as expected. Time points A and E are more similar than time points A and B,

suggesting recovery of a similar microbial community. However, we note that time point B occurred after the completion of the transplant and engraftment process, while the patient was exposed to several antibiotic agents. Sample E also has significant representation of species not found in time point A, including a 16% fraction of *Lactobacillus rhamnosus*. Recovery of diversity and original microbial community structure after HCT could occur through persistence of microbes in very low fractions, acquisition of new microbes following the HCT process, or a combination of both. To evaluate these options, we examined if microbial genomes assembled from identical organisms at multiple timepoints had high nucleotide similarity (see Methods). Of species present at a relative abundance greater than 2% in multiple samples, 8 species are present at time points A, B and E. Five out of 8 species had $> 99.9\%$ nucleotide similarity between time points A and B, likely indicating the same dominant strain is present at both time points. Lower A-B similarity for other species could be the result of different strain populations between time points or poor assembly, as these species had < 1 Mb of assembled and aligned sequence. In all cases, sequences assembled from species present at time points A and E had $< 99.5\%$ similarity. Interestingly, *Enterococcus faecium* is $>$ 99.9% similar between samples B, C and D, but much different at time point E (96% similarity, E compared to other time points). This suggests the same dominant strain of *Enterococcus faecium* is retained though the *E. coli* domination event, but a different strain is acquired or dominant by time point E. Similar results were achieved with short-read and Athena assemblies, when data were available. Taken together, these results suggest that dominant original strains are not retained in the microbiome through the clinical time course. However, this analysis cannot rule out lowly abundant strains that did not contribute to the genome assembly, which could be present either before or after the *E. coli* domination event.

2.3.2 Assembly of draft genomes

We separately performed both conventional short-read assembly (MEGAHIT) and read cloud assembly (Athena) and binned the resulting contigs into draft genomes for individual organisms present within each metagenomic sample (see Methods). We assessed the draft genome bins using CheckM and defined "high-quality" bins as attaining $> 90\%$ completeness and $< 5\%$ contamination, following a previously described standard [20]. By this standard, read cloud sequencing and Athena assembly produced 16 high-quality draft genomes for time point A (listed in Table 1), whereas conventional short-read sequencing and assembly produced 6 high-quality genomes. Binning results and assembly metrics for Athena draft genomes generated for each time point can be found in Additional file 2. Figure 3 shows a visual comparison of the *E. coli* draft genomes generated using read clouds compared to conventional sequencing for time points C and D, when *E. coli* comprises the most abundant organism in the sample. Compared to the conventional assembly, the Athena assembly demonstrated an order of magnitude increase in contig N50. An assembly's N50 is a metric of contiguity defined as the length of the shortest sequence such that 50% of the entire assembled genome

is included in contigs of greater or equal length (higher N50 indicates greater contiguity). The draft genome for sample C was the most contiguous and complete *E. coli* assembly, containing 5.16 Mb of sequence in 23 contigs with an N50 of 1.32 Mb. Overall, these results support our previous finding [19] that read cloud sequencing and Athena assembly improves the reconstruction of genomes of individual organisms within microbial mixtures.

2.3.3 Detection of resistance genes

We aligned the predicted protein-coding sequences from the Athena-assembled metagenomes for samples A, C, D, and E against the Comprehensive Antibiotic Resistance Database (CARD) database, which yielded 87 (71 unique), 72 (72 unique), 101 (86 unique) and 15 (11 unique) resistance genes, respectively. Herein, we use the term resistance gene to refer to any gene present within the CARD database, which comprises genes known to confer antibiotic resistance and regulators of such genes. In the entire metagenome assembled for sample A, we detected several resistance genes present in multiple copies: tetO (7 copies), cfxA3 (5 copies), mefA (3 copies), tetQ (3 copies), tet(40) (2 copies), and ermF (2 copies). We found that copies of identical resistance genes occurred both within the genome of the same organism and among different organisms. For instance, tetO was present on 3 contigs that all belonged to the *Lachnospiraceae* bin, and it was also present in single copy in draft genomes classified as *Blautia sp.*, *Clostridium*, *Eubacterium rectale*, and *Ruminococcus gnavus*. Inspection of the genomic regions of the 3 *Lachnospiraceae* contigs containing tetO revealed that the regions with the resistance gene share some homology but are not completely identical. Note that no resistance gene duplication was observed for sample C. For sample D, a set of 13 resistance genes (acrB, acrD, baeR, cpxA, CRP, emrB, emrR, marA, mdtB, mdtC, msbA, patA, and sul1) was detected in the draft genomes of both *E. coli* and *K. pneumoniae*. Although both organisms share this same set of genes, we did not find evidence for horizontal gene transfer because the genes themselves are not identical (different numbers of mismatch from the reference), and the contigs on which the genes are present have homology in the region of the resistance genes but are not completely identical as determined by alignment dotplots of the contig pairs. For sample E, the dfrF gene appeared in 5 distinct copies in 4 different organism bins. Positive selection for the dfrF gene may have potentially occurred given that trimethoprim was administered to the patient prior to time point E. Performing the equivalent resistance gene analysis on the conventional sequencing data for samples A, C, D and E revealed 27 (27 unique), 84 (84 unique), 94 (82 unique) and 9 (9 unique) resistance genes, respectively. Compared to read cloud assembly, a greater proportion of resistance genes detected in the conventional data are unique (in single copy) within their assembly as genes present in multiple copies are collapsed into a single sequence in the absence of barcode information. The specific resistance genes detected within each read cloud and conventional sample as well as alignment metrics are listed in Additional file 3. These results show that the ability to resolve numerous copies of the same resistance gene present in one or multiple distinct organisms

within the proper genomic context is a notable technological advantage of the read cloud sequencing over conventional methods.

2.3.4 Comparative genomic analysis of *E. coli* strains

We postulated that comparison of the *E. coli* draft genomes across time points would reveal genomic differences between the *E. coli* assemblies. Assuming that the assembled *E. coli* genome for a given time point represents the most abundant strain of *E. coli* in the sample, significant genomic differences across time could indicate acquisition of a new strain, selection and subsequent outgrowth of a previously low-abundance strain, or possible remodeling of the genome. We also hypothesized that the particular strain of *E. coli* producing the bloodstream infection could be traced back to the gut microbiome based on our previous findings in [168]. To assess *E. coli* strain similarities, we aligned pairs of *E. coli* draft genomes from the various stool time points and the bloodstream isolate against each other (see Methods). Table 2 lists the average percent nucleotide identity, total number of SNPs, and total bases aligned for each pair of genomes. We also included NCBI *E. coli* S88 reference genome in the analysis to serve as a comparison to a strain that is also a known extraintestinal pathogen but unrelated to our patient.

We discovered that the dominant intestinal *E. coli*strains present in samples A, C, and D contain relatively few SNPs and share extremely high nucleotide identity.The number of SNPs ranged from 371 to 3811 (compared to 56,513 SNPs with the S88 reference) and percent nucleotide identity ranged from 99.91 to 99.98% (compared to 98.61% identity with the S88 reference). Somewhat interestingly, the bloodstream isolate (day + 60) genome most closely matched the draft genome from sample C(day + 27) with 182 SNPs and 99.99% nucleotide identity,even though the patient's clinical manifestation of bloodstream infection occurred after time point D (day + 33)with 3742 SNPs and 99.91% identity. The low number ofSNPs and high percent identity between the stool sample*E. coli* strains and the bloodstream isolate reveal that the same *E. coli* strain existing in the patient's intestine prior to HCT likely persisted in spite of antibiotics, expanded to dominate the gut, and also eventually caused the patient's bloodstream infection. Our group initially analyzed theshort-read libraries of these samples via an orthogonal bioinformatic approach as described in [168], which also suggested that the intestine was the source of the bloodstream strain for this patient.

In order to ascertain whether any large-scale genomic island incorporation or genomic remodeling took place in the dominant *E. coli* strain over time, we visualized pairwise genome alignments of the various strains as syntenic dotplots, which can compare two genomes to each other. Each main axis represents the entire length of one genome being compared, and a colored dot is plotted at regions where the genomic sequences match between the two genomes (areas of synteny). For example, comparing two completely identical genomes would produce a dotplot with a perfectly contiguous diagonal stretching from the bottom-left to top-right corners. Figure 4 shows the synteny

dotplots comparing *E. coli* strains from sample A to sample D and comparing the bloodstream isolate to sample C. Visual inspection of the plots showed no evidence for any large genomic island incorporations. The lack of major discontinuities or inversions provide additional evidence that the strains are genetically equivalent from a genome structure perspective across the various time points and between the gut and the bloodstream.

2.3.5 Antibiotic resistance genes in pre-transplant *E. coli* strain

Given that the *E. coli* strain dominating the intestine likely originated from a single original strain that persisted through the extreme selective pressures of antibiotic administration, we hypothesized that the pre-transplant (time point A) strain harbored antibiotic resistance genes that potentially aided its survival. By aligning the predicted protein-coding regions of the Athena-assembled *E. coli* draft genome from sample A against the CARD database, we detected 46 known antibiotic resistance genes (Table 3). Functional annotations of these genes revealed that the majority of genes code for proteins related to drug efflux pumps, and others encode known resistance mechanisms to aminoglycosides, bacitracin, and polymyxin. There was also a gene (CTX-M-27) that confers extended-spectrum beta-lactamase resistance.

Next, we evaluated the fitness of the pre-transplant *E. coli* strain to other organisms present in the same stool sample at time point A by comparing the resistance gene content of *E. coli* to that of the other organisms. Out of the total 87 resistance genes detected in the entire metagenome for sample A, 46 were localized to contigs in the *E. coli* draft genome bin. The remaining 41 genes were distributed widely across many other organisms, with no individual bin containing greater than 5 resistance genes. The organisms containing the second-highest number of resistance genes (each with 5 genes) were classified at the genus level as Lachnospiraceae and Eubacterium. Because all organisms with a near-complete draft genome possessed no more than 5 resistance genes, our results support a model in which the particular *E. coli* strain present in the subject's microbiome prior to transplant was able to achieve gut domination over other organisms due to the selective pressures applied by antibiotics.

2.4 Discussion

Our results show that the metagenomic read cloud sequencing methodology allows for more comprehensive and contiguous recovery of individual bacterial genomes from a sequenced community within the gut microbiome of an HCT patient. The improved assemblies allow for augmented detection of antibiotic resistance genes that are present in multiple copies in the metagenome and facilitates comparative genomic analysis to ascertain strain similarity. Recovery of microbial diversity is expected following HCT, but previous research has shown that the post-HCT microbiome is often different than the pre-HCT microbiome [136]. Our results corroborate these findings as microbiome diversity

is restored at time point E without the recovery of the original species and strain-level composition. We find that the assembled genomes for organisms present at time point E compared to other time points are actually quite different (< 99.5 similarity for strains of the same species). Several potential mechanisms could explain this finding: for example, a new strain (either externally acquired or a previously rare strain) may become dominant due to selective fitness advantage; alternatively, drug exposure occur- ring over the clinical time course may drive widescale mutagenesis of the dominant strain within these organisms.

Bacteroides was the most abundant genus in the subject's microbiome prior to transplantation (sample A).The patient was then administered multiple antibiotics,and the microbiome concurrently developed markedly decreased diversity until becoming dominated by *E. coli*.Previous studies have established Bacteroides to be an abundant and prevalent genus in the healthy human gut microbiome; conversely, healthy populations rarely exhibit gut domination by Proteobacteria like *E. coli* [98].By characterizing the presence of antibiotic resistance genes in the gut metagenome, we discovered that the E.coli strain present at time point A, before transplant and before any antibiotic administration, already contained avast arsenal of antibiotic resistance genes. Increased fitness due to a greater number of resistance mechanisms may have afforded this particular *E. coli* strain a selectivead vantage, enabling it to survive as other organisms were eliminated by the antibiotics.

In the setting of the specific antibiotics administered to the patient, the survival of the dominating *E. coli* strain may be explained in part by the resistance genes detected in its genome. Preceding the *E. coli* domination observed starting at time point C (day + 27), the patient had received the following antibiotics in chronological order: ciprofloxacin (day - 2 to + 12), cefepime (day + 2 to 3), vancomycin (day + 2 to 9), meropenem (day + 3 to 17), daptomycin (day + 9 to 11), levofloxacin (day + 17 to 32), and metronidazole (day + 21 to 33). The strain's observed resistance to ciprofloxacin and levofloxacin (members of the fluoroquinolone class of antibiotics) can potentially be explained by multidrug efflux complexes AcrAB-TolC, AcrEF-TolC, EmrAB-TolC, and MdtEF-TolC as well as multidrug resistance proteins MdtH and MdtM, which are all annotated in CARD as potentially conferring fluoroquinolone resistance. The observed resistance to Piperacillin/tazobactam (a penicillin) and cefepime (a cephalosporin) may be attributed to CTX-M-27. The patient's bloodstream infection was due to a highly resistant extended-spectrum beta-lactamase (ESBL) *E. coli* bacteria, and most of the ESBL *E. coli* infections in the U.S. are accounted for by CTX-M-type enzymes [39]. Our analysis did not identify resistance genes that can explain the ability for this particular strain of E.coli to survive despite the use of meropenem; however, a decrease in uptake of antibiotics due to a deficiency of porin expression or biofilm formation may possibly be involved [116]. *E. coli* possesses native resistance to daptomycin and vancomycin, which both target Gram-positive organisms.

While this analysis follows a single HCT patient, our findings have broader clinical implications.

We demonstrate that the intestinal microbiome of patients can act as a reservoir of antibiotic resistance genes, which may govern which organisms are most predisposed to endure and dominate the gut under the extreme selective pressure applied by antibiotics. Although broad-spectrum antibiotics remain a vital part of our medical armamentarium, the issue of increasing antibiotic resistance strongly argues for their conscientious use. Antibiotics can both select for antibiotic resistance and contribute to the loss of commensal organisms and resulting expansion of a few organisms or even a single organism to the point of gut domination. Further studies are warranted to investigate whether our findings generalize to other HCT patients as well. It is conceivable that the antibiotic resistance gene potential of organisms present prior to transplantation can be used to predict or explain eventual gut domination events or bloodstream infections. Additionally, it is important to note that the resistance genes detected in this study are limited to known antibiotic resistance mechanisms present within the CARD database, and commensals likely have mechanisms of resistance that remain unknown.

2.5 Conclusion

This case study serves as an example of how advanced DNA sequencing technologies can help to illuminate complex biological phenomena occurring within real patients. We explore a clinical application of our recently developed metagenomic read cloud sequencing and assembly approach to study gut microbiome dynamics under the intense selective pressures caused by heavy antibiotic administration in the context of HCT. Because intestinal domination has been linked to poor outcomes in this patient population, we applied read cloud sequencing to longitudinal stool samples of an HCT patient who developed *E. coli* gut domination and a subsequent bloodstream infection. Read cloud sequencing and the Athena assembler provided a higher-resolution characterization of microbiome dynamics surrounding the period of domination than conventional short-read sequencing alone, as it generated draft genomes for constituent organisms in the patient's microbiome with greater completeness and contiguity. Moreover, the improved assembly using read cloud sequencing enhanced our ability to assemble multiple copies of conserved and repeated sequences (e.g. antibiotic resistance genes) within their proper genomic context.

The generation of high-quality assemblies enabled the genomic comparison of organisms over time. We find that although microbial diversity recovers in our subject post-HCT, for most organisms the original dominant strains are not retained throughout the clinical time course. By performing comparative genomic analysis on the *E. coli* strains between the gut microbiome across time and the bloodstream, we found that a single highly resistant strain of *E. coli* originally residing within the patient's baseline microbiome prior to HCT and antibiotic treatment persisted to eventually dominate the subject's microbiome and also instigate the bloodstream infection. By detecting known antibiotic resistance genes within the assembled genomes, we discovered that the

E. coli strain present before transplant was armed with a large collection of resistance genes whereas other organisms initially present in the same intestinal community lacked such extensive resistance potential. These findings are aligned with a model in which the eventual gut domination by *E. coli* can be attributed to its increased fitness compared to other organisms, leading to its outgrowth under extreme selective pressures. A more comprehensive understanding of microbiome dynamics occurring in HCT could potentially lead to the development of personalized antibiotic regimens based on the gene content of microbial strains within an individual's microbiome or microbiome-related treatments to improve patient outcomes by preserving or enhancing microbiota diversity during the course of HCT.

2.6 Methods

2.6.1 Sample preparation and sequencing

As part of our original previously published investigation of bloodstream infections in HCT recipients [168], we performed a retrospective cohort study, approved by the Stanford institutional review board under IRB protocol #42053 (principal investigator: A.S.B.). Informed consent for weekly stool sample collection on all Stanford HCT patients was obtained under protocol #8903 (principal investigator: David Miklos). All fresh stool samples were placed at 4°C immediately upon collection, aliquoted into 2 mL cryovial tubes within 24 h, and stored at -80 °C.

One study subject undergoing HCT was unique in having a simultaneous *E. coli* and Methicillin-resistant *Staphylococcus aureus* (MRSA) bloodstream infection [168]. Furthermore, this patient also had a total of five longitudinal stool samples (denoted A-E) in addition to the *E. coli* isolate cultured from the bloodstream infection available for sequencing. While MRSA was not found in the patient's stool sample, the E.coli bloodstream isolate appeared indistinguishable from the same strain in the intestine using short-read sequencing [168]. We chose to further investigate this patient's samples using read cloud sequencing for even more precise longitudinal strain-level analysis.

From the frozen stool samples, we isolated microbial cells from stool debris by differential centrifugation following a previously described protocol [82]. 400 mg of frozen stool was vortexed with 1 mL 0.9% saline solution for 30 s, then centrifuged at 3000 rpm (645 g) for 2 min. The pellet containing stool debris was discarded, and the supernatant was centrifuged at 10,000 rpm (7168 g) for 3 min to spin down bacterial cells. The saline supernatant was discarded, and the differential centrifugation process was repeated with 1 mL of phosphate-buffered saline (pH 7.4) to acquire a purified microbial pellet.

For read cloud sequencing, we extracted high-molecular weight DNA from the purified microbial pellet using the Gentra Puregene Yeast/Bacteria Kit following the manufacturer's protocol with the following modifications to increase DNA yield: increased lytic enzyme volume to 5.0 μL and increased protein precipitation solution to 130 μL. For conventional sequencing, we extracted DNA directly

from frozen stool using the Qiagen QIAamp DNA Stool Mini Kit modified with an added step after addition of buffer ASL in which the samples underwent seven alternating 30s cycles of beating with 1 mm diameter zirconia beads in a bead beater (Biospec Products) and chilling on ice. The extracted DNA was visualized by agarose gel electrophoresis, and concentration estimations were performed for both Qiagen and Puregene DNA using Qubit fluorometric quantitation. The concentration of DNA extracted for time point B was too low to be used as input for read cloud sequencing; therefore, the read cloud sample for time point B was excluded from downstream processing. For all other time points, we removed small (< 10 kb) DNA fragments by size selection prior to read cloud library preparation using a BluePippin agarose electrophoresis instrument.

The size-selected high-molecular-weight DNA was used as input for read cloud library preparation. We prepared 10x Chromium libraries using the Chromium instrument and reagents from 10x Genomics (Pleasanton, CA). Additionally, we prepared conventional Illumina Truseq libraries for all five time points (A-E) as well as the bloodstream isolate according to the Illumina Truseq Nano protocol. We quantified library fragment size using a Bioanalyzer 2100 instrument (Agilent Technologies). The four 10x Chromium libraries were multiplexed and sequenced on one lane of Illumina HiSeq 4000 using 2×150 bp paired-end reads (11-16 Gb of sequence coverage per library). The Illumina Truseq stool libraries were multiplexed and sequenced on an Illumina HiSeq 4000 instrument using 2×101 bp reads (4-5 Gb of sequence coverage per library).

The bloodstream bacterial isolate of *E. coli* was collected and stored by the Stanford Health Care Clinical Microbiology lab, as part of the previously published investigation of bloodstream infections in HCT recipients [168]. We extracted isolate DNA from colonies grown in small volume liquid culture following the manufacturer's protocol for the Gentra Puregene Yeast/Bacterial Kit and sequenced the Illumina Nextera XT library on an Illumina HiSeq 4000.

2.6.2 Quality control of reads

The samples were demultiplexed using Illumina's bcl2fastq v2.19. For the read cloud libraries, we extracted the 16 bp 10x barcode from each read using the Long Ranger Basic pipeline (10x Genomics). Next, we performed identical quality control and filtering procedures for raw reads generated from all stool libraries (both read cloud and conventional): read quality was assessed with FastQC v0.11.4 [9] and quality trimming was performed with cutadapt v1.8.1 using a minimum length of 60 (-m 60), minimum terminal Phred quality cutoff of 30 (-q 30, 30), and N-end trimming (-trim-n) [104].

2.6.3 Taxonomic classification of reads and diversity calculation

To measure the microbial composition of our short-read sequencing samples, we used the Kraken2 taxonomic sequence classifier with default parameters [190] and a comprehensive database containing all bacterial and archaeal genomes in Genbank assembled to "complete genome" or "chromosome"

quality as of October 2018. Kraken2 classifies individual reads by mapping all k-mers(k = 35) to the lowest common ancestor genome in the database. Bracken [100] was then used to estimate species abundance. The Shannon diversity index was calculated for each sample at the species level using the R package Vegan (version 2.5-4) [120]. Shannon diversity was calculated on samples rarefied to 7,360,000 paired-end reads,the number in the lowest covered file.

2.6.4 Generation of organism draft genomes

We assembled the quality-controlled reads for both the read cloud and conventional libraries using the short-read assembler MEGAHIT v1.1.3 [86], which first builds a succinct de Bruijn graph from k-mers, then forms assembled contigs by finding paths through the graph. We performed no further assembly for the conventional samples (the MEGAHIT contigs constituted the final contigs comprising the draft genomes). For read cloud samples, we used BWA v0.7.10 to perform sequence alignment of the raw reads against the MEGAHIT contigs [88]. We then used the Athena assembler to further assemble the MEGAHIT seed contigs. Athena takes as input the barcoded reads (FASTQ), the seed contigs (FASTA), and the alignment file (BAM), and it returns contigs assembled with read clouds (see [19] for full details of Athena).

Next, we clustered the individual contigs generated from Athena into bins representing nearly complete organism genomes. Binning was achieved by using four established metagenomic binning tools: MetaBAT2 [73], MyCC [94], CONCOCT [6], and MaxBin 2.0 [192]. We then used DAS Tool to integrate the results from the various binning methods to yield a single set of nonredundant bins with maximal coverage of single-copy core genes [157]. We assigned a taxonomic classification to each individual contig using Kraken2 [190]. We assigned a taxonomic designation to an entire bin if greater than 60% of contigs in the bin shared the same Kraken2 identification. For each resulting bin, which represents an organism draft genome, we used QUAST to assess the size and contiguity of the assembly [60]. We used CheckM to calculate metrics of genome completeness (existence of expected core genes) and contamination (duplication of core genes expected to exist in single copy) for each draft genome [129]. We used the circlize package in R [58] to visualize and compare the assemblies and Prokka [148] to predict the protein-coding genes in each contig.

2.6.5 Comparative genomic analysis

To quantify the similarity between the various *E. coli* strains across time points (A, C, and D) and between the stool and bloodstream isolate, we used the NUCmer script within MUMmer v3.23 to perform pairwise alignment of the *E. coli* draft genomes from each pair of samples [37]. We also included the full genome for extra-intestinal pathogenic *E. coli* strain S88 (NCBI accession CU928161.2) in the analysis as a comparison. For each pair of assembled draft genomes of the various *E. coli* strains, we calculated the percent nucleotide identity, number of single nucleotide polymorphisms (SNPs), and total number of aligned bases. Additionally, we generated syntenic

dotplots for each pairwise comparison using the mummerplot script with layout option (-l), which reorders and orients the contigs to the main diagonal of the plot for optimal viewing [37].

A reference-guided assembly method was used to compare species present at multiple time points when species were too lowly abundant to obtain unbiased bins. For both conventional and read cloud sequencing, reads were aligned against the NCBI reference genome for a given species with BWA [88], mapped reads were extracted with SAMtools [89] and assembled with metaSPAdes [118]. Athena assembly was conducted on read cloud data. Resulting contigs were filtered to a minimum length of 500 bp, and pairs of time points were aligned with MUMmer. Only alignments with $>$ 100 kb 1-1 aligned sequence were reported.

2.6.6 Antibiotic resistance gene detection

We detected the presence of antibiotic resistance genes within contigs generated from each sample by aligning the predicted protein-coding genes against the Comprehensive Antibiotic Resistance Database (CARD), a curated database of genes known to be determinants of antibiotic resistance [70]. The "protein homolog" model of the CARD database was used in order to minimize false positives. We performed the alignment using DIAMOND [25] and filtered the results to sequences exceeding both 90% identity and 90% coverage of the reference sequence in CARD.

2.7 Figures

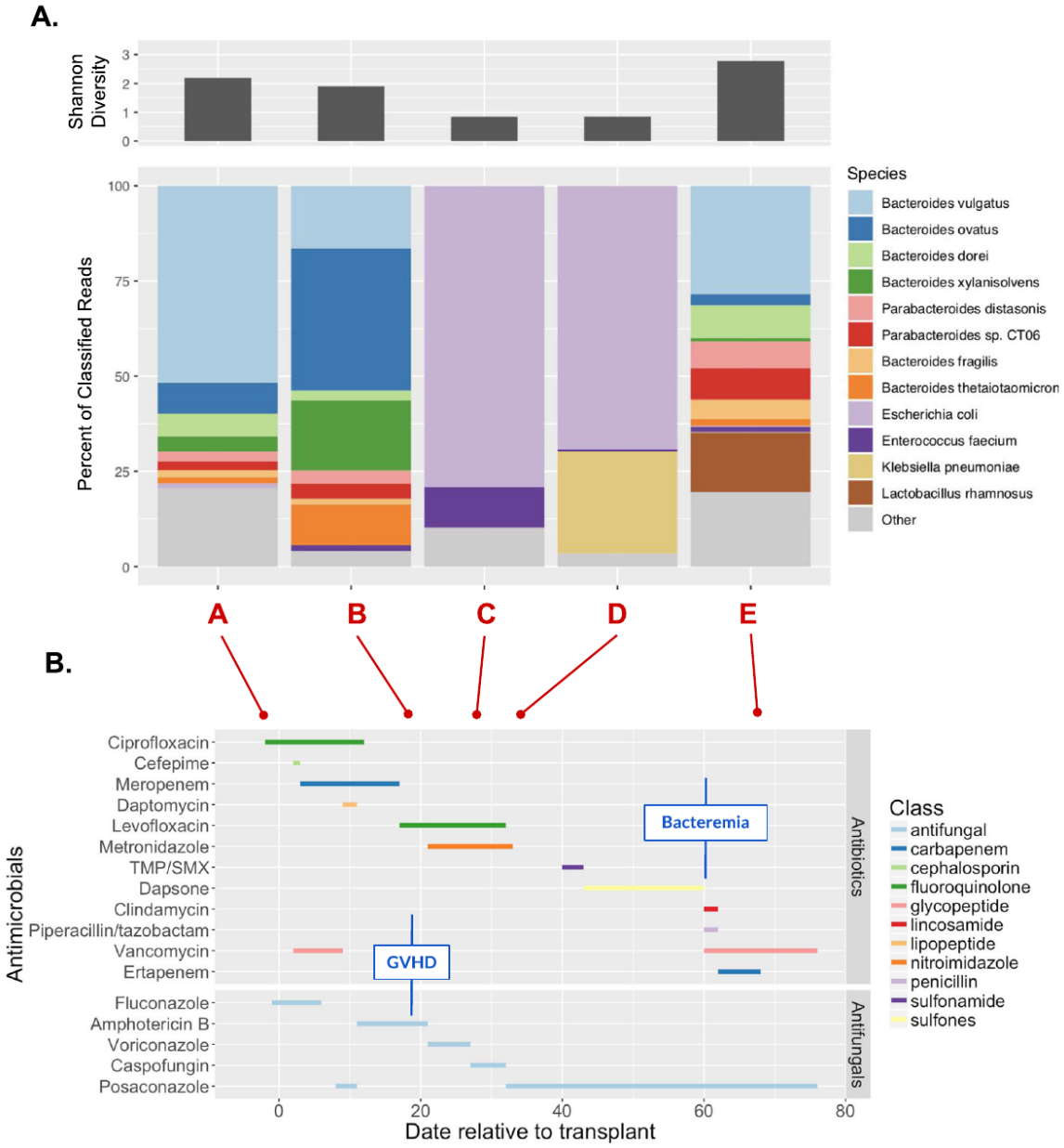


Figure 2.1: **A.** Shannon diversity and composition of the intestinal microbiome of the study subject across five time points over the course of HCT obtained from species-level taxonomic classification of conventional short-read samples. Each bar represents one stool sample, where colors represent different species and thickness indicates relative readcount attributed to that species within the sample (proportion of total reads classified to the species level). "Other" represents species comprising <2% readcount. Microbial diversity decreases to a period of domination by *E. coli* (time points C and D) followed by recovery of diversity (time point E). **B.** Clinical time course of the study subject. The x-axis denotes number of days after transplantation. Dates on which a stool sample was collected are marked by red dots. Each row portrays the start and end date of administration of an antibiotic (antibiotic class indicated by the color of the line). The timing of GVHD onset and bloodstream infection (bacteremia) are marked

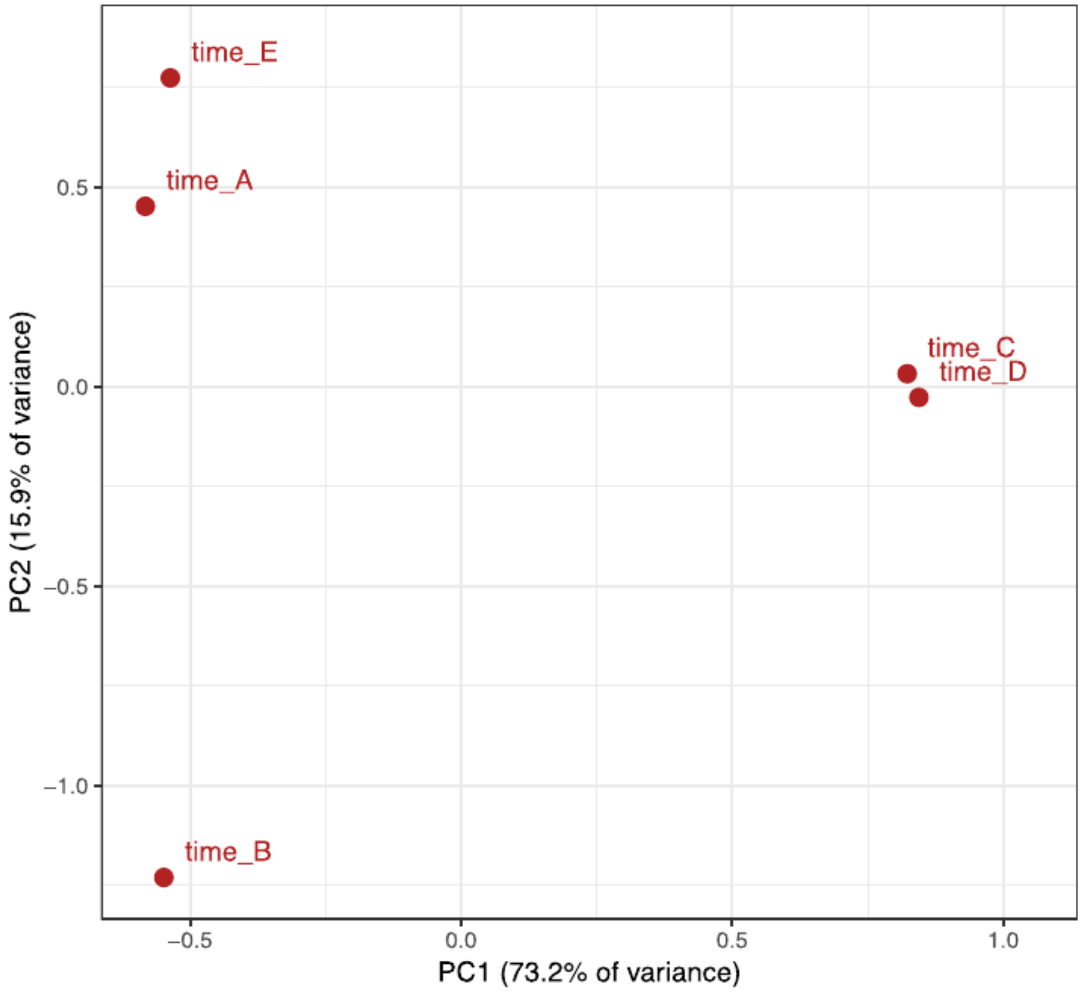
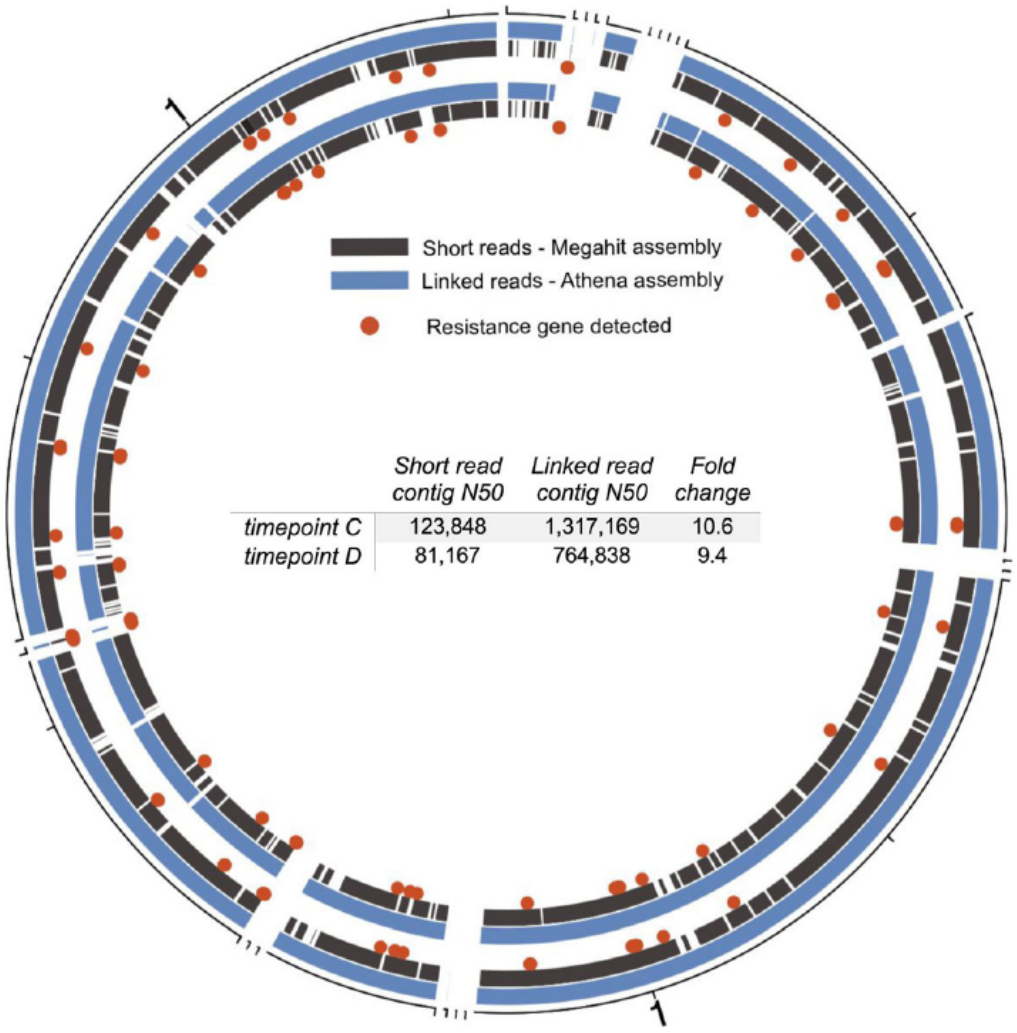


Figure 2.2: Principal Coordinate Analysis (PCoA) of microbiome content classified at the species level (Bray-Curtis beta diversity metric). Most of the variation is captured in the x-axis and separates *E. coli* dominated samples from the rest. Time points A and E are closer together than time point B, showing the recovery of a similar microbiome community following transplant.



	Short read contig N50	*Linked read contig N50*	*Fold change*
timepoint C	123,848	1,317,169	10.6
timepoint D	81,167	764,838	9.4

Figure 2.3: Circos plot showing *E. coli* draft genomes for sample C (outer track) and D (inner track) constructed with read clouds and Athena assembly (blue) compared to conventional short reads and MEGAHIT assembly (dark grey). Athena assembly demonstrates enhanced contiguity with an approximately 10-fold improvement in N50 for both samples compared to the conventional assembly. Red dots mark genomic locations where resistance genes were detected. Red dots located at breaks in the grey track identify resistance genes detected in the Athena assembly but were missing from at least one of the short-read assemblies.

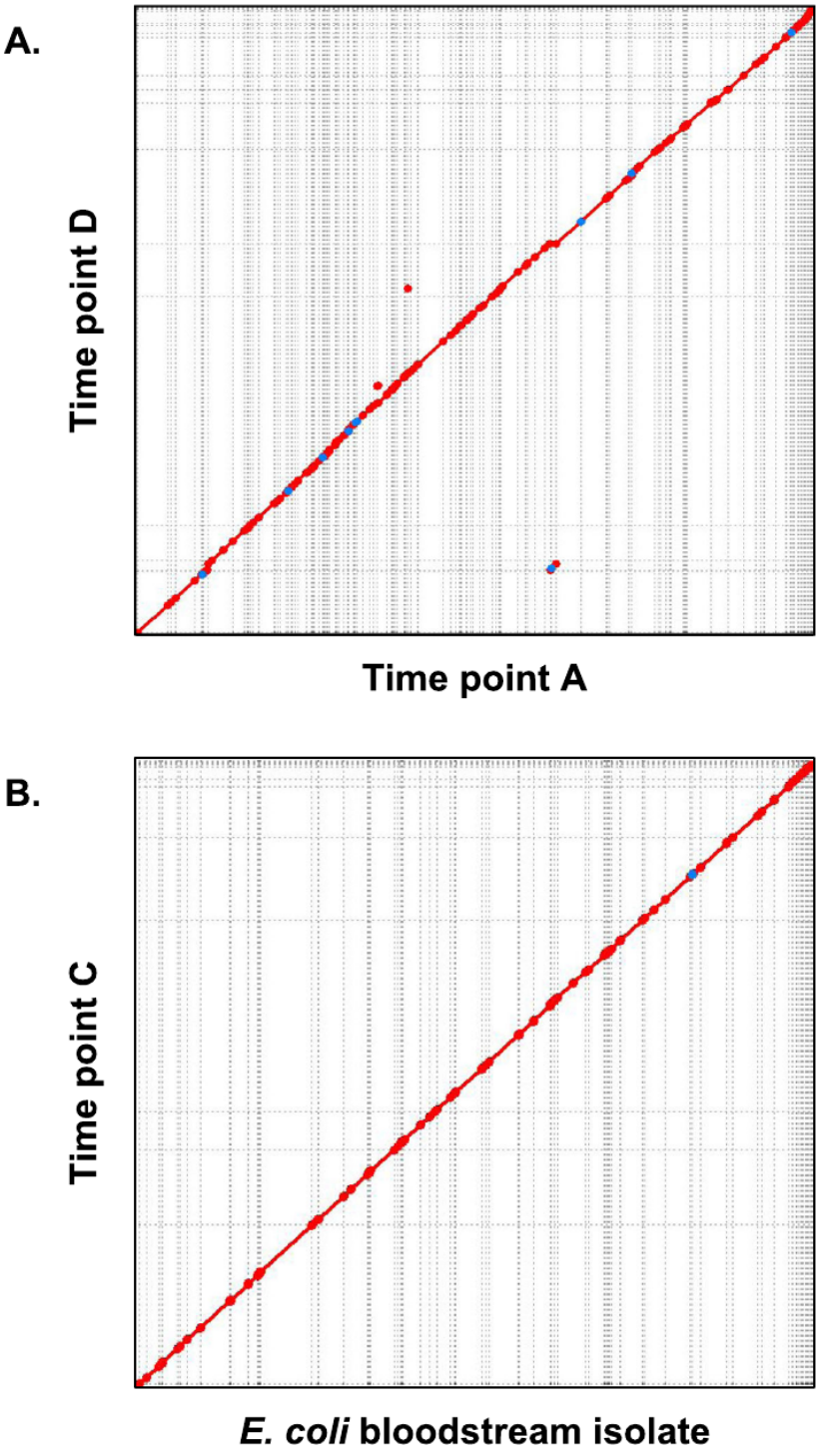


Figure 2.4: Syntenic dotplots comparing *E. coli* strains across time points and between the intestine and the bloodstream. Regions of sequence identity are marked by colored lines. **A.** Sample A draft genome (x-axis) compared to sample D draft genome (y-axis). **B.** Bloodstream isolate genome (x-axis) compared to sample C draft genome (y-axis). The near-perfect correspondence reveals that the bloodstream isolate is concordant with and thus likely originated from the intestinal microbiome.

2.8 Tables

Organism	Size (Mb)	Coverage	Completeness	Contamination	N50
Catenibacterium sp.	2.57	50.16	100	0	160,908
Erysipelotrichaceae bacterium	4.33	50.59	100	3.77	498,545
Streptococcus thermophilus	1.74	21.49	99.89	0.58	49,696
Faecalibacterium prausnitzii	2.95	45.28	99.66	3.17	292,610
Eubacterium rectale	3.32	178.86	99.52	0.72	375,749
Flavonifractor plautii	3.6	52.02	99.33	0.81	983,109
Eubacterium (Genus)	2.91	29.92	99.33	2.68	148,852
Bacteroides vulgatus	5.35	629.82	98.5	0.19	502,539
Escherichia coli	4.96	20.48	98.4	0.58	70,983
Parabacteroides distasonis	5.28	77.48	98.27	0.83	455,277
Streptococcus parasanguinis	2.1	17.92	97.89	0	46,401
Clostridium sp.	3.08	18.87	97.63	0	42,920
Bifidobacterium longum	2.47	44.91	97.62	1.08	111,224
Blautia sp.	3.09	34.29	96.2	0	272,530
Bacteroides ovatus	5.98	56.70	94.61	1.87	529,675
Blautia sp.	3.16	26.35	92.83	2.22	140,920

Table 2.1: Athena draft genome assemblies generated for sample A

Draft genome 1	Draft genome 2	Aligned bases	Percent identity	Number SNPs
Assembly A	Assembly C	4,965,009	99.98	371
Assembly C	Assembly D	5,050,613	99.91	3811
Bloodstream isolate	Assembly C	5,056,888	99.99	182
Bloodstream isolate	Assembly D	5,002,210	99.91	3742
E. coli strain S88 (NCBI)	Assembly C	4,410,742	98.61	56,513

Table 2.2: Comparison of *E. coli* strain similarities across time and spatial location

Category	Resistance Gene(s)
Beta-lactam resistance	CTX-M-27
Aminoglycoside resistance	kdpE
Polymyxin resistance	arnA, pmrC, pmrE, pmrF
Bacitracin resistance	bacA
Efflux pump complex or subunit	acrA, acrB, acrD, acrE, acrF, emrA, emrB, emrD, emrE, emrK, emrY, marA, mdfA, mdtA, mdtC, mdtE, mdtF, mdtG, mdtH, mdtM, mdtN, mdtO, mdtP, msbA, msrB, patA, TolC, YojI
Protein modulating antibiotic efflux	acrS, baeR, baeS, cpxA, CRP, emrR, evgA, evgS, gadW, gadX, H-NS

Table 2.3: Antibiotic resistance genes present in pre-transplant *E. coli* genome

Chapter 3

Acquisition, transmission and strain diversity of human gut-colonizing crAss-like phages

The work in this chapter was presented in:

Siranosian, B.A., Tamburini, F.B., Sherlock, G., and Bhatt, A.S. (2020). Acquisition, transmission and strain diversity of human gut-colonizing crAss-like phages. Nat Commun 11, 1-11.

3.1 Abstract

CrAss-like phages are double-stranded DNA viruses that are prevalent in human gut microbiomes. Here, we analyze gut metagenomic data from mother-infant pairs and patients undergoing fecal microbiota transplantation to evaluate the patterns of acquisition, transmission and strain diversity of crAss-like phages. We find that crAss-like phages are rarely detected at birth but are increasingly prevalent in the infant microbiome after one month of life. We observe nearly identical genomes in 50% of cases where the same crAss-like clade is detected in both the mother and the infant, suggesting vertical transmission. In cases of putative transmission of prototypical crAssphage (p-crAssphage), we find that a subset of strains present in the mother are detected in the infant, and that strain diversity in infants increases with time. Putative tail fiber proteins are enriched for nonsynonymous strain variation compared to other genes, suggesting a potential evolutionary benefit to maintaining strain diversity in specific genes. Finally, we show that p-crAssphage can be acquired through fecal microbiota transplantation.

3.2 Introduction

In addition to trillions of bacteria, the human gastrointestinal tract is densely populated with bacteriophages. Bacteriophages can drive bacterial community composition and mediate horizontal gene transfer[177], and alterations in the human gut virome have been associated with disease[26, 202]. However, our knowledge of the contributions of specific bacteriophages to human biology is limited, in part due to the paucity of viral sequences represented in reference databases. High-throughput sequencing and advanced genomic tools have facilitated the in silico discovery and characterization of previously unknown bacteriophages. The preeminent example of such a discovery is crAssphage (cross-Assembly phage), initially identified from human virome sequencing data[44]. A bacteriophage with a 97 kilobase circular, double-stranded DNA genome, crAssphage sequences are found almost exclusively in human fecal metagenomes in diverse populations globally[59, 66, 106, 28], and can be highly abundant. Initial estimates indicate that crAssphage is present in up to 73-77% of humans[44, 59]. Given the near ubiquity of crAssphage and its apparent specificity to the human gut, quantitative PCR assays have been developed to use crAssphage genes as markers for tracking human fecal pollution in water and environmental samples[166, 91] and in human stool[30].

More recent investigations have shown that crAssphage is one member of a wide range of crAss-like phages that exist in the human microbiome[59, 199]. In this manuscript, we adopt the taxonomic classification system for crAss-like phages used in Guerin et al.[59], which proposed 4 subfamily (Alpha, Beta, Gamma, Delta) and 10 cluster (1-10) designations based on shared protein coding genes. The crAssphage first described by Dutilh et al.[44] belongs to the Alpha subfamily, cluster 1 and is given the designation prototypical crAssphage (p-crAssphage); we use p-crAssphage in all further designations to avoid ambiguity. The genomes classified as crAss-like phages by Guerin et al.[59] are diverse - members of the same cluster share at least 40% of protein coding genes, while members of the same subfamily share only 20-40% of protein coding genes.

It is not known whether or how crAss-like phages influence host biology or disease [92, 45]. To answer higher-order questions about the role of crAss-like phages in human biology, it is necessary to establish basic principles of acquisition, persistence, and distribution. Although p-crAssphage has been detected in infant gut metagenomes [106, 92], it is not yet known how crAss-like phages are acquired in infancy. Infants acquire many of their first microbes, such as Bacteroides species, from their mother during and after delivery[13, 195, 50, 40]. By contrast, it has been demonstrated that adult twins and their mothers have unique gut viromes [140]. Given that Bacteroides species are hypothesized to be the bacterial host(s) of p-crAssphage[44, 154], and the apparent specificity of p-crAssphage to the human gut as opposed to other mammals or environmental samples, we postulated that p-crAssphage is vertically transmitted from mother to infant, similar to what is observed for many bacterial taxa and in contrast to what is reported for other members of the human virome. To test this hypothesis, we examined publicly available shotgun metagenomic data from two stool microbiome datasets15,16 consisting of samples from mothers and their infants (n =

143 mother-infant pairs).

In this study, we find that p-crAssphage and other crAss-like phages are rarely detected in the gut microbiome at birth but become detectable during the first year of life. We observe >99.5% identical genome sequences in one half of cases where mothers and infants have the same crAss-like phage, suggesting vertical transmission from mother to infant. Infants acquire a reduced diversity population of p-crAssphage compared to their mother, but strain diversity expands upon colonization. Finally, by examining shotgun metagenomic data from patients undergoing fecal microbiota transplantation (FMT), we show that FMT recipients can acquire p-crAssphage with a nearly identical genome sequence as the stool donor. These results begin to uncover the principles of acquisition and transmission of p-crAssphage and other crAss-like phages, which are the most prevalent human-associated phages described, to date.

3.3 Results

3.3.1 Presence of p-crAssphage in mother and infant microbiomes.

We evaluated the presence and abundance of p-crAssphage in the microbiomes of mothers and infants by classifying sequencing reads with Kraken2[190], using a database of all bacterial, viral and fungal genomes in NCBI GenBank assembled to complete genome, chromosome or scaffold quality as of February 2019 (see Methods). P-crAssphage is represented by a 97-kb genome (accession NC_024711.1). Assigning absolute presence or absence of an organism in metagenomic sequencing data is difficult and confounded by sequencing depth. Here, we consider samples with ≥1,000 reads classified as p-crAssphage to be evidence for presence, as this corresponds to roughly 1x coverage of the genome (assuming 100 bp reads and a 100-kb genome length). Samples from mothers and infants had an average of 8.7 M reads after preprocessing, and the 1000 read coverage threshold thus corresponds to an average relative abundance of 0.011%. Of note, this somewhat arbitrary threshold, while fairly specific, renders our approach limited in sensitivity - that is, we do not report on p-crAssphage when it is present at lower relative abundance.

Although p-crAssphage is highly abundant in the adult gut microbiome[59, 45] and has been detected in infant gut microbiomes, it is unclear when or how it is acquired. Consistent with previous studies describing low relative abundance or absence of p-crAssphage in the infant gut microbiome[106, 93], we found 0 out of 22 infants have ≥1000 p-crAssphage reads in samples collected within 24 h of birth (Supplementary Data 1). P-crAssphage increases in prevalence as infants age: it is detected in 3/35 (9%) infants from Yassour et al.[195] by three months and 16/100 (16%) infants in Bäckhed et al.[13] by 12 months. P-crAssphage is more prevalent in adult mothers, where it is detected in at least one sample from 8/35 (23%) and 25/100 (25%) of mothers in each study, respectively (Supplementary Fig. 1). P-crAssphage is detected in both the mother and her infant in ten cases, while 23/33 p-crAssphage positive mothers have a p-crAssphage negative infant, and

9/19 p-crAssphage positive infants have a p-crAssphage negative mother.

The infant gut microbiome is strongly impacted by delivery mode, and it has been shown that infants born by Cesarean section initially lack Bacteroides species[195, 40, 68, 151, 96]. In line with the described effects of Cesarean section delivery and a hypothesized Bacteroidetes host, we found that all 19 p-crAssphage positive infants were delivered vaginally, while all 21 infants delivered through Cesarean section remained p-crAssphage negative. Although not significant when each study is tested individually, the association between delivery mode and p-crAssphage presence is significant when samples from the two studies are combined ($p = 0.043$, Fisher's exact test). This result contrasts with the findings of McCann et al.[106], where the authors found no association between p-crAssphage relative abundance and delivery mode.

3.3.2 Putative vertical transmission of p-crAssphage.

Vertical transmission of gut microbes from mother to infant is common and well-described among certain bacterial taxa[13, 195, 50, 10] . To test the hypothesis that p-crAssphage can be vertically transmitted from mother to infant, we investigated metagenome-assembled p-crAssphage genomes from ten p-crAssphage positive mother-infant pairs (see Methods). In six cases, mother-infant pairs had nearly identical assembled sequences (families M0226, M0808, M1098, 335, 343, and 345; >99.7% similarity). Two mother-infant pairs had assembled sequences more similar than unrelated pairs (families 263 and 268; 98-99.2% similarity), and two mother-infant pairs had assembled sequences that were no more similar than unrelated pairs (families 184 and 272; 96% similarity) (Fig. 1). Overall, related mothers and infants harbor more closely related p-crAssphage sequences than unrelated mothers and infants (Supplementary Fig. 2a). When all samples with sufficient p-crAssphage coverage were included in the assembly comparison, no pairs from unrelated individuals had >98% similarity (Supplementary Fig. 3, Supplementary Data 3). Assembled p-crAssphage genomes were high quality and contiguous: in the 29 samples from families with p-crAssphage found in mothers and infants, the median contig N50 was 59.6 kb (standard deviation, SD = 41.4 kb), median number of contigs was 3 (SD = 17) and median total assembled length was 96.3 kb (SD = 24.2 kb) (assembly statistics reported in Supplementary Data 2). One-to-one pairwise alignments between assembled sequences from mother-infant pairs had a median length of 85.8 kb (SD = 27.8 kb). In all, 22 samples assembled a nearly complete (>95 kb) p-crAssphage genome in a single contig. The assembled genomes also share 91.2-97.6% nucleotide identity with the p-crAssphage reference genome, adding confidence that they are truly representative assemblies. Next, we used a variant calling approach to identify fixed SNPs compared to the p-crAssphage reference (see Methods). Pairs of samples were compared at genomic sites covered $\geq$10x and used to construct a heatmap of SNP similarity. The same six mother-infant pairs (M0226, M0808, M1098, 335, 343, and 345) had >99.5% SNP similarity and continued to cluster together (Supplementary Fig. 4). Mother-infant pairs had higher SNP identity than unrelated pairs on average (Supplementary Fig. 2b).

3.3.3 Strain diversity in the p-crAssphage population.

Metagenomic assembly only represents the dominant allele at each position, and a fixed SNP comparison only considers sites that are identical across all strains present in a sample. To understand the differing p-crAssphage strains present in the microbiome, we would ideally phase strain "haplotypes" with a technology like long-read sequencing. With only short reads available, we examined genomic positions that had multiple single nucleotide variant alleles called at high-quality ($\geq$5 reads for each allele, multiallelic sites) as a proxy for strain diversity. We report a normalized statistic (Fmulti) to compare multiallelic sites across samples with highly variable coverage. At a given minor allele fraction (AF), Fmulti is the proportion of multiallelic sites with a minor AF $>$ x among those sites covered well enough to detect a minor AF of x. We calculated Fmulti for minor allele fractions of 0.40, 0.30, 0.20, and 0.10 and compared across samples at a given minor AF value. At every AF tested, infants as a group had a smaller fraction of multiallelic sites when compared to mothers as a group. (Fig. 2a). In the three cases where we detected p-crAssphage in multiple samples from the same infant, we found more multiallelic sites in later samples. We were typically powered enough to detect multiallelic sites of the observed AF in earlier samples, but we cannot rule out the possibility that newly observed variants are below our limit of detection in earlier samples. Multiallelic sites in infants are often fixed sites in the p-crAssphage population of the mother (Fig. 2d, Supplemental Fig. 10). In contrast to infants, mothers from Yassour et al.[195] showed no change in the proportion of multiallelic sites over the 6 month sampling period (Fig. 2b). We then looked at multiallelic sites in mothers that are fixed in matched infant samples. In 2/3 cases where we observed putative p-crAssphage transmission and the mother had $\geq$10 multiallelic sites, major alleles are disproportionately detected in the child (Fig. 2c). Taken together, these results suggest a few potential models: one is that infants acquire a single strain or limited diversity of p-crAssphage strains. As the infant microbiome matures and diversifies with age, the p-crAssphage population can evolve and acquire genetic diversity. Alternatively, a larger spectrum of p-crAssphage strains than is detected may be harbored in a mother, with some strains below the limit of our detection. These strains may experience less selective pressure in the infant than in the mother; thus, these strains may be more numerous and easily detected in the infant than they are in the mother.

We continued to use multiallelic sites to investigate p-crAssphage strain diversity within an individual. Mothers generally have limited strain diversity, with a median of 0.41 (SD = 7.5) multiallelic sites per kb at AF $>$ 0.1; variation in 0.04% of the genome. In a small number of cases, we do observe samples with up to 100x more frequent multiallelic sites (Fig. 3a). This allelic variation could be the result of many closely related strains existing together or a smaller number of more divergent strains. Phasing strain "haplotypes" is necessary to distinguish between these possibilities. Limited strain diversity suggest an exclusion principle, which favors mono- or oligo-colonization of a particular p-crAssphage strain or closely related strains within the gut of an individual, though

notably, a minority of individuals may be simultaneously colonized by multiple diverse strains. P-crAssphage has fewer multiallelic sites on average compared to Lactococcus phages, which are the only other group of phages detected at ≥1x coverage in at least ten samples. Using the reference genome of the most frequently detected individual phage, Lactococcus phage 16802, 34 samples had at least 1x

coverage; these samples had a median of 58.8 (SD = 14.7) multiallelic sites per kb at AF > 0.1 (Supplementary Fig. 9). We did not find assembled Lactococcus phage genomes with >99% similarity between any samples from different individuals.

We next evaluated whether strain variation in the p- crAssphage population was the result of synonymous or nonsynonymous genomic changes. Variant effects were predicted using the p-crAssphage genome annotation from GenBank and SnpEff[31] (see Methods). We compared the proportion of observed variant effects to a null model of equal probability of mutation at every base in the reference genome. In samples from mothers, multiallelic sites with predicted nonsynonymous and nonsense effects were less likely than expected under the null model, while synonymous sites were more likely than expected ($p < 1e\text{-}5$ for each category, likelihood ratio test; Fig. 3b). An overrepresentation of synonymous variants suggests that strain diversity in the p-crAssphage population of mothers is enriched for neutral genetic variation, which may have been acquired over the relatively long time the phages could have been present in the microbiome. In contrast to mothers, the predicted effects of multiallelic sites in infants were indistinguishable from the null model ($p > 0.05$ for each effect category, likelihood ratio test). This may suggest that multiallelic sites in infants arise randomly and the forces acting to influence the distribution of predicted sites in mothers have not had time to act on the infant's p-crAssphage population yet. Alternatively, selective pressures acting on p-crAssphage alleles may be entirely different in the infant and mother microbiomes. Comparing mother and infant distributions showed that only the proportion of synonymous multiallelic sites was significantly different between the two ($p = 0.04$, likelihood ratio test). We note that synonymous variants may not be truly neutral, as noncoding variants have been shown to affect translation efficiency in bacteriophages[57].

In adults, we find wide variation in the number of multiallelic sites across the p-crAssphage genome, with enrichment in the number of sites and the ratio of nonsynonymous to synonymous variants corresponding to certain predicted genes (Fig. 4). Multi- allelic sites were detected in 80/88 predicted genes. When genes were ranked by the length-normalized number of nonsynonymous variants, "putative Tail sheath protein" was the top annotated gene, and other predicted tail proteins also had high ratios (Supplemen- tary Data 4). Phage tail proteins are responsible for host tropism;[55, 17, 145] therefore maintaining multiple functionally different alleles in the population may be beneficial to expand the host range of p-crAssphage. Increased variation in tail fiber genes was also found in an analysis of p-crAssphage genomes from South Africa[23]. The genes that were least likely to have nonsynonymous multiallelic sites appear to be those that are critical for phage

function, such as "putative portal protein", "putative major capsid protein" and putative RNA polymerase subunits. Interestingly, some of these genes correspond to peaks in the number of multiallelic sites detected, even though the variants had mostly synonymous predicted effects. Infant samples have fewer multiallelic sites than mothers, with sites detected in 33/88 genes. "Putative ssb single stranded DNA-binding protein" was the most enriched gene for nonsynonymous multiallelic sites. Five out of ten tail fiber proteins had at least one nonsynonymous multiallelic variant in infant samples, and the most frequently mutated tail fiber gene was "putative phage tail-collar fiber protein (DUF3751)," a top hit in the adult samples (Supplementary Data 4).

3.3.4 Acquisition and transmission of crAss-like phages.

P-crAssphage is the first described member of an expanding group of crAss-like phages. Guerin et al.[59] assembled 249 complete or near-complete crAss-like phage genomes out of metagenomic sequencing datasets, which were then classified into four subfamilies (Alpha, Beta, Gamma, Delta) and 10 clusters (1–10) based on shared protein coding genes. P-crAssphage is a member of cluster Alpha 01. Given the observed sharing of p-crAssphage genome sequences by mother-infant pairs, we were interested to determine if similar putative transmission events could also be observed for crAss-like phages. We added the crAss-like genomes from Guerin et al.[59] to the Kraken2 viral reference database in a hierarchy following the pro- posed subfamily and cluster designations (see Methods). For clas- sification and transmission analyses, we carried out analyses at the level of crAss-like phage clusters. A threshold of 1000 reads clas- sified to the same cluster (roughy 1x coverage) was treated as evidence for presence.

Broadly, crAss-like phages are more frequently detected in the microbiome of mothers and infants than p-crAssphage alone. In total, 7/36 (19%) infants from Yassour et al.[195] and 49/100 (49%) infants from Bäckhed et al.[13] have at least one crAss-like cluster detected in at least one sample. At least one cluster was detected in 33/43 (77%) and 88/100 (88%) of mothers from each study. Mothers are most likely to be colonized by a single crAss-like phage cluster, although we observe samples with up to 8 clusters detected (Supplementary Fig. 5). We do not observe any crAss-like phage cluster present in infant samples taken within 24 h of birth. However, two infants from Bäckhed et al.[13] have a crAss- like phage meeting the presence threshold in samples collected as soon as 3 days after birth (samples 385_B and 633_B). This may represent the lower time limit for the crAss-like phage and its bacterial hosts(s) to reach the detection threshold. The putative hosts of crAss-like phages, members of the Bacteroidetes phylum, are known to be vertically transmitted[13, 195, 50, 10]. Thus, we hypothesized that crAss-like phages would be more frequently transmitted to vaginally vs. Cesearan section born infants. We observed that all 19 p-crAssphage positive infants were delivered vaginally, while all Cesarean section born infants remained p-crAssphage negative for the duration of sampling. Although low sample numbers prevented this association from raising to the level of significance when each cohort was tested individually, it was significant when samples

from both cohorts were considered together (Bäckhed p = 0.12, Yassour p = 1, combined p = 0.043, Fisher's exact test). We also tested for associations in cases where at least 10 infants were positive for a given crAss-like phage cluster. Presence of cluster Delta 07 (Bäckhed p = 0.009, Yassour p = 1, combined p = 0.007, Fisher's exact test) and any crAss-like phage cluster (Bäckhed p = 0.004, Yassour p = 0.32, combined p = 0.001, Fisher's exact test) was significantly associated with vaginal delivery (Supplementary Data 1). P-values are uncorrected for multiple hypothesis testing. No significant associations between crAss-like phage presence and breastfeeding status were found.

Although a crAss-like phage similar to cluster Beta 06 phages was recently cultured on a *Bacteroides intestinalis* host[154], the hosts of other crAss-like phages have yet to be identified. We searched for bacterial taxa that were differentially abundant between crAss-like phage positive and negative infants to make inferences about potential hosts. Samples from vaginally born infants at three or four months of age were included to allow comparisons across the two studies at a similar time point. Bacterial relative abundances were transformed to centered log-ratios, and differential abundance was calculated with the R package ALDEx2[49] (see Methods). P-values were calculated with the two-sided Wilcoxon rank-sum test and corrected for multiple hypothesis testing34. Due to limited sample numbers, we considered presence of any crAss-like phage as a group.

At the genus level, Collinsella was the most enriched taxon in infants positive for any crAss-like phage (corrected p = 0.0167, twosided Wilcoxon test) (Supplementary Fig. 6, Supplementary Data 6). Several members of the Collinsella genus, including *Collinsella aerofaciens* (corrected p = 0.0239, two-sided Wilcoxon test) were also the most enriched species in these infants. Certain species of the genus Bacteroides were also significantly enriched to a lesser degree, such as *Bacteroides massiliensis* (corrected p = 0.0499, twosided Wilcoxon test). Collinsella is a member of Actinobacteria, an entirely different phylum than the posited Bacteroidetes hosts[44]. Collinsella was previously identified as a signature of the developing anaerobic infant microbiome[13] , but further work is necessary to determine if these species have a direct or indirect influence on the acquisition of crAss-like phages.

Next, we searched for putative mother-infant transmission of crAss-like phages. Depending on the cluster, 0–16 families (median 2.5, SD = 5.1) have the same crAss-like phage cluster detected in at least one mother and matched infant sample (Supplementary Data 5). We extended the metagenomic assembly analysis above, using a pangenome compiled from all crAss-like genomes in each cluster in place of the p-crAssphage reference. We find cases where mothers and infants share an assembled genome with >99.7% identity and at least 20 kb of aligned sequence in 7/10 clusters. As expected, candidate transmission events in cluster Alpha 01 matched p-crAssphage. Overall, we observe putative transmission in 21/42 (50%) cases where mother and infant have the same crAss-like phage cluster (Supplementary Data 5). In two families, we observed putative transmission of two separate crAss-like phage clusters. 50% is likely an underestimate of the true transmission rate, because many comparisons were limited by low sequencing depth and poorly assembled draft genomes. Interestingly, a subset of infants have a crAss-like phage but their mothers do not. This

could be due to waxing/waning amounts of crAss-like phages in mothers, as has been described in adults[45]. If this is the case, the crAss-like phage may have been present in the mother at a level lower than our limit of detection and thus may have been transmitted to the baby. Alternatively, the baby may have acquired the crAss-like phage from an altogether different source, such as another housemate or environmental source.

3.3.5 Similar p-crAssphage genomes found in FMT donors and recipients.

Another example of a perturbation where the gut microbiome, typically stable in adults, may acquire a large number of new microbes is when individuals experience infection with the gut pathogen *Clostridium difficile*, are treated with antibiotics, and subsequently receive fecal microbiota transplantation (FMT) [162, 111]. Draper et al.[41] found that p-crAssphage relative abundance is decreased in individuals with recurrent *C. difficile* infection and that p-crAssphage could be transplanted from donor to recipient. However, strain-level transmission of p-crAssphage has not been explored in this patient population. We examined metagenomic sequencing data from Smillie et al.[162] and viral metagenomic sequencing data from Draper et al.[41] using the classification, assembly and comparison methods described above.

In the data from Smillie et al.[162], we detect p-crAssphage at $\geq$1x coverage in samples from two donors, MGH06D and MGH03D (Supplementary Data 1). 12 patients received stool preparations from either of those two donors. After FMT, 8/11 (73%) patients who received material from donor MGH03D were positive for p-crAssphage, while the individual who received material from donor MGH06D remained negative (Fig. 5a). Zero recipients who received FMT from a p-crAssphage-negative donor acquired p-crAssphage during the sampling period. We compared assembled p-crAssphage genomes from donors and recipients and found >99.8% identical sequences in samples from MGH03D and recipients of this donor's material, while samples from donor MGH06D had a distinct p-crAssphage sequence (96.7% identity to MGH03D) (Supplementary Fig. 7). Interestingly, individuals MGH11R and MGH12R experienced dynamic p-crAssphage presence, with the phage falling below and rising above the detection limit in subsequent samples. The assembled genomes remained highly similar in each case, suggesting a waxing/waning p-crAssphage population in individual MGH12R, who did not receive an additional FMT. In pre-FMT samples from recipients from this study, p-crAssphage was not detected, and only one sample was positive for a single crAss-like phage (Supplementary Data 1). This suggests the phages are substantially diminished in abundance when individuals are treated with drugs such as metronidazole, which has high activity against Bacteroides species.

Draper et al.[41] specifically sequenced the amplified viral content of the metagenome, so we adjusted the detection threshold to 10,000 reads (10x coverage) to reduce the number of false positives. P-crAssphage was detected in all 16 samples from donor D3 and 0/17 samples from donors D1 and D2. No pre-FMT samples from C. difficile colitis affected patients were positive for p-crAssphage

or any crAss-like phage (Supplementary Data 1). In total, 7/7 patients who received FMT from donor D3 material became p-crAssphage positive; most remained positive for the 12 month duration of sampling (Fig. 5b). Assembled p-crAssphage genomes from donor D3 and the seven recipients had >99.5% nucleotide identity, suggesting colonization with the specific donor p-crAssphage strain (Supplementary Fig. 8). Patient P7 became p-crAssphage positive with a genome 92% identical to the other donor and patients. P7 received material from donor D1, who was p-crAssphage negative, and therefore could have acquired the phage from a population below the detection limit in the donor or another source following reestablishment of host bacterial populations. Patient P13 had p-crAssphage present at 40x in a single sample, but the assembled genome only had a total length of 12 kb and N50 of 2.8 kb. Samples with lower coverage assembled nearly complete p-crAssphage genomes with N50 > 60 kb. It is thus possible that P13's p-crAssphage detection is an artifact of PCR amplification that is often used in the sequencing of virus-enriched samples. These data show that p-crAssphage is frequently and efficiently transplanted via FMT and that p-crAssphage can stably engraft in FMT recipients for up to one year.

3.4 Discussion

The in silico discovery of p-crAssphage and recent publication of hundreds of crAss-like phage genomes has highlighted the diversity and global prevalence of these phages in human gut microbiomes[153]. CrAss-like phages have even been found in non- human primates[45], suggesting these phages have been evolving alongside humans for millions of years. However, it is currently unknown when and how an individual typically acquires crAss- like phages, as well as what level of strain diversity exists within the microbiome of an individual. The datasets examined here[13, 195] contain mother-infant pairs sampled extensively during the first year of life and represent a unique opportunity to answer these questions.

We first characterized p-crAssphage and found no samples collected from infants within 24 h of birth met our 1x coverage threshold. P-crAssphage becomes increasingly prevalent as infants age, but does not reach the levels found in mothers by one year of life. The host(s) of p-crAssphage may not be present or have reached sufficient abundance in some infants by the end of sample collection. Infants acquire many of their gut bacteria through direct transmission from their mother, while gut viromes have been shown to remain unique between family members and twins[140]. In contrast to other members of the gut virome, we found nearly identical assembled p-crAssphage genomes in 6/10 cases where mothers and infants both harbor the phage, suggesting vertical transmission. However, we cannot rule out alternative possibilities, such as transmission from a different family member or from a common environmental source. We also observed cases where mothers and infants had unrelated p-crAssphage genomes and cases where infants had p-crAssphage but it was undetectable in the mother, which argue that the infants acquired p-crAssphage from an undetectable population

in the mother or from another source.

It is currently unknown if individuals are typically colonized by a single or multiple p-crAssphage strains, or how similar or different these strains may be. We characterized the strain diversity of the p-crAssphage population in an individual by examining positions in the p-crAssphage genome where we detected mul- tiple high-quality alleles. We found most mothers have a limited number of variable sites, with a median frequency of 0.04% across the 97-kb p-crAssphage genome, arguing that most mothers have a limited diversity of p-crAssphage strains. We did observe one mother with 100x more frequent variable sites, however. Infants generally have an order of magnitude fewer variable sites than mothers, suggesting a population that is further reduced in strain diversity, which may be the result of a bottleneck event upon acquisition or transmission. P-crAssphage is significantly less diverse than the second most abundant phage in these samples, Lactococcus phages 16802, where variable sites are detected with a median frequency of 5.9%. In cases where we observed putative mother-infant transmission, major alleles in the mother are pri- marily found in the infant, suggesting the mother's dominant strain is primarily responsible for colonizing the infant. The p-crAssphage population in infants develops additional variable sites over time, often at positions where only single alleles were detected in the mother. This could be due to the different bac- terial hosts, nutritional sources and selective pressures in the infant microbiome, or simply due to random mutations.

In the most reductionist sense, two p-crAssphage genomes could differ at a single position and be considered different strains. However, we are most interested in strain variation that has functional consequences for the phage, its host or other members of the gut microbiome. Strain diversity in the p-crAssphage population of mothers is enriched for variants with predicted synonymous effects. However, we do observe enrich- ment for nonsynonymous (i.e. functional) variants in key genes, including predicted tail fiber proteins. This suggests that there may be a benefit to maintaining nonsynonymous allelic diversity in these genes, such as the ability to infect a broader range of hosts. One isolated crAss-like phage[154] was noted to have a very specific host range, so variation in tail fiber genes may allow these phage to infect an increased range of bacteria. Laboratory experiments are necessary to further investigate this hypothesis, but could use exiting variation in the tail genes as a starting point to screen for expanded host range.

P-crAssphage in infants has variable sites that are enriched for synonymous changes com- pared to mothers, but limited sample numbers made it difficult to determine enrichment for specific genes. P-crAssphage is the first described member of a diverse group of crAss-like phages5, with four "family" level and ten "genus" level classifications. Similar to p-crAssphage, we observe a trend of increasing prevalence with infant age for many clusters of crAss-like phages. Some clusters, such as Alpha 03, are prevalent in mothers but rarely or never observed in infants, suggesting the hosts of these phages have yet to reach sufficient abundance in the infant microbiome. We first observe a crAss-like phage at 1x coverage in samples collected three days after birth. In the case of family 633, the mother and three-day old infant have a Delta 07 phage with 99.3% alignment identity. Since

we did not observe such early potential transmission events with p-crAssphage, this may represent the first detectable transmission of any crAss-like phage from mother to infant, and a lower limit for the time for a crAss-like phage to colonize the infant microbiome. Alternatively, crAss-like phages may not colonize the infant microbiome at such an early time, rather, they may be acquired through routes other than actual parturition. For example, the phages might be present in yet understudied niches, such as the mother's breast milk or the shared built environment of the baby and mother. Of note, Bacteroides species, which are posited to be the natural host of p- crAssphage and are known to be the host of a crAss-like phage[154] have previously been detected in breast milk[71]. Overall, we find nearly identical genomes in 50% of cases when we detect the same cluster crAss-like phage in both mother and infant, suggesting a transmission rate similar to p-crAssphage.

Regardless of the crAssphage status of the mother, we found a strong association of p-crAssphage and crAss-like phage presence with vaginal delivery, in contrast to what has been described previously[106]. One potential explanation is that vaginal birth is responsible for transmitting the phage from mother to infant. However, this is less likely in cases where infants harbor a phage undetected in the mother. Another possible explanation is that vaginal birth is responsible for seeding bacteria necessary for later colonization by crAss-like phages. Previous research found maternal seeding of bacteria from the class Bacteroidia was inhibited by C-section birth, supporting this hypothesis[151, 79]. Future research with more balanced cohorts will likely clarify whether or not birth mode affects crAss-like phage acquisition and transmission. Unexpectedly, microbiomes of vaginally born infants positive for crAss-like phages were strongly enriched in Collinsella species. It is doubtful that this finding suggests new hosts for crAss-like phages, rather, Collinsella may be a hallmark for a developing and increasingly anaerobic infant microbiome that is capable of harboring these phages.

Finally, we observe that p-crAssphage is frequently transmitted via fecal microbiota transplantation (FMT) and can engraft stably in FMT recipients for up to one year. Engraftment of bacteria and phages has been well-studied in the case of FMT treatment for recurrent Clostridium difficile infection, and transplantation of p- crAssphage has been identified previously[162, 111, 41]. Our strain-level findings add new insight into the transmission of lytic bacter- iophages. We assembled nearly identical genomes from both donors and recipients, highly suggestive of transmission of the specific p-crAssphage strain. Taken together, the results from both populations suggest that infants and patients receiving FMT have relatively unpopulated, naive microbiomes, providing an open niche for p-crAssphage to engraft into.

While this study suggests new principles about acquisition and transmission of crAss-like phages in the gut microbiome, it does have several limitations. First, we examined publicly available metagenomic data and were therefore limited to the available study cohort and sample size. In the mother-infant studies, stool samples were not collected from family members other than mothers, which could help determine other contributions to crAss-like phage acquisition in infants. Also,

many infant birth samples were limited by low sequencing depth. As such, the estimates for acquisition and transmission presented here are likely underestimates. Sampling time points, processing techni- ques and study populations were different between the two stu- dies, although both were conducted in Northern European individuals. Second, short-read sequencing data limited our ability to phase strain variants in the p-crAssphage genome. If the samples were resequenced with long-read sequencing approa- ches41 , we could obtain single reads spanning many variable sites. This would allow us to determine if the observed variants are the result of a smaller number of more divergent strain populations, or a high number of closely related strains. Finally, our group has become aware of false positive strain sharing results due to "barcode swapping" in dual-indexed Illumina sequencing librar- ies generated in our lab, which was first described in 2017[107]. As the indexing strategy was not reported for the public data we analyzed in this manuscript, we cannot be certain that the find- ings presented are not the result of this artifact. However, we believe our results, where only matched mother-infant pairs and matched FMT donor-recipient pairs share highly related crAss-like phage sequences, are unlikely to be explained by barcode swapping alone. The detrimental effect of barcode swapping also highlights the importance of reporting index sequences as a key part of making data publicly available.

Future work expanding on our findings should be directed towards answering several important questions. How stable are crAss-like phages transmitted from mother to infant over time? Are they lifelong inhabitants that, barring heavy antibiotic use, can be transmitted for generations? Are there exclusion principles that prevent the acquisition of a second, more divergent p-crAssphage strain? Additionally, our strain diversity analysis focused on p-crAssphage, but a wealth of diversity is also present in crAss-like phages. Better genome annotations and more con- crete principles surrounding the identity and taxonomy of crAss- like phages will enable this research, and isolating, culturing and characterizing new crAss-like phages is a key next step. Finally, long-read metagenomic sequencing will enable better analysis of the strain populations among crAss-like phages in the mixed community of the microbiome. The ubiquity, distinct genome composition and ease of computational analysis with crAss-like phages may render them useful models for querying microbial transmission more broadly. Future work remains to determine precisely whether, and how, crAss-like phages influence the gut ecosystem and ultimately human health.

3.5 Methods

Sequence read preprocessing. Raw sequencing reads from Bäckhed et al.[13] and Yassour et al.[195] were downloaded from SRA from each sample and preprocessed in a consistent way: TrimGalore version 0.5.0[81] was used to perform quality and adapter trimming with the flags "-clip_R1 15-clip_R2 15-length 60". SeqKit version 0.9.1[152] was used to remove duplicates with the command

"seqkit rmdup-by-seq". Reads were mapped against the human genome using BWA version 0.7.17-r1188[87] and only unmapped reads were retained. Many infant samples at birth had low read counts after preprocessing, and samples with fewer than 10,000 reads were removed from all subsequent analyses. This left 135 families with sufficient depth in at least one sample from mother and infant.

3.5.1 Kraken2 classification.

For classification of p-crAssphage, we built a Kraken2[190] database containing all bacteria, viral and fungal genomes in NCBI GenBank assembled to complete genome, chromosome or scaffold quality as of February 2019. Human and mouse reference genomes were also included in the database. A Bracken[100] database was also built with a read length of 150 and k-mer length of 35. P-crAssphage is represented by a 97-kb genome (accession NC_024711.1). Multiple crAss-like phages are present in GenBank and would cause reads mapping to multiple genomes to be classified at the least common ancestor of "crAss-like viruses." To prevent this from happening, other crAss-like genomes were removed from the database.

For classification of crAss-like phages, we added to the viral database, replacing the original "crAss-like viruses" clade with genomes in the proposed subfamily and cluster hierarchy described in Guerin et al.[59]. Kraken2 was used with default classification parameters on paired-end reads. For testing associations between crAss-like phage presence and other bacterial taxa, reads were classified by using Kraken2 with default parameters on paired-end reads, and Bracken was used for abundance estimation with the parameters "-r 150 -l S -t 10".

3.5.2 Assembling and comparing crAss-like phage genomes.

Preprocessed sequencing reads were assembled with SPAdes version 3.13.1[117] using the '-meta' flag. Contigs $\geq$500 bp were aligned with BWA against either the p-crAsspahge reference genome or composite genomes from all the crAss-like phages in a cluster from Guerin et al.[59]. Resulting contigs were assessed for their N50 and total assembly length. Pairwise comparisons were conducted with nucmer version 4.0.0beta2[102] and the average identity and total length of 1-1 aligned segments was reported. The heatmap in Fig. 1 was clustered on the euclidean distance between samples with the ward.d2 clustering method and plotted with the heatmap.2 function in the gplots package for R[183].

3.5.3 SNPs and multiallelic sites.

SNPs were called with Snippy[147] with freebayes[53] as the variant caller using the p-crAssphage reference at sites covered $\geq$10x. Filtering, decomposition and normalization of variants was necessary to compare between samples and was conducted with vt version 0.5[169] and bcftools version 1.9[89]. The output of snippy, snps.raw.vcf, was used in this command: "vt decompose -s snps. raw.vcf —

vt decompose_blocksub -a - — bcftools norm -f crassphage_reference.fasta -m -any — bcftools view-include 'QUAL ≥ 100 && FMT/DP ≥ 10 && (FMT/AO)/ (FMT/DP) ≥ 0'". We calculated the transition/transversion ratio of detected variants using vcftools[34] version 0.1.16. Considering all detected variants agnostic of samples, the transition/transversion ratio is 2.91 for all SNPs, 2.41 for fixed SNPs after variant decomposition and 2.42 for multiallelic SNPs at >0.1 AF after variant decomposition. Considering samples individually, the median Ts/Tv ratios are 3.40, 2.80 and 3.23 (SD = 2.3, 1.5, 3.4), respectively. The median sample numbers are higher because of samples with few detected transversions producing a comparatively high ratio. To compare fixed SNPs between samples, we only consider sites covered ≥10x in both samples. The reported SNP % identity is 1 - (the number of fixed SNPs different between samples/number of sites covered ≥10x in both samples). Multiallelic sites were called as sites with two alleles and ≥5 reads supporting each allele. We report a normalized statistic (Fmulti) to compare multiallelic sites across samples with highly variable coverage. At a given minor allele fraction(AF), Fmulti is the proportion of multiallelic sites with a minor AF > x among those sites covered well enough to detect a minor AF of x. Effects of multiallelic variants were predicted with SnpEff version 4.3[31] using the p-crAssphage genome annotation available on GenBank and the flags "ann -noLog -noStats -no-downstream -noupstream -no-utr -t". When mothers had multiple samples, we used the one with the highest p-crAssphage coverage for multiallelic site analysis.

3.5.4 CrAss-like phage correlation with bacterial abundance.

We used the outputs of Bracken to test for differential abundances in taxa between groups. Matrices of reads classified to each taxon in each sample were filtered to keep only taxa with an abundance of at least 0.001 and nonzero values in at least 30% of samples. Zeros in the data were replaced with the Geometric Bayesian multiplicative method in the zCompositions version 1.3.2-1 package for R[127]. Differential abundance between groups was calculated with the ALDEx2 version 1.16.0 package for R[49].

3.5.5 FMT data analysis.

Data from the FMT studies [162, 41] were processed, assembled and compared in the same way as mother/infant data. Sample MGH06R35 was excluded from the FMT cohort analysis as it could not be definitively determined whether sample designated as pre-FMT was actually collected prior to transplantation (personal communication with authors).

3.6 Figures

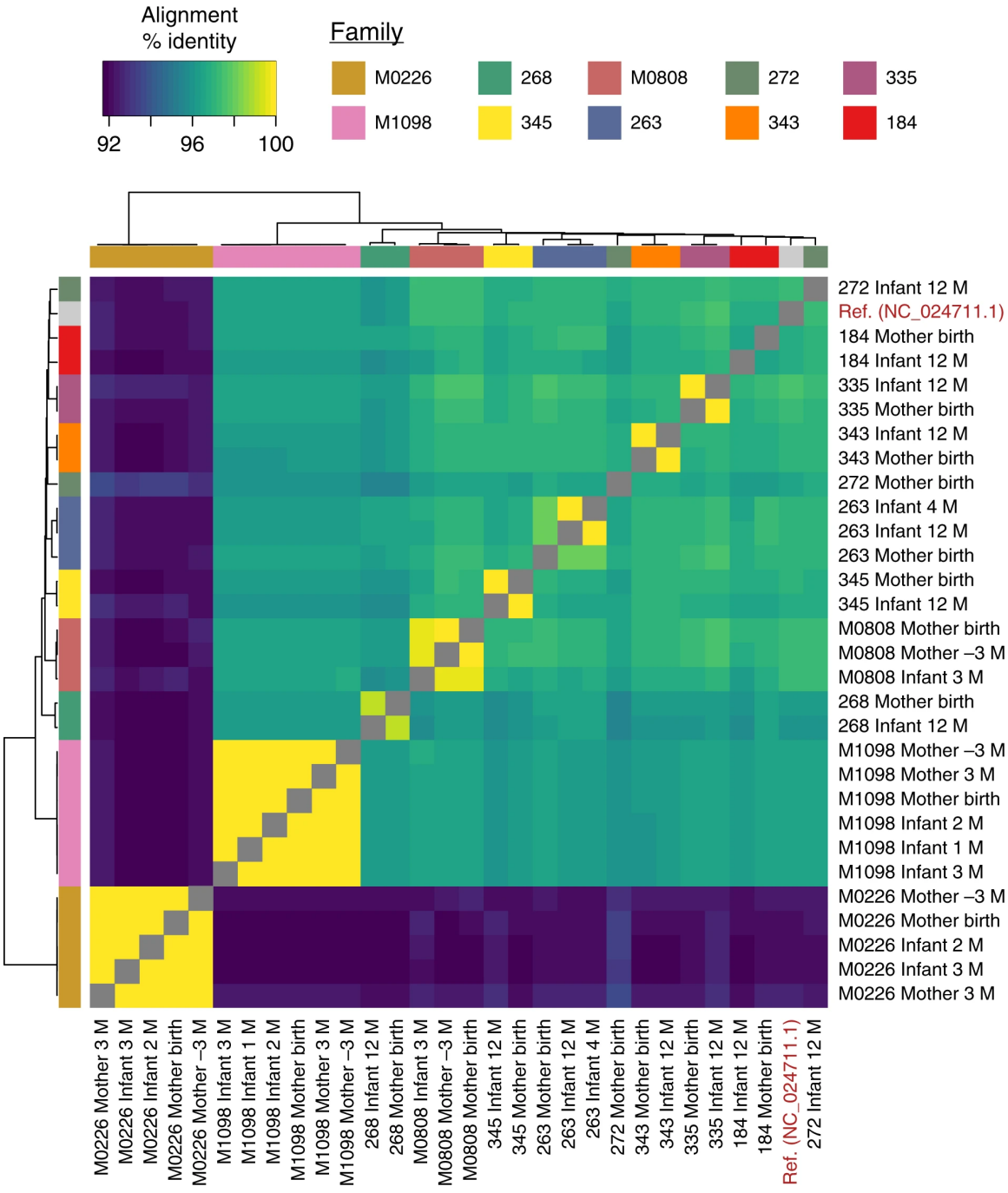


Figure 3.1: **Mother-infant pairs share > 99.7% similar p-crAssphage genomes in 6/10 cases.** Heatmap of pairwise alignment percentage identity of metagenome-assembled p-crAssphage genomes from mothers and infants. Only families with p-crAssphage detected in at least one mother and infant sample are shown. The p-crAssphage reference genome is also included as a comparison.

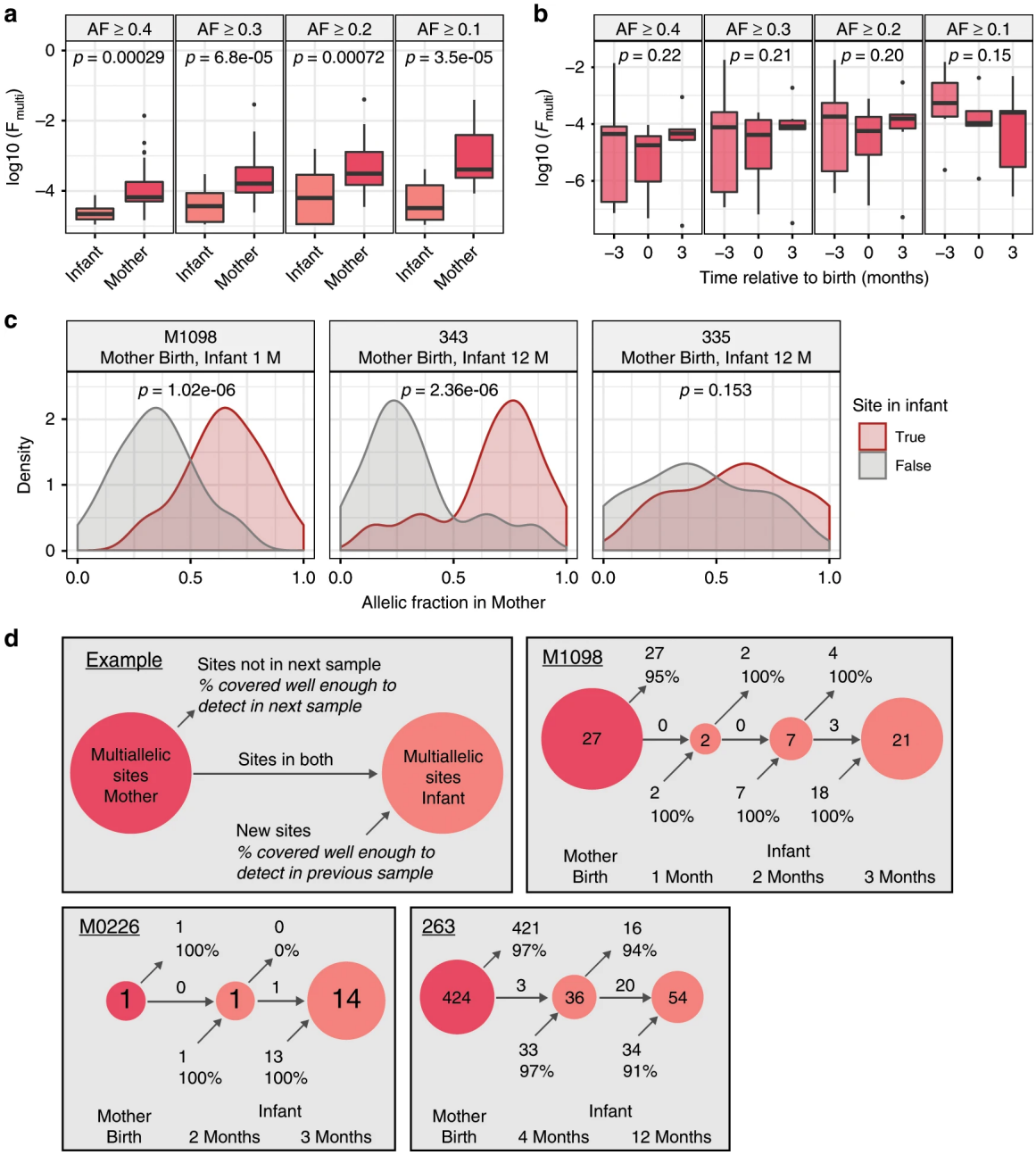


Figure 3.2: **P-crAssphage populations in mothers and infants differ in strain diversity.** **a** The p-crAssphage population in mothers has more multiallelic sites than the p-crAssphage population in infants. Fmulti (fraction of the p-crAssphage genome with multiallelic sites detected at the given allelic fraction threshold) in all mother and infant samples with at least one multiallelic site detected. P-values were calculated with the two-sided Wilcoxon rank-sum test and are uncorrected for multiple hypothesis testing. **b** P-crAssphage populations in mothers do not change in the number of multiallelic sites over time. Fmulti for mother samples from Yassour et al.[195]. P-values were calculated with a linear mixed model to account for repeated sampling of the same individual. **c** Allelic fraction of multiallelic sites in the p-crAssphage genome from mothers that are fixed in her infant. The distribution is separated by alleles that are present in the infant's p-crAssphage or not. P-values were calculated with the two-sided Wilcoxon rank-sum test. **d** Schematic depicting multiallelic sites in mother and infant samples over time. In the three cases where p-crAssphage was detected in the mother and multiple samples from the same infant, infants develop more multiallelic sites over time.

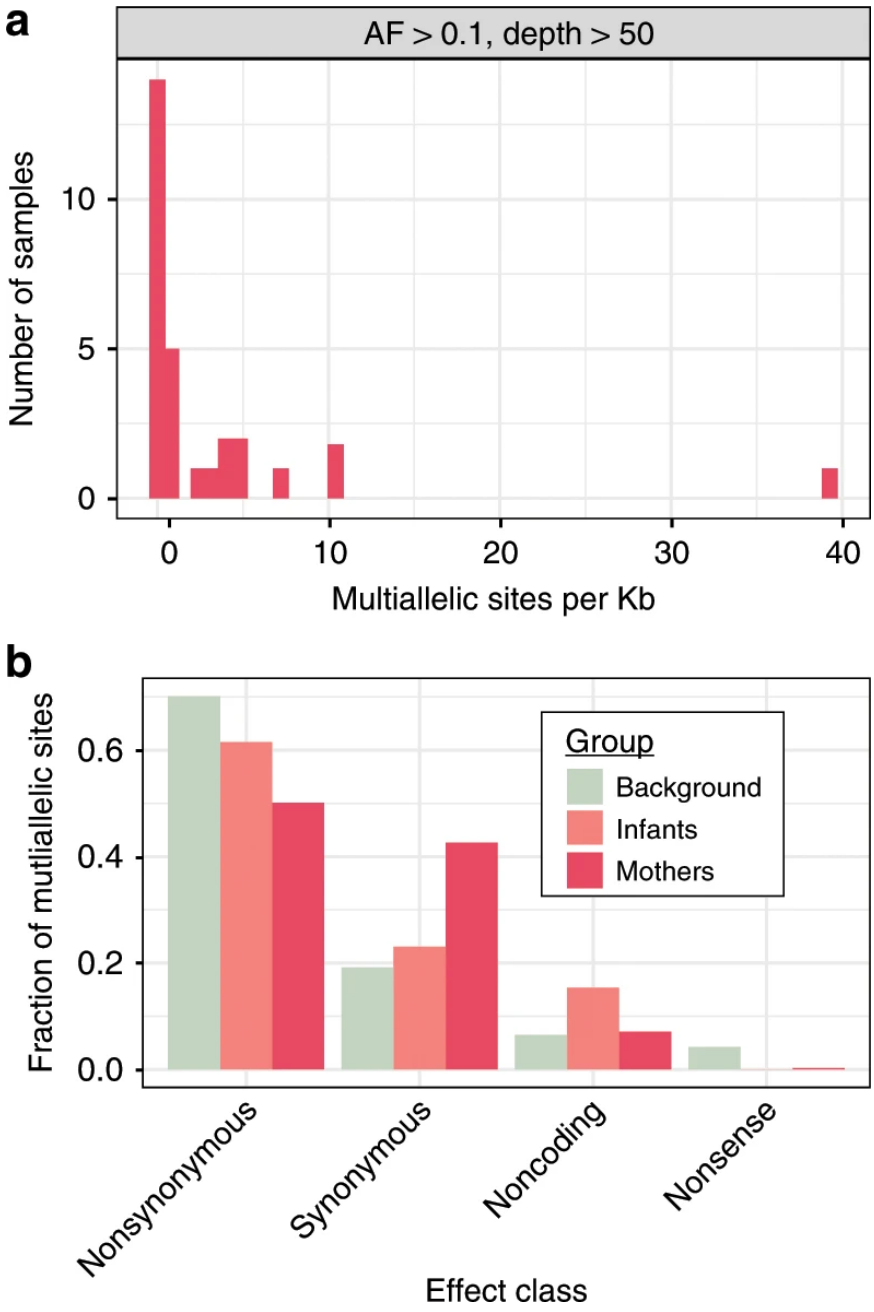


Figure 3.3: **Predicted effects of multiallelic sites differ in the p-crAssphage population of mothers and infants. a** Distribution of multiallelic sites per kilobase in samples from mothers. **b** Distribution of predicted effects of multiallelic sites from mother and infant samples, compared to a background distribution of equal probability of each DNA change at each position in the p-crAssphage reference genome.

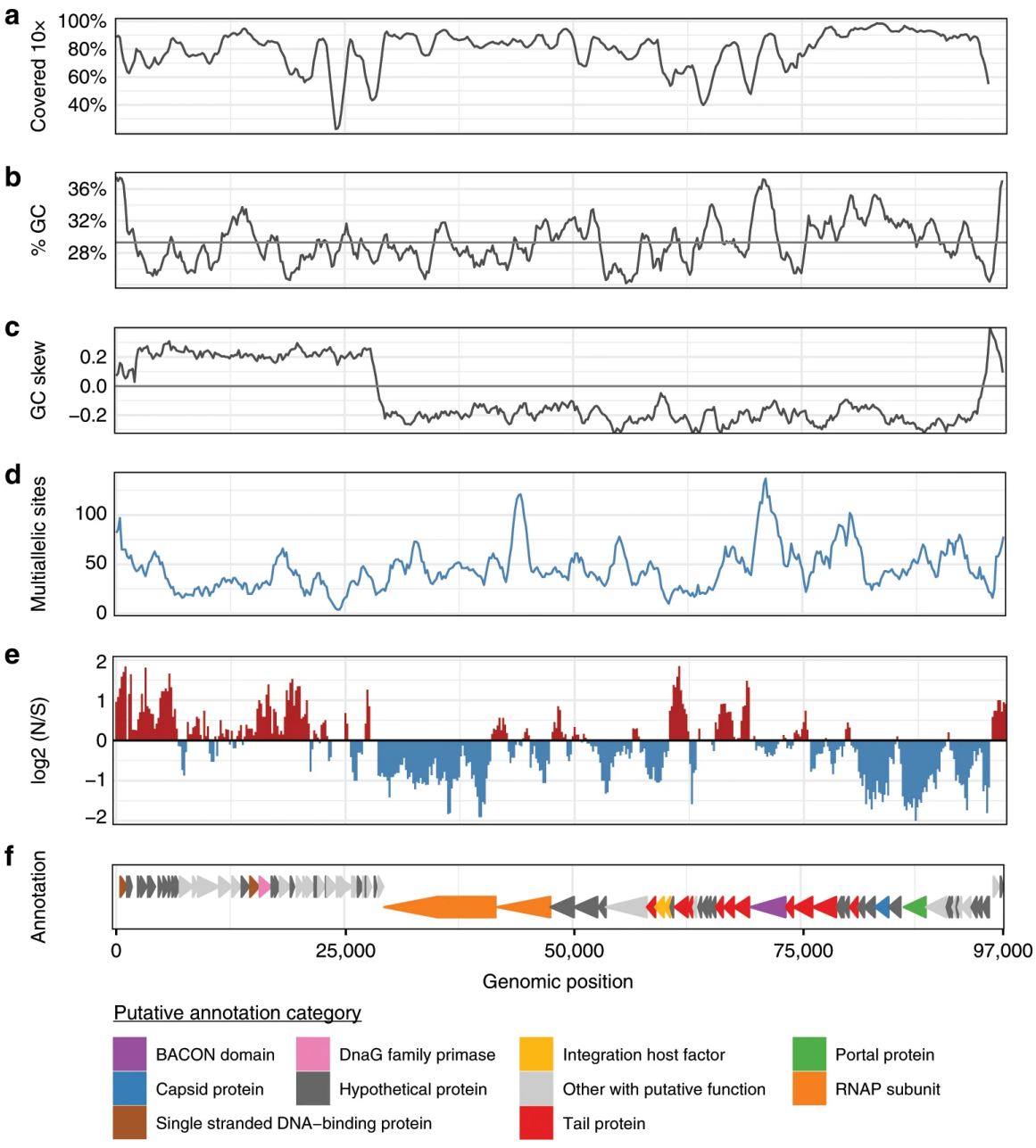


Figure 3.4: **The frequency and predicted effects of multiallelic sites vary across the p-crAssphage genome.** **a** Fraction of samples from mothers covered at least 10x. All values are calculated with a sliding window of size 1500bp with step size 200. **b** %GC content of the p-crAssphage reference genome. **c** GC skew of the p-crAssphage reference genome. **d** Total count of multiallelic sites (AF > 0.1) in the window. **e** Log base 2 ratio of nonsynonymous (N) to synonymous (S) multiallelic sites (AF > 0.1). f Annotation and selected predicted functions of genes in the reference genome.

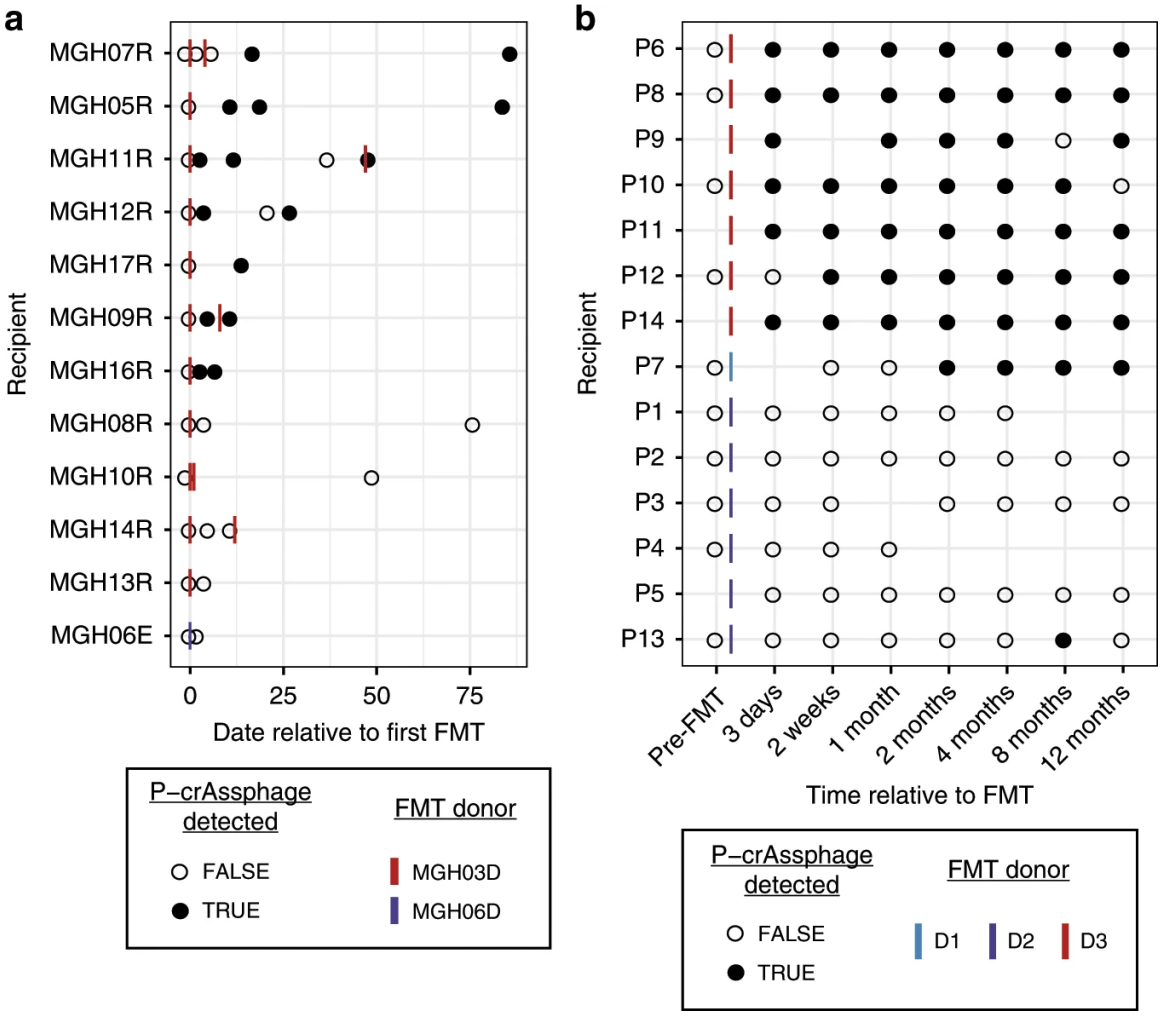


Figure 3.5: **P-crAssphage status in patients receiving FMT over time.** **a** P-crAssphage detection at 1x coverage in samples from Smillie et al.[162] Both donors shown were p-crAssphage positive. Open circles represent a p-crAssphage negative samples and closed circles represent p-crAssphage positive samples. **b** P-crAssphage detection at 10x coverage in samples from Draper et al.[41] Donor D1 was p-crAssphage positive, while donors D2 and D3 were p-crAssphage negative.

3.7 Supplementary Figures

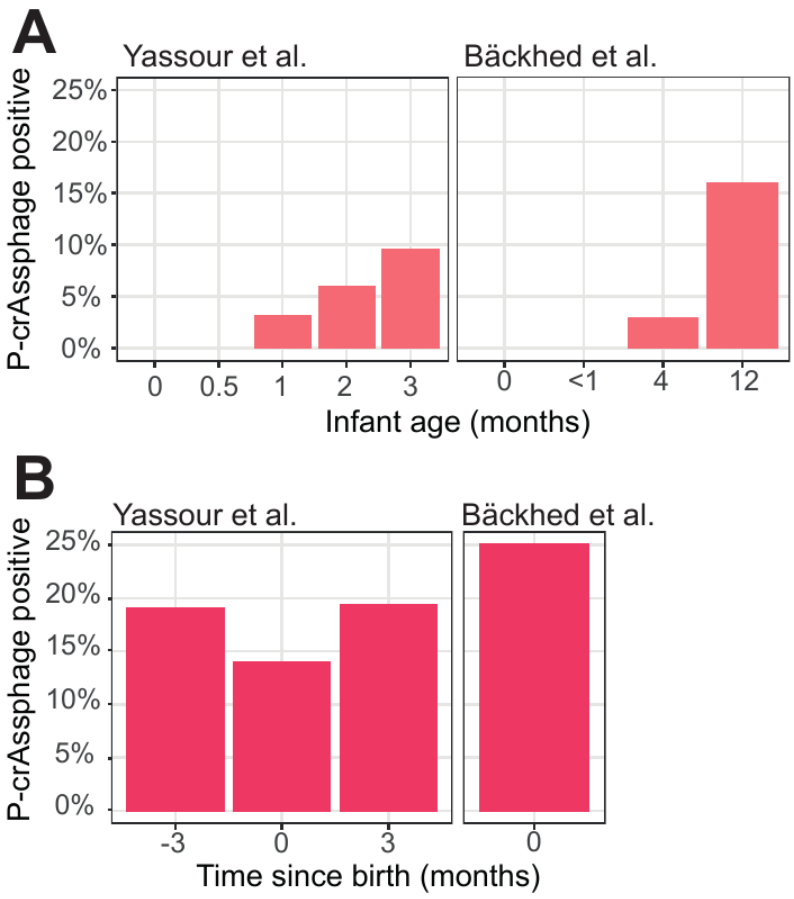


Figure 3.6: **P-crAssphage presence at 1x coverage in infant and mother samples.** P-crAssphage presence at 1x coverage in infant (a) and mother (b) samples.

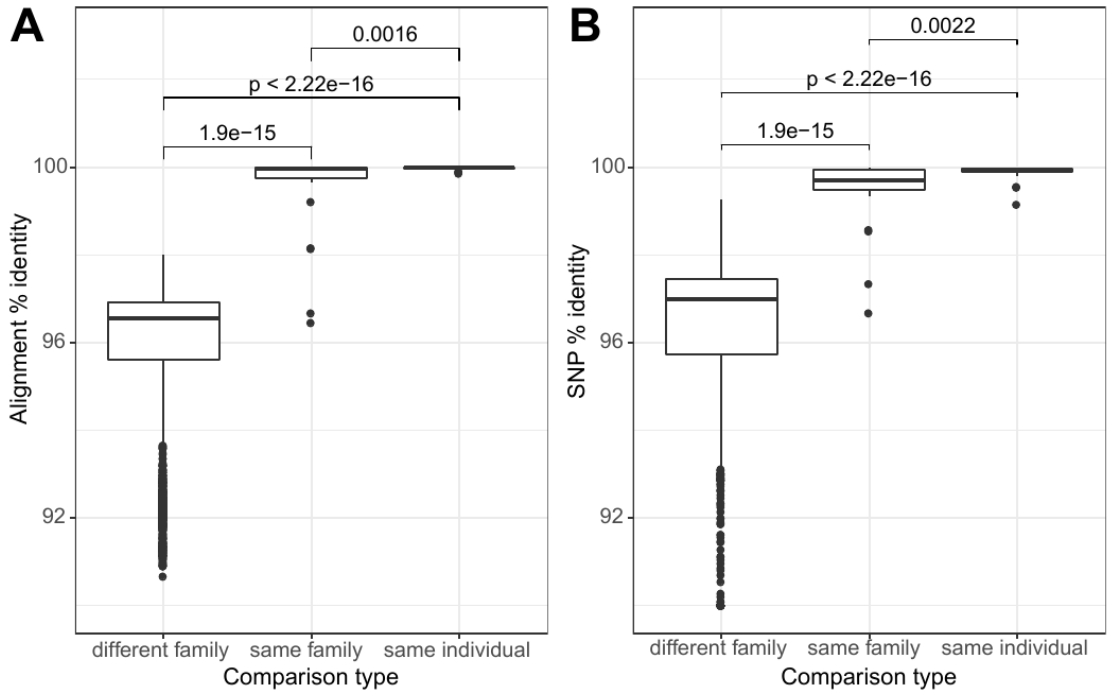


Figure 3.7: **P-crAssphage is more closely related in samples from mother-infant pairs than in samples from unrelated individuals a.** Distribution of pairwise alignment % identity of metagenome-assembled p-crAssphage genomes. Groups are separated by family relationships. P-values were calculated with the two-sided Wilcoxon rank sum test. **b.** Distribution of pairwise SNP % identity of p-crAssphage genomes. Groups are separated by family relationships. P-values were calculated with the two-sided Wilcoxon rank sum test. Boxes extend to the first and third quartile, whiskers extend to the upper and lower value within 1.5*IQR from the box. Outliers are shown as points.

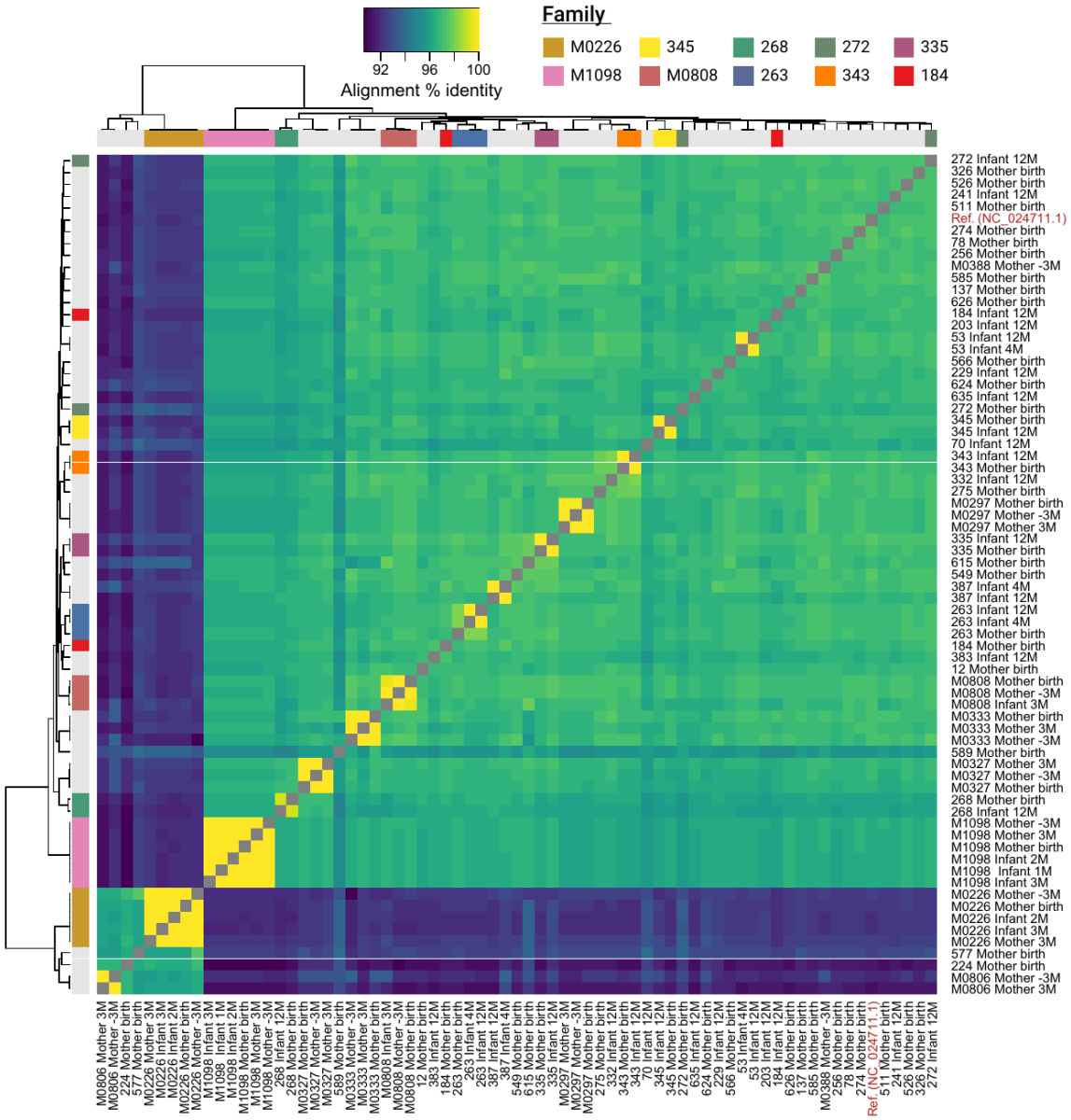


Figure 3.8: **Metagenome-assembled p-crAssphage genomes are highly similar in samples from matched mother-infant pairs.** The heatmap shows pairwise alignment % identity in all mother-infant samples with p-crAssphage detected. The p-crAssphage reference genome is also included as a comparison.

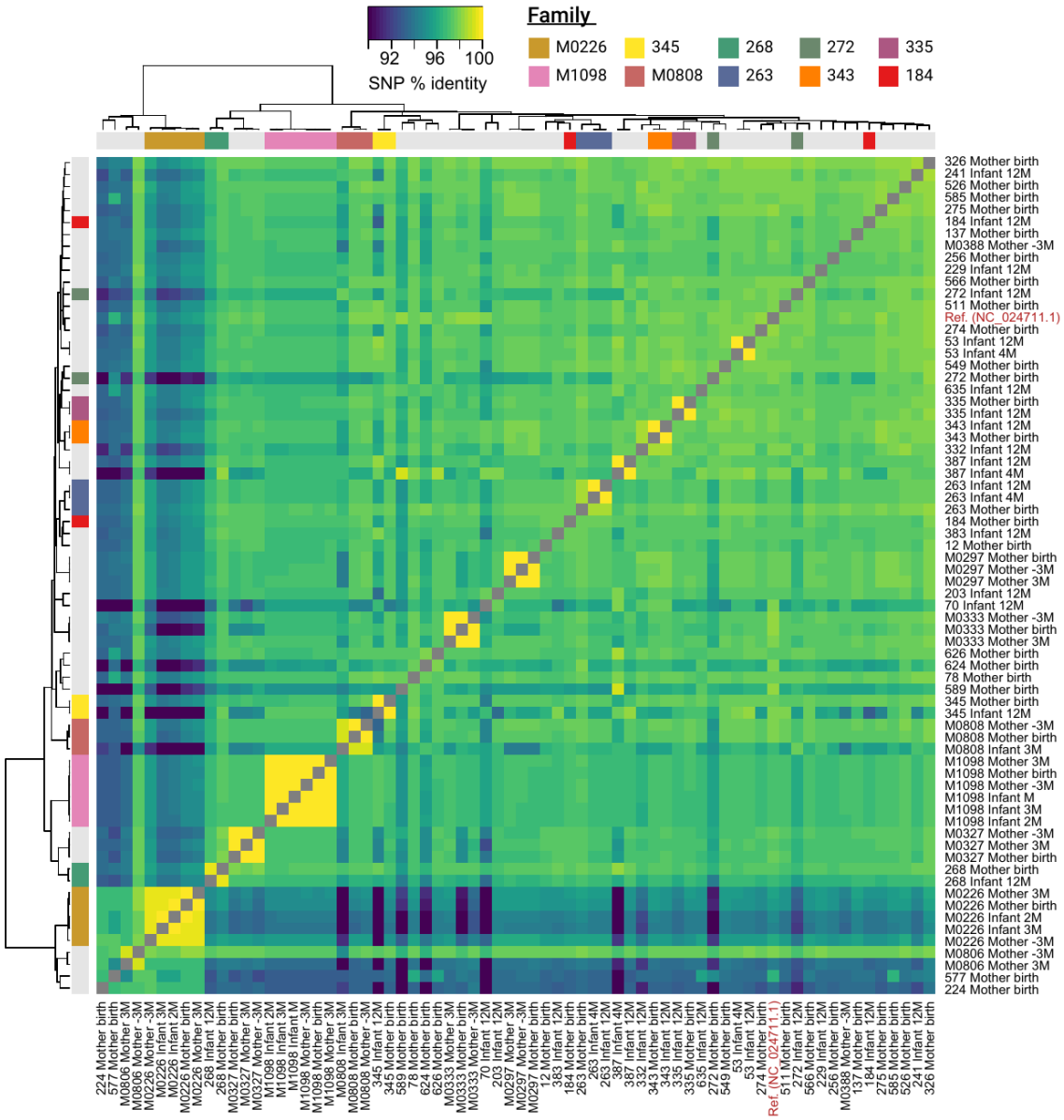


Figure 3.9: **P-crAssphage is highly similar at the SNP level in samples from matched mother-infant pairs.** The heatmap shows pairwise SNP % identity in all samples with p-crAssphage detected.

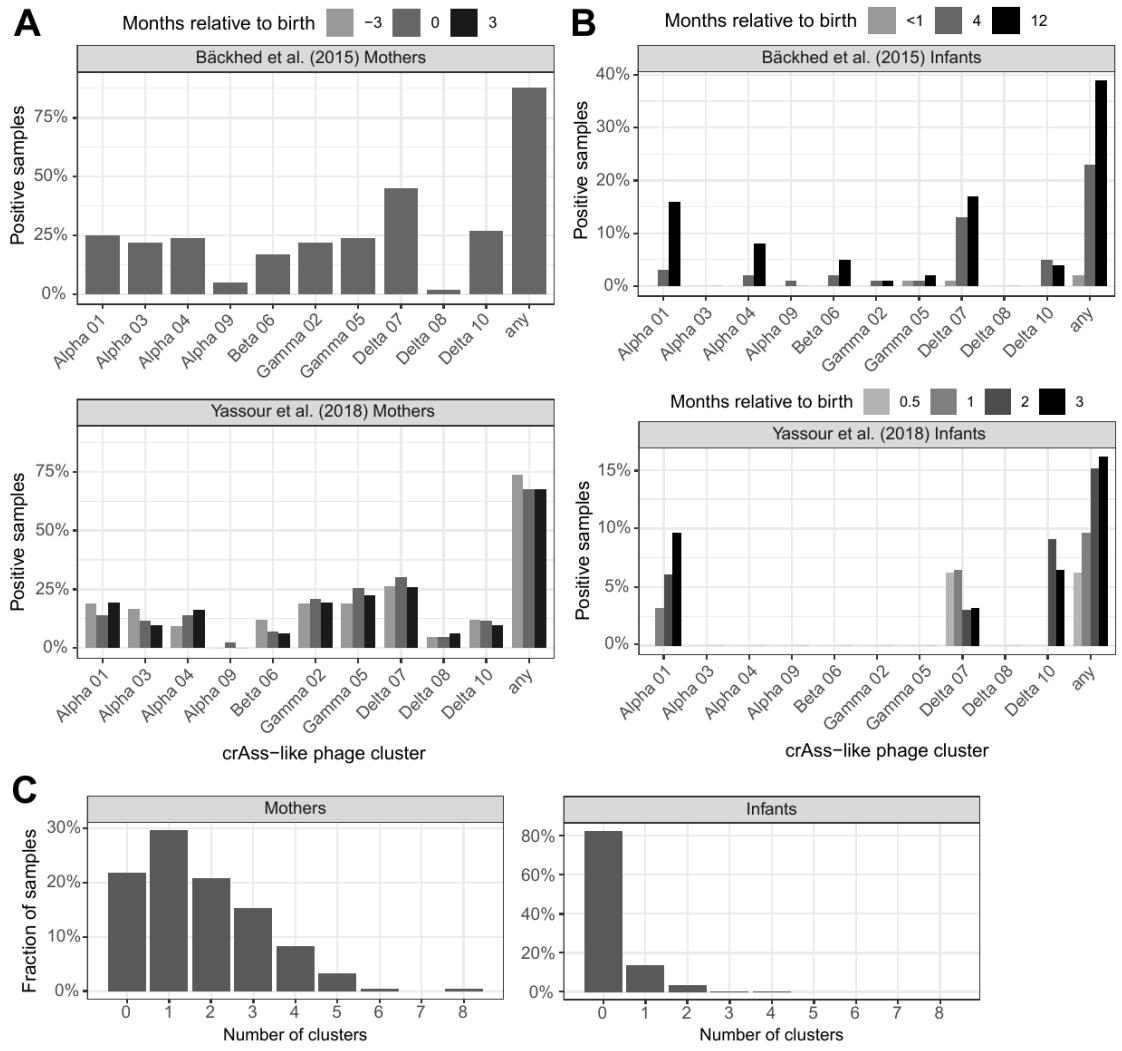


Figure 3.10: **CrAss-like phages detected at 1x covereage in mother and infant samples. a.** CrAss-like phages detected in samples from mothers in each study. **b.** CrAss-like phages detected in samples from infants in each study. c. Number of crAss-like phage clusters detected in each sample from mothers and infants.

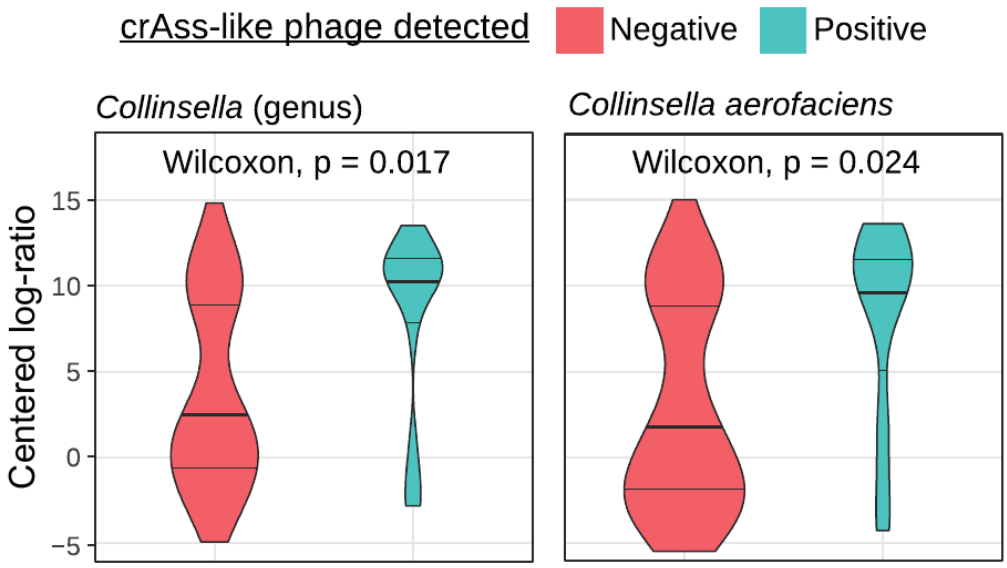


Figure 3.11: **Collinsella and Collinsella aerofaciens are at higher relative abundances in crAss-like phage positive vaginally delivered infants at 3-4 months of age, compared to crAss-like phage negative infants.** P-values calculated with the two-sided Wilcoxon rank sum test and corrected for multiple hypothesis testing. Boxes extend to the first and third quartile, whiskers extend to the upper and lower value within 1.5*IQR from the box. Outliers are shown as points.

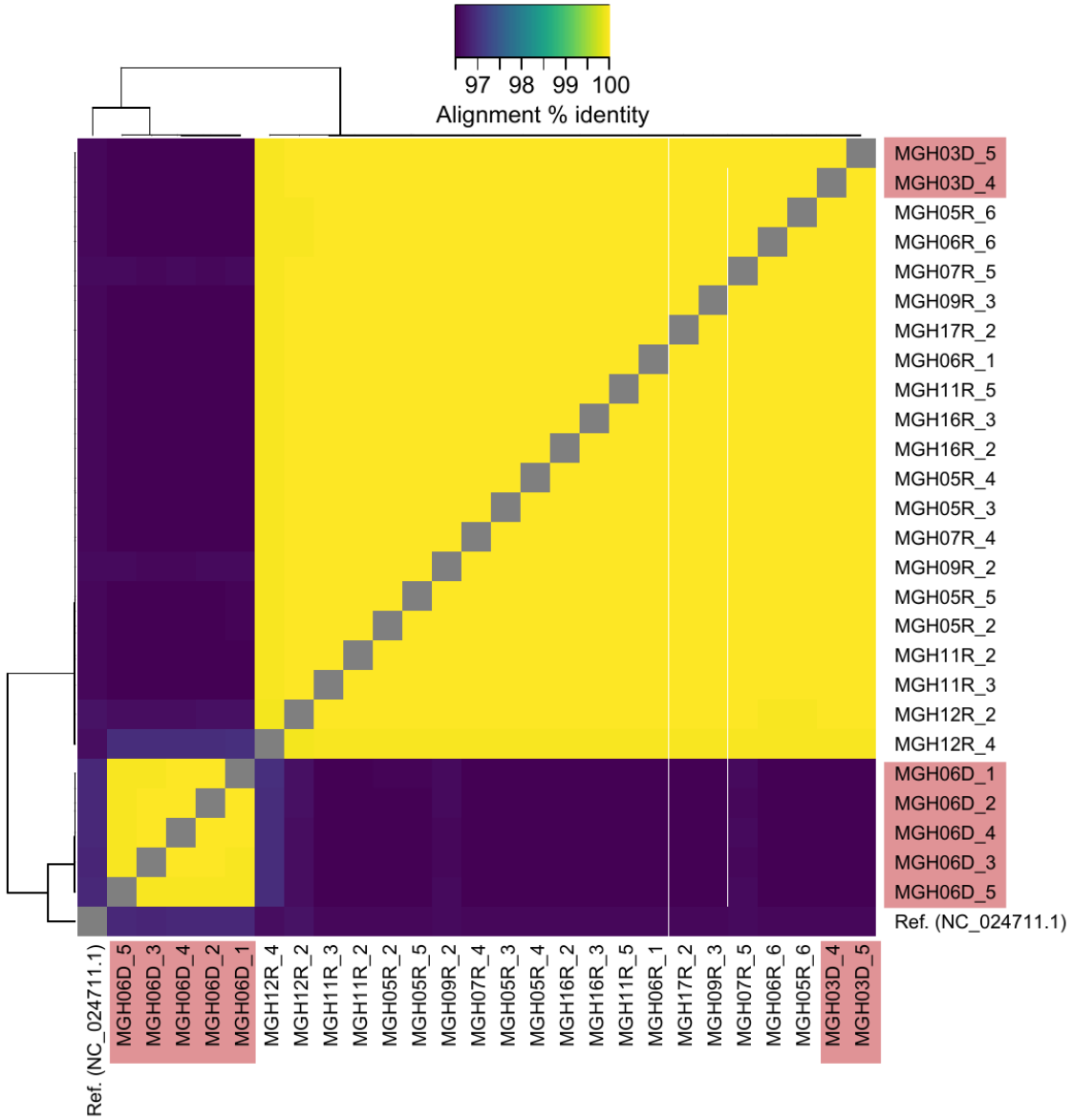


Figure 3.12: **Metagenome-assembled p-crAssphage genomes are highly similar in samples from matched FMT donor-recipient pairs in Smillie et al.** The heatmap shows pairwise alignment % identity in all samples that assembled >50kb p-crAssphage sequence. Assembled genomes from donor samples are highlighted in red. The p-crAssphage reference genome is also included as a comparison.

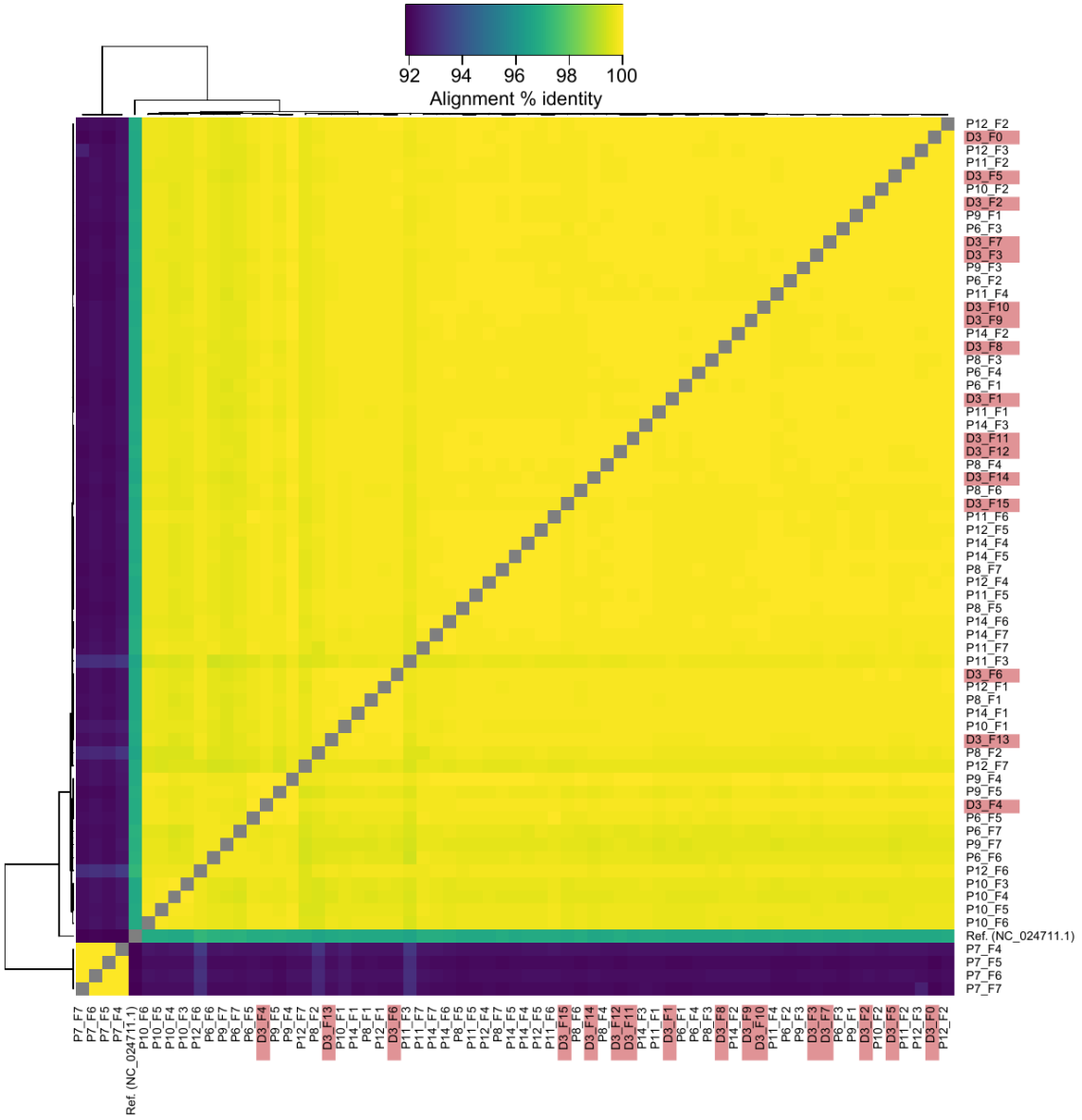


Figure 3.13: **Metagenome-assembled p-crAssphage genomes are highly similar in samples from matched FMT donor-recipient pairs in Draper et al.** The heatmap shows pairwise alignment % identity in all samples that assembled >50kb p-crAssphage sequence. Assembled genomes from donor samples are highlighted in red. The p-crAssphage reference genome is also included as a comparison. The donor for patient P7 was p-crAssphage negative; this patient may have acquired their p-crAssphage from the environment or another source.

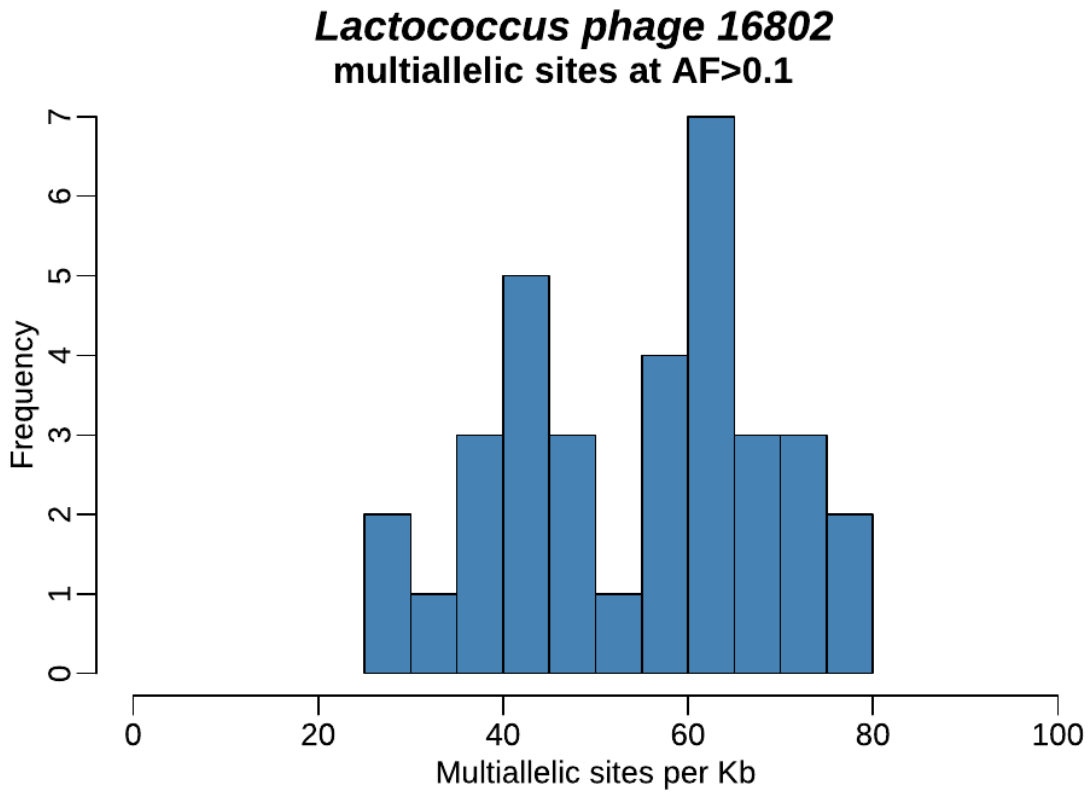


Figure 3.14: **Lactococcus phages detected in mother and infant samples.** Lactococcus phages were the only other group of phages detected with at least 1x coverage in at least ten mother and infant samples. This best represented member of this group, Lactococcus phage 16802, was detected in 34 samples and has more multiallelic sites than p-crAssphage on average, with a median of 58.8 multiallelic sites per kb.

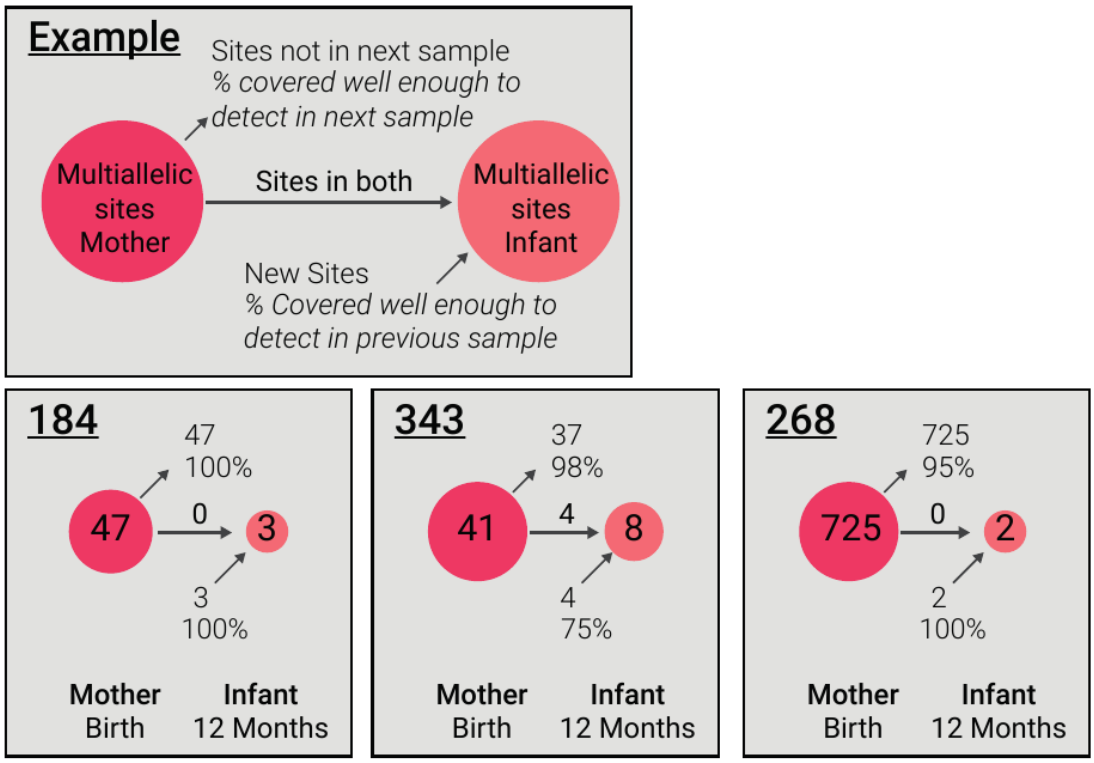


Figure 3.15: **Additional cases of multiallelic sites in mothers and infants with one p-crAssphage positive sample.**

Chapter 4

Rare transmission of commensal and pathogenic bacteria in the gut microbiome of hospitalized adults

The work in this chapter was presented in:

Siranosian, B.A., Brooks, E.F., Andermann, T., Rezvani, A.R., Banaei, N., Tang, H., and Bhatt, A.S. (2021). Rare transmission of commensal and pathogenic bacteria in the gut microbiome of hospitalized adults.

4.1 Abstract

Bacterial bloodstream infections are a major cause of morbidity and mortality among patients undergoing hematopoietic cell transplantation (HCT). Although previous research has demonstrated that pathogenic organisms may translocate from the gut microbiome into the bloodstream to cause infections, the mechanisms by which HCT patients acquire pathogens in their microbiome have not yet been described. We hypothesized that patient-patient transmission may be responsible for pathogens colonizing the microbiome of HCT patients, and that patients who share time and space in the hospital are more likely to share bacterial strains.

Here, we used linked-read and short-read metagenomic sequencing to analyze 401 stool samples collected from 149 adults undergoing HCT and hospitalized in the same unit over five years. We used metagenomic assembly and strain-specific comparison methods to investigate transmission of gut microbiota between individuals. While patients who shared time and space in the hospital did not converge in overall microbiome composition, we did observe four pairs of patients who harbor identical or nearly identical *E. faecium* strains in their microbiome. These strains may be the result

of transmission between patients who shared a room and bathroom, acquisition from a common source in the hospital or transmission from an unsampled source.

We also observed identical *Akkermansia muciniphila* and *Hungatella hathewayi* strains in two pairs of patients. In both cases, the patients were roommates for at least one day, the strain was absent in the putative recipient's microbiome prior to the period of roommate overlap and the putative recipient had a microbiome perturbed by antibiotic treatment for a bloodstream infection. Finally, we identified multiple patients who harbored identical strains of several species commonly found in commercial probiotics and dairy products, including *Lactobacillus rhamnosus*, *Lactobacillus gasseri* and *Streptococcus thermophilus*. Overall, our findings indicate that pathogenic organisms from a single source are not frequently colonizing the gut microbiome of multiple patients. However, the potential transmission of commensal microbes with immunomodulatory properties raises questions about the recovery of microbiome diversity after HCT, and indicates that patients in this setting may acquire new microbes by sharing space with others.

4.2 Introduction

Patients undergoing hematopoietic cell transplantation (HCT), a potentially curative treatment for a range of hematologic malignancies and disorders, are at increased risk for bloodstream infections (BSIs) and associated morbidity and mortality[198]. While the bacterial pathogens that cause BSIs in HCT patients are well understood, their routes of transmission are often unclear. Determining these transmission pathways involves identifying two critical elements: the source of the infection, i.e., how the pathogen was introduced into the patient's bloodstream, and the origins of the particular pathogen causing the BSI.

The most common ways bacterial pathogens can be introduced into an HCT patient's bloodstream include contaminated central intravenous lines and translocation of intestinal microbiota across a damaged epithelium[146]. Indeed, research from our group and others has shown that strains of bacteria isolated from the blood of HCT patients with BSIs may be indistinguishable from the strains in the intestinal microbiota of these patients prior to infection[76, 168, 200]. In addition, HCT patients with a microbiome dominated by a single bacterial taxon, such as Enterococcus or Streptococcus, are at increased risk for not only BSI[170, 181], but also graft-versus-host disease[69, 105] and death[132, 155, 171, 185].

Identifying the source of the BSI is only the first step. To fully understand the transmission pathways of bacterial pathogens in hospital settings, it is also essential to determine the origin of the pathogen that caused the BSI. For gut-based pathogens, there are three possibilities. First, they may exist in the HCT patient's microbiome upon admission to the hospital. Second, hospital environments and equipment may serve as unintentional reservoirs of pathogens, thereby infecting multiple patients through exposure[84]. Lastly, a pathogen could originate from the microbiome

of other patients, healthcare workers or hospital visitors and be transmitted via shared spaces. In cases where traditional epidemiological links cannot be found, this patient-patient transmission of gut microbes may be the "missing link" that explains the persistence of BSIs in hospital environments[137].

Transmission of gut bacteria and phages between individuals is known to occur in specific cases, such as from mothers to developing infants[13, 159, 195]. By contrast, adults have a microbiome that is relatively resistant to colonization with new organisms even after perturbation by antibiotics[99, 167, 204]. While adults living in the same household or in close-knit communities may have more similar microbes than those outside the group[21], to our knowledge, direct transmission of gut microbes between adults has not been observed with high-resolution metagenomic methods, with the notable exception of fecal microbiota transplantation[162, 184, 90], a drastic reshaping of the gut microbiota often used in response to *Clostridiodes difficile* infection. Transmission of gut microbiota is thought to occur by a fecal-oral route, which could happen in the hospital environment by exposure to contaminated surfaces or equipment, sharing a room or bathroom, contaminated hands of healthcare workers or other sources. The perturbed microbiomes of HCT patients, often lacking key species to provide colonization resistance, may be primed to acquire new species from these sources.

Previous studies of the microbiome in HCT patients have often used 16S rRNA sequencing[132, 7, 136, 156], which is sufficient for taxonomic classification but cannot differentiate specific strains in a mixed community. By contrast, short-read shotgun metagenomic sequencing can capture information from all bacterial, archaeal, eukaryotic and phage DNA in a stool sample. While short-read sequencing data is accurate on a per-base level, it is often insufficient to assemble complete bacterial genomes due to the presence of repetitive genetic elements. Linked-read sequencing captures additional long-range information by introducing molecular barcodes in the library preparation step. This technology allows for significant increases in assembly contiguity[19, 74] while retaining high per-base accuracy. Both of these technologies also capture information about strain diversity, genetic variation within the population of a species[178, 182], which is critical for measuring transmission between microbiomes.

Here, we use a collection of short-read and linked-read metagenomic sequencing datasets from 401 stool samples to analyze bacterial transmission between HCT patient microbiomes at a single, high-volume hospital. We apply strain-resolved comparison methods to show that transmission of bacteria between adults hospitalized in the same unit at the same time is likely a rare event, usually occurring when recipients have extremely perturbed microbiomes, such as after exposure to broad-spectrum antibiotics. Bacterial strains shared between individuals include both pathogenic and commensal organisms, demonstrating that transmission may depend more on niche availability than pathogenicity or antibiotic resistance capacity. We find that pathogens colonizing HCT patient microbiomes are present in the first sample in a time course roughly 60-70% of the time in our cohort. This suggests that in most cases, prior colonization, rather than direct transmission from other

patients or the hospital environment, is responsible for pathogenic organisms in the gut microbiomes of this patient population. Even though patients were frequently placed into double occupancy hospital rooms with a shared bathroom, we observe relatively few putative transmission events. This implies that sharing a room with another patient may not place a patient recovering from HCT at a greatly increased risk of acquiring pathogens in their gut microbiome.

4.3 Results

4.3.1 Sample characteristics and patient geography

We collected weekly stool samples (see methods) from adult patients undergoing hematopoietic cell transplantation (HCT) at Stanford University Medical Center from 2015-2019. At the time of the study, our biobank contained over 2000 stool samples from over 900 patients. Samples from October 2015 to November 2018 were considered for this study. Relevant patient health, medication, demographic, hospital admission and room occupancy data were extracted from electronic health records (Table 1, Table S1). All patients stayed in a single ward of the hospital during treatment, which contained 14 single-occupancy and four double-occupancy rooms, the latter of which included shared bathrooms (Figure S1a). Patients spent a median of 18 days on the ward and were frequently moved between rooms: 42% of patients spent at least one day in three or more rooms during the course of treatment (Table 2, Figure S1c). 73% of patients shared a room with a roommate for $\geq$24 hours. Over the course of their hospital stays, many patients had several roommates, though never more than one at a time (Figure S1d). Patients with multi-resistant Gram-negative infections were always placed into single rooms with contact precautions.

To understand how geographic overlap may influence transmission of gut microbes, we created a network from patient-roommate interactions (Figure S1b). 535 patients (77% of patients with at least one roommate, 56% of all patients) fell into the largest connected component of the network. Although the largest component was not densely connected (mean degree 2.2 $\pm$ 1.6 standard deviation (SD)), it links together patients over three years and may represent a risk for infection transmission. We used the network to select samples for further analysis with metagenomic sequencing, as described in the methods.

4.3.2 Metagenomic sequencing, assembly and binning

For an overview of the steps used in the generation and processing of sequence data, see Figure 1a. 328 stool samples from 94 HCT patients were subject to short-read metagenomic sequencing as part of previous projects (for references and SRA IDs of these samples, see Table S2). 96 additional samples from 62 patients were selected for linked-read sequencing to span periods of roommate overlap between patients. Samples were subjected to bead beating-based DNA extraction and bead-based

DNA size selection for fragments ≥2 kb (see methods). We prepared linked-read sequencing libraries with the 10X Genomics Chromium platform from 89 samples with sufficient DNA concentration. Samples were sequenced to a median of 116 million (M) (± 37 M SD) read pairs on an Illumina HiSeq4000. In total, 401 stool samples from 149 patients were sequenced (Table 3), with a median of 2 and maximum of 13 samples per patient (Figure 1b).

We processed all existing short-read data and newly generated linked-read data by first trimming and then removing low quality reads, PCR duplicates (short-read data only) and reads that aligned against the human genome (see methods). After quality control, newly sequenced linked-read samples had a median 104 M (± 40 M SD) read pairs, while short-read data had a median 7.6 M (± 4.4 M SD) read pairs. Metagenomic assembly was conducted using metaSPAdes[118] for short-read data, and MEGAHIT[86] followed by Athena[19] for linked-read data. Short-read assemblies had a median N50 of 17.2 kb ± 24.8 kb, while linked-read assemblies had a median N50 of 147.6 ± 165.8 kb. We binned metagenome-assembled genomes (MAGs) using Metabat2[72], Maxbin[192] and CONCOCT[6] and aggregated across results from each tool using DASTool[157]. MAG completeness and contamination was evaluated using CheckM[129] and MAG quality was determined by previously established standards[20]. The vast majority of short-read and linked-read MAGs were at least medium quality, and 27% of linked-read MAGs contained the 5S, 16S and 23S rRNA genes and at least 18 tRNAs to be considered high-quality (Figure 1c, Table S3). Linked-read MAGs had higher quality than the 4,644 species-level genomes in the Unified Human Gastrointestinal Genome collection[5], where 573 genomes (12.3%) are high-quality, and only 38 (6.6%) of those came from metagenomes rather than isolates. Sequencing dataset type (short-read vs linked-read) did not have a linear relationship with MAG length (linear regression, $p > 0.9$); the increase in quality was mainly due to the inclusion of ribosomal and transfer RNA genes in the linked-read MAGs, which often do not assemble well with short-read sequencing data alone. To understand the diversity of strains present in the microbiomes of our patients, we clustered all medium- and high-quality MAGs at 95% and 99% identity thresholds (roughly "species" and "strain" level, see methods) using dRep[121], yielding 1615 unique genomes representative of the microbial diversity in this sample set.

4.3.3 Classification of abundant Healthcare-associated Infection organisms

We performed taxonomic classification of sequencing reads with Kraken2[189] and abundance estimation with Bracken[100] using a custom database of bacterial, fungal, archaeal and viral genomes in NCBI Genbank (see methods) (Table S4, S5). A median of 33% ± 15% SD reads were classified to the species level with Kraken2 (72% ± 15% SD at the genus level), which was improved to 96% ± 7% SD using Bracken (97% ±8% SD at the genus level). Organisms that cause healthcare-associated Infections (HAI) were identified from the CDC list of pathogens[38]. Here, we report organisms as present if they achieve 1% relative abundance, but acknowledge that many microbes typically exist

at lower concentrations, which may be undetectable with metagenomic sequencing.

Many HAI organisms were prevalent in the microbiomes of the studied HCT patients. 152 samples (38%) from 79 patients (53%) had at least one HAI organism identified at 1% relative abundance or above (Figure 1e). *Escherichia coli* was the most common HAI organism (present at ≥1% in 42/149 patients, 28.2%; 0.1%=80/149, 53.7%), followed by *Klebsiella pneumoniae* (39/149 patients, 26.2%; 0.1%=70/149, 47%) and *Enterococcus faecium* (26/149 patients, 17.4%; 0.1%=73/149, 49%). Rates of colonization with HAI organisms were much higher than in stool samples from healthy individuals in the Human Microbiome Project[97] where *E. coli* reaches 1% relative abundance in 2.1% of samples, and *K. pneumoniae* and *E. faecium* are never found at greater than 1% (present at 0.1% in 18.4%, 1.4% and 0.7% of samples, respectively).

HCT patient microbiomes can become dominated by HAI organisms, often as a result of antibiotic usage. 24 patients (16%) have at least one sample with a dominant HAI organism (≥ 30% relative abundance), which may place them at increased risk for bloodstream infections (BSI)6. BSI in this cohort of HCT patients is most frequently caused by *E. coli*, viridans group Streptococci and *E. faecium*; these organisms less frequently cause BSI among the entire inpatient population at our hospital (Figure 1f). We focused further analysis on *E. coli* and *E. faecium*, as these species are both frequently detected in stool and frequently cause BSIs. While viridans group Streptococci frequently cause BSIs in HCT patients, these species are typically much more prevalent in the oral cavity[146, 1] compared to the gut microbiome (individual species in the group only reach 1% relative abundance in 8/149 patients, 5%).

4.3.4 Detection of *E. coli* and *E. faecium* becomes more common during a patient's hospital stay

We investigated the detection of *E. coli* and *E. faecium* in patients with time course samples (82/149 patients, 55%) as a proxy for understanding if these organisms were acquired or became more abundant during the observed hospital stay. Of the 1615 de-replicated MAGs, nine were identified as *E. coli* and five were identified as *E. faecium*. We mapped sequencing reads from all samples to these MAGs and evaluated the maximum coverage breadth, the fraction of the reference genome covered with at least one sequencing read. We used a breadth cutoff of 50% to determine "detection" or "absence" in a sample. This threshold is likely specific (it is difficult to achieve 50% breadth by read mis-mapping or homology with a different organism) but not extremely sensitive (it will likely miss very lowly abundant organisms that are truly present).

64/82 (78%) patients with time course samples have *E. coli* present at 50% breadth in at least one sample. Of these, 10/64 (16%) have *E. coli* below 50% breadth in the first sample. 60/82 (73%) patients with time course samples have *E. faecium* present at 50% breadth in at least one sample. Of these, 17/60 (28%) have *E. faecium* below 50% breadth in the first sample (Figure 1d). In these patients, *E. coli* or *E. faecium* may have been newly acquired into the gut microbiome during

the hospital stay. Alternatively, the organism could have also been present at low abundance and below our limit of detection in the first sample. It is possible that antibiotic use in the weeks after HCT kills off many of natural microbiome colonizers, allowing antibiotic resistant HAI organisms to increase in relative abundance past our limit of detection. Overall, the relative abundance of *E. coli* and *E. faecium* was not significantly different in first samples compared to later samples (Wilcoxon rank-sum test). While more patients have newly detectable *E. faecium* than *E. coli*, the difference was not statistically significant (p=0.09, binomial likelihood ratio test).

4.3.5 Antibiotic use and its effect on HCT patient microbiomes

HCT patients are frequently prescribed antibiotic, antiviral and antifungal drugs, especially in the days immediately after transplant. These drugs can have a significant impact on the microbial populations in the gut and contribute to the loss of microbial diversity frequently observed following HCT6. Antibiotic use likely impacts the dynamics of bacterial transmission in this patient population, as both natural colonizers, which may provide resistance to newly invading species, and potentially transmitted species can be killed by the drugs. We gathered electronic health record data to understand the characteristics of antibiotic prescription in our patient cohort and its potential impact on the gut microbiome composition.

Patients were prescribed a median of five different antibiotics (range 1 - 10) and had a median of 90 cumulative antibiotic-days (range 14 - 416). The most common antibiotics prescribed were ciprofloxacin (98% of patients prescribed for at least one day), IV vancomycin (80%), cefepime (66%) and piperacillin-tazobactam (57%). Prescription of most antibiotics peaked in the 14 days following HCT, while administration of antifungal drugs like posaconazole and antiviral drugs like ganciclovir were higher up to 50 days following HCT (Figure 2a). We found that antibiotic usage in the prior seven days before a stool sample was collected was strongly negatively associated with bacterial diversity (linear regression on cumulative antibiotic-days and Shannon diversity, R2 =0.18, 0.14, p=8.9e-11, 2.6e-8 for species and genus level, respectively). Samples with a single species dominant at 30% relative abundance or higher also typically had higher antibiotic usage in the past seven days (p=0.001, Wilcoxon rank-sum Test) (Figure 2b-d).

4.3.6 Patients who share time and space in the hospital do not converge in microbiome composition or frequently share strains

To understand the overall impact of temporal and spatial overlap on patient microbiomes, we first studied the taxonomic (Bray-Curtis) similarity between samples from different patients. For each sample pair, we calculated the maximum time the two patients had overlapped in the hospital or as roommates, determined by the earlier sample time. There was not a significant linear association between time of overlap and taxonomic similarity (linear regression, p=0.42 for roommate, p=0.068 for

hospital, Figure 2e,f), indicating that the broad taxonomic composition of patient microbiomes may not converge under temporal and geographic overlap. However, there may be isolated strains that are shared between patients. We used the strain diversity-aware, SNP-based method inStrain[122] to conduct a sensitive analysis of bacterial strains shared between patients. InStrain compares alignments of short reads from multiple samples to the same reference genome and reports two metrics: Consensus ANI (conANI) and PopulationANI (popANI). ConANI counts a SNP when two samples differ in the consensus allele at a position in the reference genome, similar to many conventional SNP calling methods. PopANI counts a SNP only if both samples share no alleles. For example, if A/T alleles were found at frequencies of 90/10% and 10/90% in two samples, a consensus SNP would be called because the consensus base is different. A population SNP would not be called because both samples share an A and T allele.

We mapped sequencing reads from all samples against the collection of 1615 unique MAGs and compared strains in samples from different patients with >50% coverage breadth at a depth of five reads (see methods). This coverage breadth threshold ensures ANI is calculated across the majority of the strains being compared and is recommended by the authors of inStrain. The maximum log-scaled popANI value was taken as representative of the maximal strain sharing between all pairs of patients. Length of hospital overlap was significantly associated with having a more similar microbiome strain (linear regression on log-scaled popANI values, p=0.0017), while time of roommate overlap was not significantly related (Figure 2i,j). However, we note that this comparison excludes pairs of patients with strains below 97% ANI or with very lowly abundant strains, as they would not be detected by inStrain. Taken together, these results suggest that the microbiomes of patients who share time and space in the hospital are not converging, either at the broad taxonomic level, or frequently at the specific strain level. We followed up on the few cases of high-identity strains as evidence for possible transmission events.

4.3.7 HAI organisms that colonize HCT patient microbiomes are part of known, antibiotic resistant and globally disseminated clades

E. coli and *E. faecium* are common commensal colonizers of human microbiomes[42, 128, 175]. These species can also be pathogenic and contribute to inflammation, dysbiosis and infection in the host[78, 149]. The specific strain of these species is key in determining the balance between a healthy and diseased state in the microbiome. We compared patient-derived *E. coli* and *E. faecium* MAGs with several reference genomes (Table S6) to identify the closest strains or sequence types.

Escherichia coli

The 95% identity, or species-level, cluster of *E. coli* MAGs contained 47 genomes from 26 patients (Figure 3). Within this alignment average nucleotide identity (ANI) based tree, we observed two

clades of genomes where MAGs from multiple patients had >99.9% ANI. These clades were investigated further as they may represent common sequence types. The first clade contained 15 MAGs from 7 patients; these MAGs had >99.9% ANI to pathogenic *E. coli* sequence type (ST) 131 clade C2 genomes, including EC95854 and JJ188655. ST131 is an extraintestinal, pathogenic, multidrug resistant *E. coli* strain which frequently causes urinary tract infections[61]. *E. coli* ST131 often carries extended-spectrum -lactamase (ESBL) genes which convey a wide range of antibiotic resistance. This sequence type is believed to colonize the intestinal tract even in healthy individuals without antibiotic exposure[186], and there are reports of this pathogen causing urinary tract infections in multiple individuals within a household[101].

The second clade contained 12 MAGs from 5 patients with >99.8% ANI to the pathogenic ST648 representative IMT16316[143]. ST648 is also an ESBL-producing *E. coli* strain, but it is not as widespread as ST131. Both STs have been isolated from wastewater[131] and ST648 has been isolated from the gut of humans[112] and other mammals[48]. Our finding that *E. coli* ST648 is also prevalent in HCT patient microbiomes suggests that it may become a pathogen of interest in this patient population in the future.

To understand the antibiotic resistance capabilities of the *E. coli* strains colonizing these HCT patients, we searched for β-lactamase genes in the *E. coli* MAGs (see methods). The most commonly detected genes were ampH and ampC, which are part of the core *E. coli* genome and likely do not contribute to antibiotic resistance[65] (Figure S2A). CMY-132 was detected exclusively in MAGs in the ST131 clade, and mutations conveying resistance in gyrA were detected in MAGs in multiple clades. CTX-M-type β-lactamases were detected in several samples, but often within the metagenome rather than within the *E. coli* MAG, indicating they may be on plasmids or mobile genetic elements that did not bin with the rest of the *E. coli* genomes.

Enterococcus faecium

The species-level cluster of *E. faecium* MAGs contained 30 genomes from 20 patients (Figure 4a). All MAGs were ≥99% identical, suggesting a single ST is present in most patients50. These genomes matched closest to *E. faecium* ST117, a well-described vancomycin-resistant strain that frequently causes bloodstream infections[174]. Notably, five other MAGs had 94% ANI (below the 95% clustering threshold, and therefore not shown in Figure 3a) to the ST117 clade and >99% ANI to commensal *E. faecium* strains, including strains Com15 and Com1250. To understand the vancomycin resistance capabilities of these strains, we searched for the seven genes in the vanA operon[3]. Only 2/30 samples had the full operon present within the *E. faecium* MAG (Figure S2b). When we looked in the entire metagenome, 22/30 samples had the full operon present, and the vanA genes were usually detected on a contig that was not assigned to any MAG. Van genes are often carried on mobile genetic elements or plasmids in *E. faecium*, which frequently do not assemble and bin well due to repetitive sequences.

Interestingly, no van genes are present in sample 11342_02, but the full operon is present in the *E. faecium* MAG in sample 11342_03, collected 14 days later from the same patient. In sample 11342_03, the vanA operon appears on a contig 8.7 kb long with 143x coverage, much lower than the 2000x coverage of the rest of the *E. faecium* MAG. Mapping reads against this contig revealed scattered coverage in 11342_02, leading us to believe the operon is present, but not at high enough coverage to be assembled in the earlier sample. We attempted to culture vancomycin-resistant Enterococcus (VRE) from these samples (see methods), and positively identified *E. faecium* by MALDI coupled to time-of-flight mass spectrometry in samples 11342_02 and 11342_03. Taken together, these results suggest that there were at least two strains of *E. faecium* in the gut microbiota of patient 11342. The low relative coverage of the vanA operon indicates that the VRE strain may have been a small fraction of the total *E. faecium* population. Patient 11342 was never prescribed vancomycin (Figure S7a), so the VRE strain may have not had an advantage in this environment.

VanA genes were also not detected in sample 11349_01, where we observed a nearly identical *E. faecium* strain to patient 11342 after the patients shared a room for 11 days. When we attempted to culture VRE from 11349_01, a bacterium grew poorly on plates containing vancomycin and was identified as *Klebsiella pneumoniae*. Therefore, we believe the *E. faecium* strain in the microbiome of this patient was vancomycin sensitive. If transmission from patient 11342 was responsible for colonization of patient 11349, the vancomycin-sensitive strain may have been transmitted.

4.3.8 Nearly identical strains indicative of putative patient-patient *Enterococcus faecium*, but not *Escherichia coli* transmission

We used the results from the inStrain comparison to search for nearly identical bacterial strains, which may be indicative of transmission between patients. To determine a threshold for putative transmission, we examined comparisons in several "positive control" datasets where we expect to find identical strains, either as the result of persistence or transmission: time course samples from the same HCT patient, stool samples from mother-infant pairs[195] and samples from fecal microbiota transplantation donors and recipients[162]. We often observed 100% popANI in these "positive control" comparisons, indicating that there were no SNPs that could differentiate the strain populations in the two samples (Figure S3, S4). Due to expected noise and errors in sequencing data, we set the lower bound for transmission in our HCT cohort at 99.999% popANI, equivalent to 30 population SNPs in a 3 megabase (Mb) genome. The same threshold was used to identify identical strains by the authors of inStrain[122].

Escherichia coli

While *E. coli* genomes in samples collected from the same patient over time were always more similar than the putative transmission threshold, in no case did we observe a pair of samples from different patients with $\geq$99.999% popANI (Table S8). This result suggests that all *E. coli* strains

observed are patient-specific, and argues that there are not common strains circulating in the hospital environment or passing between patients. Alternatively, patient-patient transmission or acquisition of common environmental strains is either notably rare or rapid genetic drift after a patient acquires a new strain is reducing popANI levels below the threshold. Deeper metagenomic sequencing or isolation and sequencing of *E. coli* strains may allow us to detect transmission in previously missed cases.

Enterococcus faecium

We performed the same analysis in *E. faecium* and observed four examples where two patients shared a strain with ≥99.999% popANI (Figure 4b,c). In one case, the two patients were roommates and direct transmission appears to be the most likely route. In the other three cases, epidemiological links were less clear, suggesting the patients may have acquired a similar strain from the hospital environment or through unsampled intermediates. In the following descriptions, samples are referred to by the day of collection, relative to the first sample from patients in the comparison.

Case 1: Patients 11342 and 11349 overlapped in the ward for 21 days and were roommates for 11 days (Figure 5a). Patient 11342 had a gut microbiome that was dominated by *E. faecium*; the two samples from this patient have 60% and 87% *E. faecium* relative abundance. A single sample from patient 11349 was obtained 14 days after starting to share a room with patient 11342. This sample is dominated by *Klebsiella pneumoniae*, and *E. faecium* is at 0.4% relative abundance. InStrain comparisons between the *E. faecium* strains in 11342 (the presumed "donor") and 11349 (the presumed "recipient") of the strain revealed 0-2 population SNPs (popANI 100% - 99.9999%) with 87% of the reference MAG (2.24 Mb) covered ≥5x in both samples. MAGs from each patient were also structurally concordant (representative dotplots in Figure S6a). These genomes were the most similar out of all *E. faecium* genomes compared from different patients. Samples from these patients were extracted in different batches and sequenced on different lanes, minimizing the chance that sample contamination or "barcode swapping" [115] (see Supplemental Note) could be responsible for this result. No other strains were shared between these two patients.

Case 2: Patients 11575 and 11568 overlapped on the ward for 36 days but were never roommates (Figure 5b). Samples from patient 11575 span 97 days, during which this patient experienced a BSI with *Klebsiella pneumoniae* and treatment with intravenous (IV) vancomycin, ciprofloxacin, meropenem (Figure S7c). Antibiotic treatment likely resulted in a reduction in microbiome diversity and domination by *E. faecium* in samples collected on days 16 and 28. Two samples were collected from patient 11568 on days 28 and 119. The first sample from 11568 was also dominated by *E. faecium*, but strains from the two patients were distinct (99.95% popANI). 91 days later, the second sample from 11568 has a lower relative abundance of *E. faecium* but a nearly identical strain to patient 11575. Five population SNPs (99.9997% popANI) were detected with 88% of the reference MAG covered ≥5x in both samples (representative dotplot in Figure S6b). This suggests that the *E.*

faecium strain in 11568 was replaced by a different strain with high identity to the strain in 11575. Patient 11568 was discharged from the HCT ward during the period between the two samples. The shared strain may represent an acquisition from a common environmental source or transmission from unobserved patients, rather than a direct transmission event between these two patients. While the *E. faecium* strain was different at the two time points from patient 11568, an E. faecalis strain remained identical.

Case 3: Patients 11605 and 11673 did not overlap in the ward (Figure 5c). Two samples were collected from 11605 on days 0 and 14. This patient experienced a BSI with *E. faecium* and treatment with meropenem (Figure S7e) prior to a sample dominated by the same species on day 14. Patient 11673 experienced a BSI with *E. coli* and treatment with vancomycin, meropenem and cefepime (Figure S7f) prior to the single sample we collected from this patient. Comparing *E. faecium* strains between the two patients revealed 2 population SNPs (99.9998% popANI) with 48% of the reference MAG covered ≥5x in both samples (representative dotplot in Figure S6c). Although slightly below the 50% coverage threshold, the high degree of similarity caused us to consider this result. While *E. faecium* strains in the two patients were nearly identical, the samples were collected 161 days apart and the patients had no overlap in the ward. This suggests both patients may have acquired the strain from the hospital environment, through transmission from unsampled patients, or another source such as healthcare workers.

Case 4: Patients 11360 and 11789 did not overlap in the ward. *E. faecium* remained at relatively low abundance in all samples. Comparing *E. faecium* strains between patients revealed 5-10 population SNPs (99.9993% - 99.9996% popANI) with 50%-57% genome coverage. Neither patient had a BSI during the sampling period. As these samples were collected at least 428 days apart, a shared source again may be the most likely explanation.

Comparisons with *E. faecium* and *E. coli* in published data

The *E. faecium* and *E. coli* strains we observe in our patients may be unique to this patient population and hospital environment. Alternatively, they may be hospital acquired strains that are present in other settings around the globe. We searched through several published datasets to differentiate between these possibilities. Our comparison dataset included metagenomic shotgun sequence data from 189 stool samples from adult HCT patients[144], 113 stool samples from pediatric HCT patients[76, 32, 158], 732 stool samples from hospitalized infants[123] and 58 vancomycin-resistant *E. faecium* isolates[67]. Sequence data were downloaded from SRA and processed in the same manner as other short-read data. Each sample was aligned against the *E. faecium* and *E. coli* MAGs used in the inStrain analysis above, profiled for SNPs, and compared against samples collected from our HCT patients. Comparisons within our data and comparisons within individual external datasets frequently achieved popANI values of ≥99.999%, typically from comparisons of samples from the same patient over time. Meanwhile, comparisons between our samples and external samples had

lower popANI values (Figure S5).

Comparisons of *E. faecium* strains in samples from patient 11346 in our dataset and patient 688 in the HCT microbiome dataset collected at Memorial Sloan Kettering Cancer Center[144] demonstrated a maximum of 99.9993% popANI (16 population SNPs detected in 2.3 Mb compared). While direct transmission is likely not involved here, this observation does align with the nearly identical *E. faecium* strains we observed in patients with no geographic or temporal overlap (case 3 and 4) and speaks to the global dissemination of vancomycin-resistant *E. faecium* ST 117. Comparing *E. coli* to external datasets revealed a maximum of 99.996% popANI (200 population SNPs detected in 5.0 Mb compared).

4.3.9 Putative transmission of commensal bacteria

Next, we extended the inStrain analysis to compare all species that were present in multiple patients. We found nearly identical genomes of commensal organisms that may be the result of transmission between patients, as well as several species shared between patients without clear explanations.

Hungatella hathewayi

Patients 11639 and 11662 overlapped in the ward for 34 days and were roommates for a single day, after which 11639 was discharged (Figure 6a). *Hungatella hathewayi* was at 5-10% relative abundance in the two samples from 11639. Patient 11662 developed Streptococcus mitis BSI on day 10 and was treated with IV vancomycin and cefepime (Figure S8b). The microbiome of this patient recovered with markedly different composition, including an abundant *H. hathewayi* strain reaching 54% and 17% relative abundance on days 58 and 100, respectively. Comparing *H. hathewayi* genomes between these two patients revealed 0-1 population SNPs (100% - 99.99998% popANI) with 94%-98% coverage $\geq$5x (6.9 - 7.1 Mb sequence covered in both samples). This was the single highest ANI comparison among all strains shared between patients. *H. hathewayi* MAGs from these patients were also structurally concordant and had few structural variations (Figure S6d). No other strains were shared between these patients.

Patient 11662 had *H. hathewayi* in the first two samples at 1.2% and 0.3% relative abundance, respectively. Although we were limited by coverage, comparing early to late samples with inStrain revealed 472 population SNPs in 3% of the genome that was covered at least 5X, implying 11662 was initially colonized by a different *H. hathewayi* strain, which was eliminated and subsequently replaced by the strain present in 11639. Given that samples were collected weekly, determining the direction of transmission is challenging. However, 11639 to 11662 appears to be the most likely direction, given the sampling times and perturbation 11662 experienced. However, it is possible that transmission occurred in the opposite direction or from a common source. Interestingly, 11662 was also re-colonized with *Flavonifractor plautii* in later samples. This strain was different from the

strain in earlier samples from this patient, as well as all other Flavonifractor strains in our sample collection.

H. hathewayi is known to form spores and is able to persist outside a host for days[24]. Although these patients were only roommates for a single day, 11662 remained in the same room for 4 days after 11639 was discharged, increasing the chance that a *H. hathewayi* spore could be transmitted from a surface in the shared room or bathroom. The question remains as to why transmission of *H. hathewayi* is not more common, given it is found at $\geq$1% relative abundance in 31 patients. Perhaps the earlier colonization of the microbiome of 11662 with a different *H. hathewayi* strain was key - the microbiome in this patient was "primed" to receive a new strain of the same species, despite the significant perturbation this patient experienced.

Notably, *H. hathewayi* was recently reclassified from *Clostridium hathewayi*[75], and was previously shown to induce regulatory T-cells and suppress inflammation[11]. Although the interaction of this microbe in HCT is not known, it may be interesting to investigate further given that the microbe may be transmitted between individuals and may contribute to inflammation suppression that may be relevant in diseases such as graft-vs-host disease. However, *H. hathewayi* may not be entirely beneficial or harmless and has been reported to cause BSI and sepsis in rare cases[95, 188].

Akkermansia muciniphila

Patients 11742 and 11647 overlapped in the ward for 11 days and were roommates for nine days (Figure 6b). Patient 11647 experienced a BSI with *Klebsiella pneumoniae* (perhaps related to previous *K. pneumoniae* domination of the microbiome) and was treated with piperacillin-tazobactam and cefepime (Figure S8d). The final sample from 11647 has *Akkermansia muciniphila* at 9.4% relative abundance, while the single sample from 11742 was dominated by *A. muciniphila* (85% relative abundance). Comparing these genomes revealed 0 population SNPs and 7 consensus SNPs with 90% coverage, as well as concordant MAGs from each sample (Figure S6e). No other strains were shared between these two patients.

In contrast to *H. hathewayi*, *A. muciniphila* is not known to form spores, which may reduce the chance of this microbe being transmitted. However, it is an aerotolerant anaerobe that may survive in oxygen for short periods of time[139]. The microbiome domination of 11742 with *A. muciniphila* and the relatively long overlap period of nine days in the same room may provide a greater "infectious dose" (abundance * exposure time) to the recipient patient.

4.3.10 Widespread strain sharing of commercially available probiotic organisms

Several organisms were found with identical or nearly identical genomes across multiple patient microbiomes without clear epidemiological links. The largest clade was found for *Lactobacillus rhamnosus*, which included 11 samples collected from eight patients over a span of 2.5 years (Figure

7a,d). All 11 samples in this clade had pairwise popANI of ≥99.999%, and in a subset of eight samples from seven patients, all pairs were identical from a popANI perspective (100%). Of the eight patients in this clade, only two pairs were roommates or overlapped in the hospital (patients 11537/11547 and 11647/11662, roommates for three days and one day, hospital overlap for 20 and 44 days, respectively). All 11 samples in this clade were collected after HCT (median time from HCT to first sample 43 days, range 12-93 days), and *L. rhamnosus* always had <0.1% relative abundance in pre-HCT samples, when present. Five of eight patients were discharged from the hospital after HCT and prior to acquiring *L. rhamnosus*, which we observed in a sample collected during a subsequent admission. We also observe *L. rhamnosus* falling below 0.1% relative abundance in a subsequent sample in five patients, suggesting that this strain may be a transient colonizer of the microbiome (Figure 7E). We evaluated antibiotic prescriptions in these patients and found that acquisition and loss of *L. rhamnosus* typically occurs independently of antibiotic use. For example, in patient 11537, *L. rhamnosus* is first detected and expands to 23.4% relative abundance while the patient is prescribed ciprofloxacin and cefepime. *L. rhamnosus* declines in relative abundance after the antibiotic prescription ends (Figure S9). Similar clades of high-identity genomes from different patients were found for *Lactobacillus gasseri* (Figure 7b) and *Streptococcus thermophilus* (Figure 7c).

Given that we did not observe hospital or roommate overlap between most patients in the *L. rhamnosus* cluster, the most likely explanation is that patients acquired this strain from a common source. *L. rhamnosus* is a component of several commercially available probiotic supplements, is present in certain live active-culture foods such as yogurt, and is among the most commonly prescribed probiotic species in US hospitals[197]. However, HCT recipients were not allowed to take probiotics or consume high-bacteria dairy products, such as probiotic yogurt or soft cheese, while inpatients on the HCT ward. We also verified that no prescriptions were written for probiotics by examining electronic health records. A majority of patients were discharged from the hospital between HCT and acquiring the *L. rhamnosus* strain, which may have provided them with the opportunity to consume a probiotic supplement or dairy product. Contact with a family member or other individual who had the strain in their microbiome could also be responsible for colonization of the HCT patient.

If this *L. rhamnosus* strain is a commonly used probiotic supplement or is found in commonly consumed dairy products, it may be found in other gut microbiome sequencing datasets. Comparing MAGs from this cluster against all Genbank genomes revealed a maximum alignment-based ANI of 99.95% to *L. rhamnosus ATCC 8530*[134]. Instrain-based comparisons against this reference had a maximum popANI of 99.98%, below the putative transmission threshold. We then searched against all genomes in the Unified Human Gastrointestinal Genome collection41 and identified two genomes that were nearly identical to the strain found in HCT patients. These genomes were originally from the Human Gastrointestinal Bacteria Culture Collection[51] (accessions ERR2221226

and ERR1203919, belonging to the same isolate per a personal communication with the authors). Assembled isolate and patient-derived genomes had ≥99.99% ANI; inStrain-based SNP comparisons had ≥99.999% popANI. This suggests that a *L. rhamnosus* strain that is nearly identical to the genomes in our HCT patients has been isolated from human stool in the past.

4.4 Discussion

Our investigation using high-resolution metagenomic sequencing attempts to quantify if and when patient-patient microbiome transmission is involved in the spread of pathogenic organisms. We first found that hospitalized HCT patients frequently harbor HAI organisms in their gut microbiome, validating previous studies which used culture-based approaches or 16S rRNA sequencing[170, 7, 136]. MAGs created from patient samples had high identity to several globally disseminated and antibiotic resistant sequence types, including *Escherichia coli* ST131 and ST648. Interestingly, whereas ST131 is a well-recognized multi-drug resistant pathogen, ST648 is nearly as prevalent as ST131 in our sample collection, and may thus be an emerging pathogen in this patient population.

Despite the high degree of temporal and spatial overlap between patients in our study, we found no association between these factors and the taxonomic similarity of patient microbiomes, and only a weak association between hospital overlap and maximum strain identity. These findings suggest that patient-patient transmission is not driving microbiome composition, but individual strains may still be shared between patients. We did not identify any pairs of patients harboring *E. coli* strains with popANI values above the 99.999% popANI transmission threshold. Taken together with the observation that *E. coli* is commonly detected in the patient's microbiome upon admission, this finding argues that patients usually enter the hospital with an "individual-specific" *E. coli* strain and do not frequently transmit it to others. An exclusion principle may be at play, where an *E. coli* niche can only be filled by a single strain and new strains are unlikely to engraft when the niche is already occupied. In contrast to *E. coli*, we observed four pairs of patients with *E. faecium* strains that were more similar than 99.999% popANI. In one case, the two patients spent 11 days sharing a room and bathroom prior to observing the shared strain. Direct links between patients were less clear or non-existent in the other three cases, and transmission through unsampled intermediates or acquisition from an environmental source may have been responsible. We also found evidence for an *E. faecium* strain in a published dataset from HCT patient microbiomes[144] that had above 99.999% popANI to a strain in our sample collection. While this finding is likely not the result of patient-patient transmission, it does indicate that very similar strains may exist within patients in different geographic locations.

We then expanded the transmission analysis to examine all species that were present in the microbiome of multiple patients. Identical *Hungatella hathewayi* strains were found in two patients who were in the hospital together for 34 days and roommates for a single day. Earlier samples

from patient 11662 had a significantly different *H. hathewayi* strain than the strain present in later samples. It is possible that the earlier colonization with the same species exhibited a priority effect[52] and primed this individual to be re-colonized. In another set of patients who overlapped in the hospital for 11 days and were roommates for nine days, we identified identical *Akkermansia muciniphila* strains. In both examples, the likely "recipient" patient experienced BSI prior to the putative transmission event. The subsequent antibiotic treatment initiated for the treatment of BSI resulted in vast microbiome modification and simplification, which may have opened a niche for the new organism to engraft into. Both *H. hathewayi* and *A. muciniphila* can survive outside the host for periods of time, but *H. hathewayi* can form spores that enable it to live in aerobic conditions for days[24]. In the case of *H. hathewayi* transmission, patient 11662 remained in the same room with a shared bathroom after the single day of overlap with patient 11639. Spores surviving on surfaces may be responsible for transmission, given the relatively short period of overlap. In these cases, patient-patient transmission may help in the recovery of microbiome diversity following BSI and may play a role in ameliorating post-HCT inflammatory processes, such as acute graft-vs-host disease. Finally, we observed identical probiotic species in multiple patients without clear geographic or temporal links, including *Lactobacillus rhamnosus*, *Lactobacillus gasseri*, and *Streptococcus thermophilus*. Acquisition from a commercially available probiotic or live-active culture food appears to be the most likely explanation. While patients hospitalized for HCT were not allowed to consume probiotics or high-bacterial dairy foods, a majority of patients were discharged after HCT and prior to a subsequent admission, upon which *L. rhamnosus* was detected. Many patients lost the strain in later samples, independent of antibiotic prescription, suggesting that *L. rhamnosus* was a transient colonizer. This matches the observation that abundance of probiotic species in the gut often declines after supplementation ends[167].

The healthy adult gut microbiome is relatively resistant to perturbation and colonization with new strains or species[99]. In contrast, mother-to-infant transmission of bacteria and phages is common and well-described[13, 159, 195]. Patients in our study often shared spaces, were exposed to dramatic "niche clearing" therapies and were often immunosuppressed. We frequently observed patients acquiring new organisms into their gut microbiome during their hospital stay, especially following BSI. Still, we found that patient-patient transmission of gut microbes is relatively rare. This suggests that age, rather than perturbation or microbial exposure, may play the largest role in microbiome transmission. There are also several alternative explanations for the relative lack of transmission between patients. First, the adult gut microbiome may remain densely colonized even when dramatically perturbed by antibiotics and chemotherapy, and thus resistant to invasion with new strains. Strains we observed in later samples may have existed at low very levels in earlier samples from the individual, therefore evading detection. Second, it is possible that the microbiome of healthcare workers, hospital visitors, or other staff serve as the source of newly colonizing strains. Third, it is possible that the built environment, equipment used in the care of these individuals,

and other environmental sources such as food and personal items harbored the microbes that were later transmitted. As we did not sample these other potential sources, it is difficult to know the extent to which they contributed to the collective reservoir of potentially transmitted organisms. Fourth, genetic drift and adaptive evolution may rapidly act on newly acquired microbes, as has been described for *E. faecium* in the human gut[43] and *E. coli* in the mouse gut[56, 126, 14]. Rapid drift or evolution would move the genomes of transmitted microbes below the popANI transmission threshold. Finally, transmitted organisms may be killed by antibiotics before they can establish a community within the host and thus be detected with metagenomic sequencing.

Our findings have important implications for hospital management and infection prevention. 55/149 patients (37%) in our study experienced BSIs, which is comparable to the rate of BSI in other transplant centers[33]. Our findings suggest that microbiome transmission does not play a large role in spreading infections among HCT patients, and that established contact precautions and procedures for patient isolation were working as intended. Recently, the HCT ward at our hospital moved to a new location with exclusively single rooms, which may further reduce the opportunity for transmission.

Our analysis of transmission of microbes between HCT patients does have several limitations. We analyzed hundreds of samples collected over many years, but most sampling was done on a weekly basis. We did not explicitly collect samples on the day of admission or discharge. Our sample collection also ignores previous hospital stays, either in a different ward in our hospital, or other hospitals entirely, that may be responsible for the acquisition of HAI organisms. Many patients in our study had one or two sequenced samples, limiting our inference about microbiome changes over time. By contrast, the second largest study using shotgun metagenomics to study the microbiome of HCT patients had much more dense sampling, analyzing 395 samples from 49 patients[144]. We also did not perform any sampling of the hospital environment, healthcare workers or visitors, which would allow us to track transmission patterns in detail and more conclusively state where newly acquired microbes originated[22]. Our work is also entirely based on metagenomic sequencing data, which has its own challenges and sources of bias, including "barcode swapping"[115], which could contribute to false positive transmission findings. To address this, we measured the impact of barcode swapping in linked-read data, and eliminated linked-read and short-read comparisons where a finding of identical strains could be the result of barcode swapping (see supplemental note). Additionally, metagenomic sequencing may fail to detect lowly-abundant colonizers, especially when samples are contaminated with host DNA[133], which is often the case in stool samples from HCT patients. These challenges undoubtedly affect our sensitivity and specificity in measuring acquisition and transmission of both pathogenic and commensal microbes. While our comparison methods were sensitive to strain populations in the gut microbiome, we did not attempt to phase strain haplotypes. Haplotype phasing with long-read sequencing technology like Nanopore or PacBio[18, 110, 179] could help us determine whether sets of SNPs occurred in the same or different strains.

Our study leaves several questions unanswered that we hope future work on microbiome transmission will attempt to answer. First, our findings need to be validated in an external cohort in a different hospital. Collecting stool samples during an infection outbreak may lead to more transmission events being identified and may implicate microbiome transmission in perpetuating the outbreak. Similar experiments in a pediatric patient population may reveal more gut-to-gut transmission, as young children have microbiomes that are still developing and more susceptible to colonization with new species. Our work did not investigate any possible sources of microbial transmission other than the gut microbiota of HCT patients. A more granular study where samples are collected from the hospital environment, as well as hand swabs and stool samples from healthcare workers, visitors, and family members, is clearly indicated by these early results. As more transmission events are observed with high-resolution genomic methods, we will start to uncover the general principles governing community assembly in the human microbiome. These new insights may help prevent infections and other co-morbidities in this patient population in the future.

4.5 Methods

4.5.1 Cohort selection

Hematopoietic cell transplantation patients were recruited at the Stanford Hospital Blood and Marrow Transplant Unit under an IRB-approved protocol (Protocol #8903; Principal Investigator: Dr. David Miklos, co-Investigator: Drs. Ami Bhatt and Tessa Andermann). Informed consent was obtained from all individuals whose samples were collected. Stool samples were placed at 4°C immediately upon collection and processed for storage at the same or following day. Stool samples were aliquoted into 2-mL cryovial tubes and homogenized by brief vortexing. The aliquots were stored at -80°C until extraction.

We identified all samples that had been sequenced previously by our group. Samples were selected for linked-read sequencing to augment this collection. We examined the network of patient roommate overlaps to find cases where we were likely to uncover transmission events, if they were happening. These included patient pairs from whom we ideally had samples before and after the roommate overlap period. 96 samples that provided the best coverage of roommate overlaps were selected for linked-read sequencing.

The following clinical data were extracted from the electronic health record: demographic information, underlying disease, type of transplantation (allogeneic vs. autologous), date and type of bloodstream infection, medication prescriptions, time of admission and discharge and location of patients (rooms) over time. Hospital-wide BSI data were obtained from an electronic report generated by the clinical microbiology laboratory. Medication prescription data was filtered by the following criteria:

1. Only entries for antibiotics, antifungals, antivirals, antibacterials, and *Pneumocystis jirovecii*

pneumonia prophylaxis were retained.

2. Medications with a missing start or end date were excluded.
3. Medications with a frequency of "PRN" (pro re nata, or as needed) or a prescription status of "Canceled" were excluded.
4. Medications with a difference between start and end date of less than one day were excluded.
5. Medications prescribed to be taken by eyes, ears, topical application, or "swish and spit" were excluded.
6. Medication prescriptions occurring outside the window of HCT date +/- 100 days were excluded for the aggregated analysis.

We do acknowledge the challenge of working with electronic health record data, and recognize that there is a disconnect between medications prescribed and medications consumed by a patient.

4.5.2 DNA Extraction, library preparation and sequencing

DNA was extracted from stool samples using a mechanical bead-beating approach with the Mini-Beadbeater-16 (BioSpec Products) and 1-mm diameter zirconia/silica beads (BioSpec Products) followed by the QIAamp Fast DNA Stool Mini Kit (Qiagen) according to manufacturer's instructions. Bead-beating consisted of 7 rounds of alternating 30 s bead-beating bursts followed by 30 s of cooling on ice. For samples subjected to linked-read sequencing, DNA fragments less than approximately 2 kb were eliminated with a SPRI bead purification approach[135] using a custom buffer with minor modifications: beads were added at 0.9×, and eluted DNA was resuspended in 50 μl of water. DNA concentration was quantified using a Qubit fluorometer (Thermo Fisher Scientific). DNA fragment length distributions were quantified using a TapeStation 4200 (Agilent Technologies).

Short-read sequencing libraries were prepared with either the Nextera Flex or Nextera XT kit (Illumina) according to manufacturer's instructions. Linked-read sequencing libraries were prepared on the 10X Genomics Chromium platform (10X Genomics). Linked-read libraries have a single sample index, and were pooled to minimize the possibility of barcode swapping between samples from patients who were roommates (see supplemental note). Libraries were sequenced on an Illumina HiSeq 4000 (Illumina).

4.5.3 Sequence data processing

TrimGalore version 0.5.0[114] was used to perform quality and adapter trimming with the flags "–clip_R1 15–clip_R2 15–length 60". SeqKit version 0.9.1[152] was used to remove duplicates in short-read data with the command "seqkit rmdup–by-seq". Due to excessive processing time, this step was skipped for linked-read data. Reads were mapped against the GRCh38 assembly of the

human genome using BWA version 0.7.17-r1188[87] and only unmapped reads were retained. Quality metrics were verified with FastQC version 0.11.8[12]. Bioinformatics workflows were implemented with Snakemake[80].

4.5.4 Short-read classification with Kraken2

We classified all short-read data with a Kraken2[189] database containing all bacteria, viral and fungal genomes in NCBI GenBank assembled to complete genome, chromosome or scaffold quality as of January 2020. Human and mouse reference genomes were also included in the database. A Bracken[100] database was also built with a read length of 150 and k-mer length of 35. Classification results were processed into matrices and taxonomic barplots with the workflow available at https://github.com/bhattlab/kraken2_classification. Bray-Curtis distances were calculated with the R package vegan[119] version 2.5-7.

4.5.5 Assembly and binning

Short-read sequencing samples were assembled using SPAdes version 3.14.0[118] using the '–meta' flag. Linked-read sequencing samples were assembled with Megahit version 1.2.9[86] to generate seed contigs, which were then assembled with the barcode-aware assembler Athena[19]. Metagenome-assembled genomes (MAGs) were binned with Metabat2 version 2.15[72], Maxbin version 2.2.7[192] and CONCOCT version 1.1.0[6] and aggregated using DASTool version 1.1.1[157]. MAG completeness and contamination was evaluated using CheckM version 1.0.13[129] and MAG quality was evaluated by the standards set in[20]. All assembled contigs were classified with Kraken2 as described above. To generate bin identifications, contig classifications were pooled such that contigs making up at least two thirds the length of the bin were classified as a particular species. If a classification could not be assigned at the species level, the process was repeated at the genus level, and so on.

4.5.6 Genome de-replication and SNP profiling

MAGs were filtered to have minimum completeness 50% and maximum contamination 15% as measured by CheckM, then were de-replicated with dRep version 2.6.2[121] with default parameters except the primary clustering threshold set to 0.95. In further steps, a single de-replicated genome will be referred to as a cluster. Reads from all samples were mapped against the de-replicated set of genomes with BWA. Clusters that had greater than 1x average coverage in at least two samples were retained for further analysis. Individual bam files were extracted for each sample-cluster pair with at least 1x coverage. Bam files were randomly subsetted to a maximum of 2 million reads for computational efficiency. Alignments were profiled and then compared across samples with inStrain

version 1.3.11[122] using default parameters. A Snakemake workflow for dRep and inStrain analysis is available at https://github.com/bhattlab/bhattlab_workflows.

4.5.7 Building phylogenetic trees

MAG Average Nucleotide Identity (ANI) trees (Figure 2 and 3) were created using the pairwise alignment values from dRep, which uses the MUMmer program[83]. MAGs were filtered to be at least 75% the length of the mean length of reference genomes used in the tree. Reference genomes were selected by searching literature for collections of well-described isolates with genomes available. References that were not relevant and clustered in isolated sections of the tree were removed. Pairwise ANI values were transformed into a distance matrix and clustered using the 'hclust' function with the 'average' method in R version 4.0.3[173]. Heatmaps were created using pairwise popANI values from inStrain, transformed into a distance matrix, and hierarchically clustered using the 'ward.D2' method.

4.5.8 Determining transmission thresholds

To determine the ANI threshold to call a comparison a "putative transmission event" we evaluated the distributions of ANI values for within- and between-patient comparisons for different species (Figure S3). We often detected zero population SNPs in time course samples from the same patient, including *E. faecium* in a pair of samples collected from the same patient 323 days apart. Meanwhile, between-patient comparisons typically had lower ANI values. To verify that transmission events would also result in population ANI values near 100%, we examined external datasets where transmission of bacteria in the microbiome is known to occur as a "positive control". We gathered sequencing data from stool samples of matched mother-infant paris[195] and fecal microbiota transplantation donors and recipients[162] and processed them with the same methods. In these datasets, we regularly observed genomes with 100% popANI between matched individuals, and did not find cases of 100% popANI between unmatched individuals (Figure S4). In the ideal cases, we expect transmission of bacteria between the microbiomes of HCT patients to result in genomes with 100% popANI. However, the measured genomes may not reach this level of identity, due to mutations or genetic drift since the transmission event, sequencing errors, or other factors. Therefore, we set the transmission threshold at 99.999% popANI, equivalent to 30 population SNPs in a 3 megabase (Mb) genome. Although this threshold is stringent, we recognize that it may allow for false positives where two closely related strains exist in different patients solely by chance.

Despite our efforts to minimize the impact of barcode swapping on detecting transmission (see Supplemental Note), we still identified many comparisons with >99.999% popANI that we believe to be false positives. These were filtered out based on the following criteria: short read samples from different patients that were sequenced on the same lane and shared one index sequence or linked read samples that were sequenced on the same lane and reads mapping to the organism shared >40% of

barcodes. We also removed pairs that could be affected by "secondary" swapping, where the two samples were not directly affected, but an interaction between other samples from the two patients could cause false positives. In total, we removed 31 comparisons from the final table with >99.999% popANI.

4.5.9 Pairwise MAG comparison

MAGs were aligned with the mummer program using default settings[83] and filtered for 1-1 alignments. Dotplots were visualized with the "Dot" program[113] filtering for non-repetitive alignments ≥1 kb.

4.5.10 Antibiotic resistance gene detection

Antibiotic resistance genes (ARGs) were profiled in contigs from all samples using Resistance Gene Identifier (RGI) and the Comprehensive Antibiotic Resistance Database (CARD)[4] with default parameters. Genes were counted if they met the "strict" or "perfect" threshold from RGI. ARGs were annotated both if they occurred on a contig in the MAG of interest, or anywhere in the metagenomic assembly.

4.5.11 Isolation, culture and identification of VRE organisms

Stool samples from patients 11342 and 11349 were resuspended in glycerol and streaked on SpectraVRE (R01830, ThermoFischer Scientific) plates and incubated at 35 °C overnight. The following day, four colonies from each of the plates that displayed growth were picked and streaked out on a separate quadrant of fresh SpectraVRE plates. These Round 1 (R1) plates were incubated at 35 °C overnight and checked for growth the following day. For each R1 plate, a colony was picked from each quadrant and streaked out on a new quadrant of a fresh SpectraVRE plate. These Round 2 (R2) plates were incubated at 35 °C overnight and checked for growth the following day. For each R2 plate, a colony from each quadrant was subjected to MALDI-TOF bacterial species identification analysis on a Bruker Biotyper (Burker) per manufacturer instructions

4.6 Supplemental note: Mitigation of laboratory contamination and barcode swapping

Any study of transmission is susceptible to confounders that may introduce false positives. Two major sources are laboratory contamination and barcode swapping, both of which can make it appear as if identical strains were present in multiple samples. To minimize the chance of laboratory contamination, samples selected for linked-read sequencing were randomized prior to extraction into groups of 16, subject to the constraint that the number of samples from roommate pairs in the same extraction batch were minimized. These groupings were carried out through library preparation. Similar constraints were used when preparing pooled libraries for sequencing.

It is a recognized phenomenon that pooled Illumina sequencing libraries experience "barcode swapping" or "index hopping"[115] when libraries are differentiated by a single sample index. While this issue is avoided by using unique dual index sequences for all samples in a pool, our laboratory was not aware of the issue until 2018, and older libraries were prepared without a unique dual indexing strategy. Linked-read libraries only contain a single sample index sequence, which makes it impossible to eliminate the effect of barcode swapping, other than the costly option of devoting an entire lane to each sample.

In linked-read sequencing libraries, we were able to estimate the impact of barcode swapping. There are 10 million possible 10X barcodes (these are the barcodes which convey long-range information, different from the sample index barcodes). While a subset of 10X barcodes will overlap between two samples, the fraction of barcodes from reads mapping to a single organism should be limited. We mapped reads from all linked-read samples against the uniquely identifiable p-crAssphage genome[44]. Then we looked at the fraction of 10X barcodes that overlapped between samples. Samples sequenced on different lanes typically had 0-30% 10X barcode overlap. Samples sequenced on the same lane had 60-100% of barcode overlap in some cases. EC95853 We set a threshold of 40% overlap of barcode sets to call a comparison "swapped" and remove it from analysis. By counting reads believed to be assigned to improper samples because of barcode swapping, we estimate the rate in our linked-read data to be 0.1-0.2%.

While this rate may seem small, at high sequencing depth and with abundant organisms, it quickly results in enough reads being swapped to assemble a genome or conduct an inStrain comparison. Indeed, we found cases where multiple species (instead of the single species believed to be the result of transmission events) were shared between linked-read samples sequenced on the same lane that were likely the result of barcode swapping.

We also attempted to measure the degree of barcode swapping in dual-indexed lanes of short read Illumina sequencing. Using the uniquely identifiable p-crAssphage genome as a marker for swapping, we observed roughly 0.5% of sequencing reads swapped between samples on the same lane that shared one index sequence. Samples on the lane that shared no sequencing indices often had

p-crAssphage below 1e-5%. Simple relative abundance metrics cannot distinguish between barcode swapping and a true difference in abundance between samples. However, even with the 0.5% rate of swapping, we regularly observed >5x coverage of the p-crAssphage genome in what we believe to be the truly negative samples, and the resulting inStrain comparisons revealed sufficient paired genome coverage and 100% popANI. We never observed identical p-crAssphage genomes between samples from different patients sequenced with unique dual indices or on different lanes.

An example of p-crAssphage relative abundance measured with Kraken2 in a single lane of sequencing is shown in the figure below. Here, we believe the sample 11713_98 in blue is a "true positive" for p-crAssphage. While measured abundance in the other samples is likely some combination of true abundance and barcode swapping, the separation between samples that share one or zero index sequences is clear. For short-read sequencing samples, we know which pairs of samples share one of two index sequences and have the possibility of being impacted by swapping. We cannot estimate the impact of barcode swapping like was done for linked-read datasets. We simply eliminated all comparisons where two samples had the possibility of barcode swapping, and all comparisons that could be affected by "secondary" swapping, where the samples were not directly affected, but an interaction between other samples from the two patients could cause false positives. While this filtering may discard legitimate transmission events, we believe it is necessary to lower the number of false positives.

Previous DNA extraction and short-read sequencing efforts did not follow the randomization constraints above and we cannot guarantee that laboratory contamination did not happen at some point in the process. However, we note that cases of laboratory contamination or barcode swapping would result in the entire microbiome composition of one sample being transferred to another. After our stringent filters, we only discovered one case where patients shared two separate species. As these were both Lactobacillus species, our hypothesis about probiotic consumption is a possible explanation.

4.7 Figures

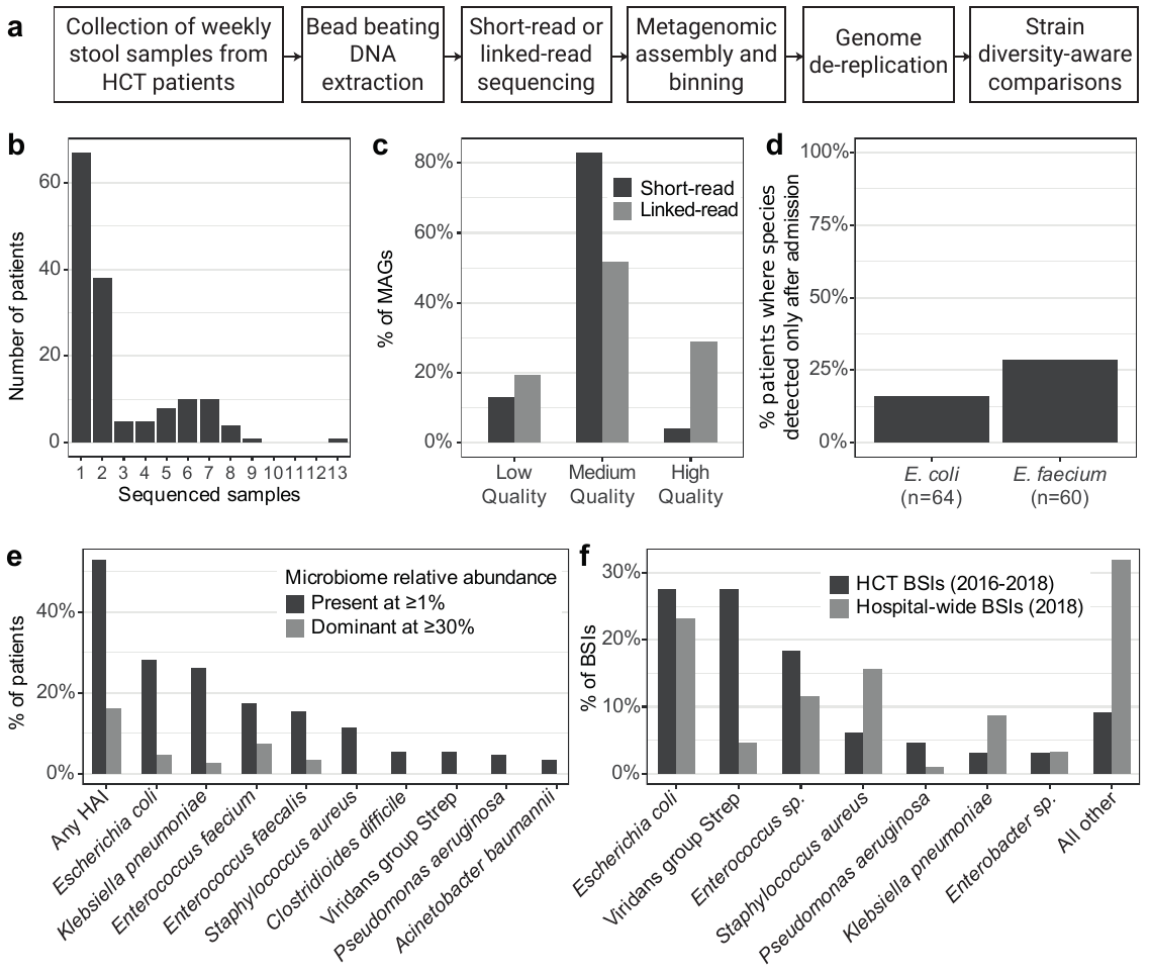


Figure 4.1: **Overview of the methods, data generated, and clinical features of this sample set. a)** Overview of the wet lab and computational workflow used to generate sequencing datasets, bin MAGs and compare strains between patients. **b)** Number of stool samples sequenced per patient. **c)** Percentage of MAGs meeting each quality level, stratified by sequencing method. **d)** Of patients who have the given organism detected (≥50% coverage breadth) in a time course sample, percentage of patients where the organism was below the detection threshold (<50% coverage breadth) in the first sample. **e)** Percentage of patients with at least one sample positive with (≥1% relative abundance) or dominated by (≥30% relative abundance) hospital acquired infection (HAI) organisms, as identified by Kraken2 and Bracken. **f)** Percentage of bloodstream infections (BSIs) identified with each organism or group in HCT patients and hospital-wide.

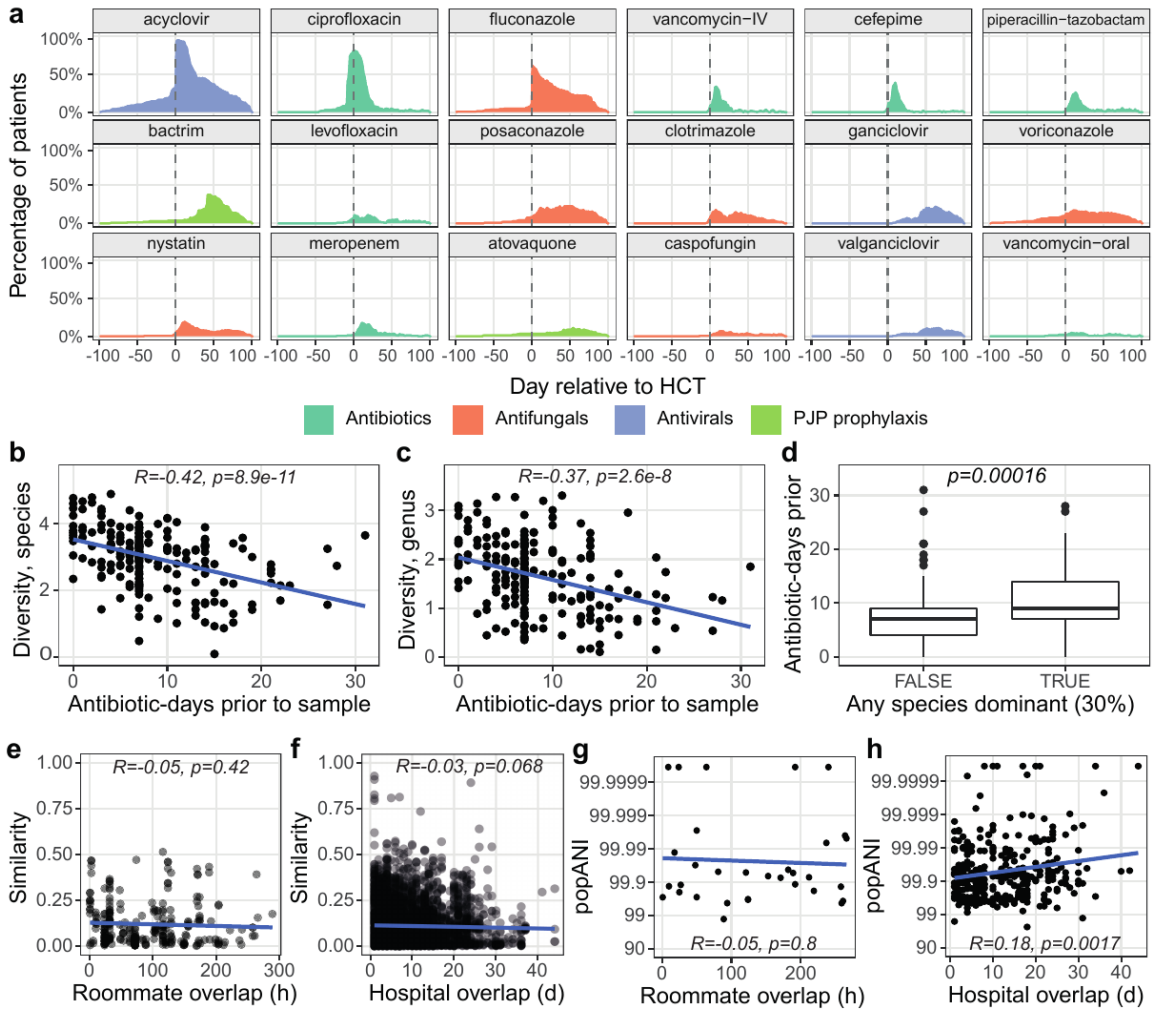


Figure 4.2: **The impact of antibiotic prescription and geographic overlap on patient microbiomes. a)** Aggregated prescription history of 20 of the most frequently prescribed antibiotic, antifungal and antiviral drugs. Each panel shows the percentage of patients who were prescribed a drug at the given day, relative to the date of HCT. Shannon diversity at the species **(b)** or genus **(c)** level compared to total antibiotic-days in the seven days prior to sample collection. **(d)** Samples with or without a single species dominant ($\geq$30%), compared with total antibiotic-days in the prior seven days. Taxonomic similarity at the species level (1 - Bray-Curtis dissimilarity) between samples from different patients, evaluated against days of hospital overlap **(e)** or hours of roommate overlap **(f)** prior to the sample. Maximum inStrain popANI achieved by comparing all strains in all samples from two patients, evaluated against hours of roommate **(g)** or days of hospital overlap **(h)** prior to the earlier sample. In all panels, trend lines are calculated as the best-fit linear regression between the X and Y variables. R and p values are the pearson correlation coefficient and correlation p-value, respectively.

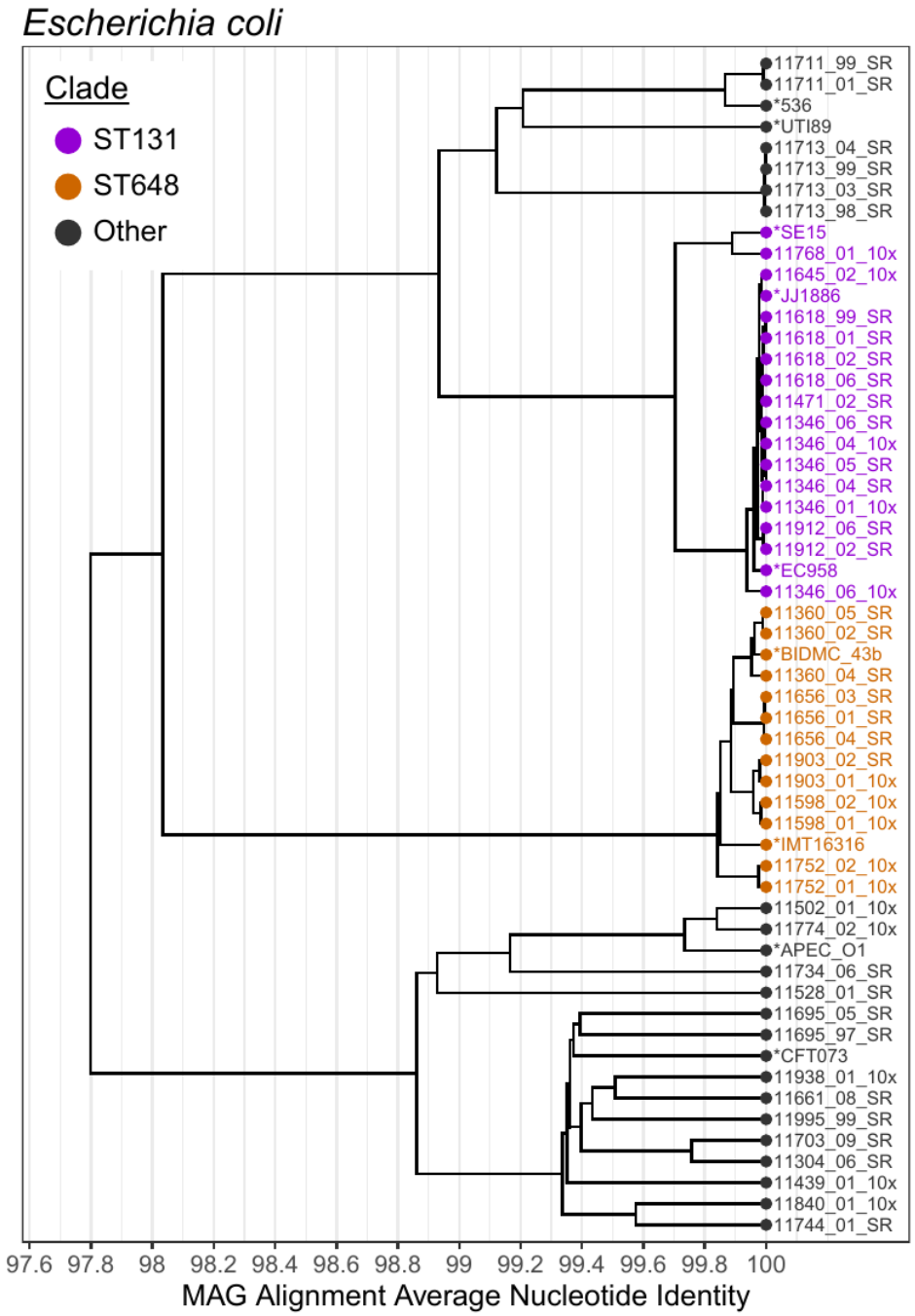


Figure 4.3: **Alignment average nucleotide identity (ANI) tree of *Escherichia coli* MAGs.** MAGs identified as *E. coli*, medium quality or above and at least 75% the mean length of the reference genomes are included. Several reference genomes are included and labeled with an asterisk. Clusters at the 99% ANI level corresponding to ST131 (purple) and ST648 (orange) are highlighted. Alignment values used to construct this tree can be found in Table S7.

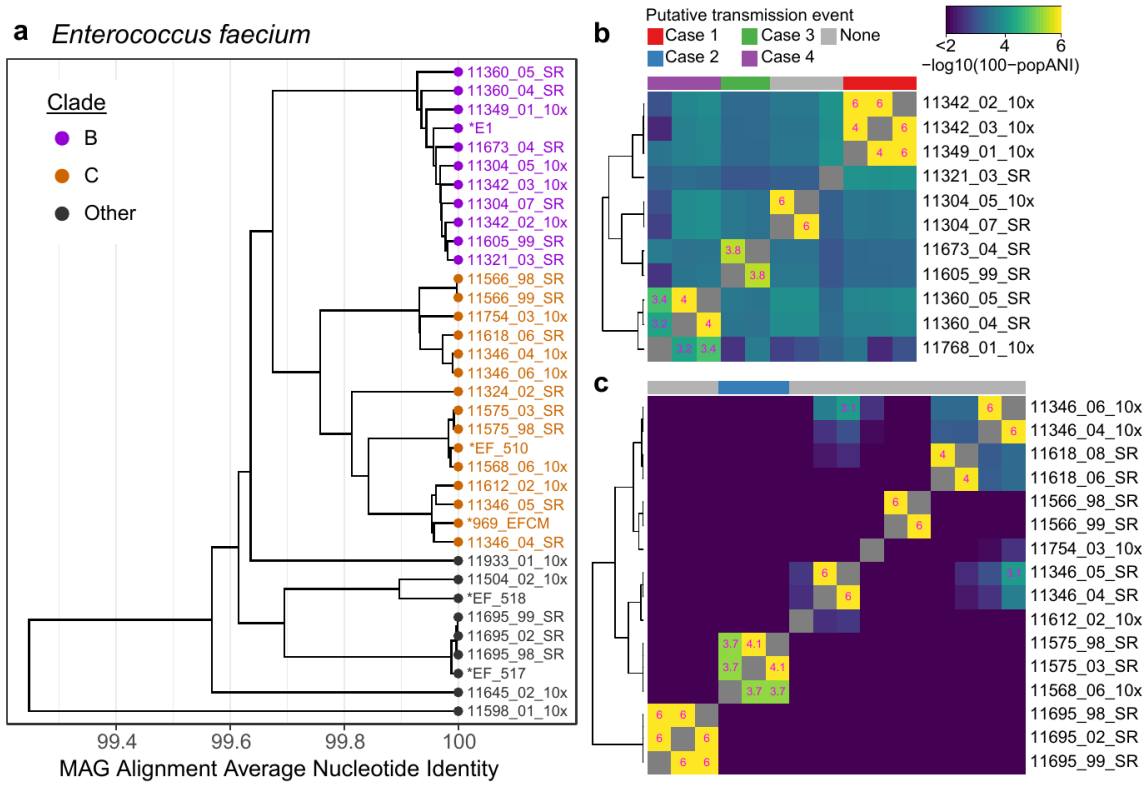


Figure 4.4: ***Enterococcus faecium* strains compared between patients. a)** Alignment average nucleotide Identity (ANI) based tree of *E. faecium* MAGs. MAGs identified as *E. faecium*, medium quality or above and at least 75% the mean length of the reference genomes are included. Several reference genomes are included and labeled with an asterisk. Two clades containing samples from multiple patients are highlighted for further comparison. Alignment values used to construct this tree can be found in Table S7. **b, c)** Heatmaps showing pairwise popANI values calculated with inStrain for clades B and C. Color scale ranges from 99.99-100% popANI and is in log space to highlight the samples with high popANI. Cells in the heatmap above the transmission threshold of 99.999% popANI are labeled. Four groups containing samples from multiple patients with popANI values above the transmission threshold are highlighted on the top of the heatmaps.

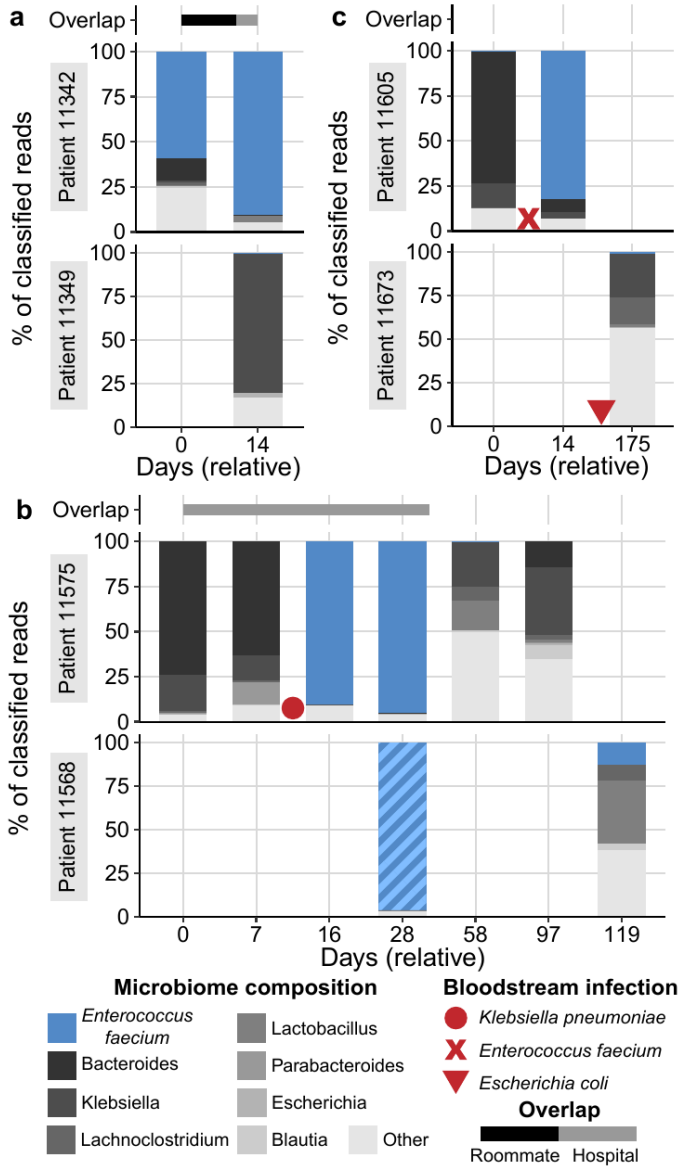


Figure 4.5: **Microbiome composition of patients with putative *Enterococcus faecium* transmission events.** Each panel shows the composition of two patients over time. The height of each bar represents the proportion of classified sequence data assigned to each taxon. Samples are labeled relative to the date of the first sample in each set. Bars above each plot represent the approximate time patients spent in the same room (black bars) or in the hospital (grey bars). Red symbols indicate approximate dates of bloodstream infection with the specified organism. Hypothesized direction of transmission progresses from the top to the bottom patient. Fractions of the bar with >99.999% popANI strains in each panel are indicated with solid colors, and different strains are indicated with hashed colors. All taxa except *E. faecium* are shown at the genus level for clarity. **a)** Case 1: Putative transmission from patient 11342 to 11349. **b)** Case 2: Putative transmission from patient 11575 to 11568. **c)** Case 3: Putative transmission from patient 11605 to 11673.

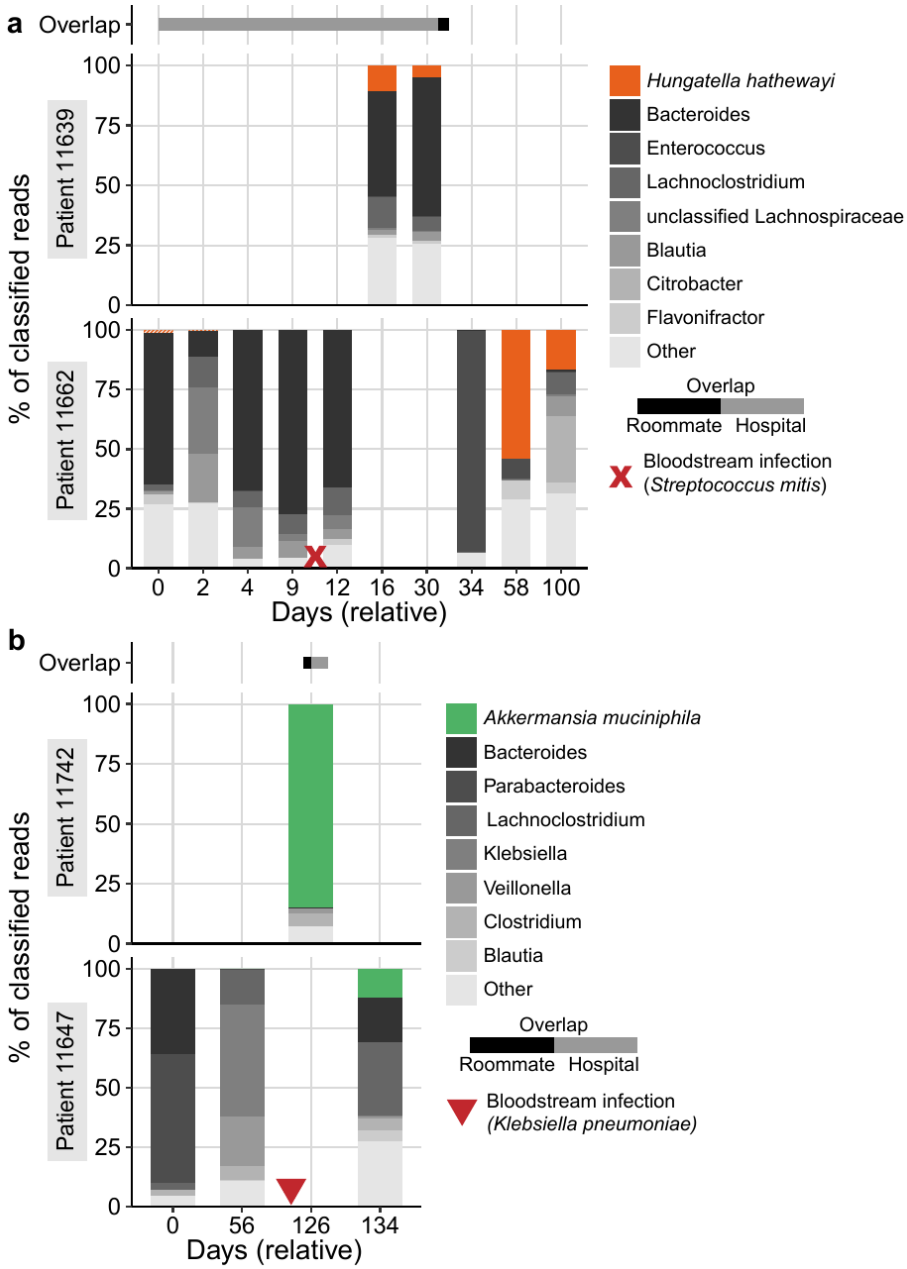


Figure 4.6: **Microbiome composition of patients with putative *Hungatella hathewayi* or *Akkermansia muciniphila* transmission events.** Each panel shows the composition of two patients over time. The height of each bar represents the proportion of classified sequence data assigned to each taxon. Samples are labeled relative to the date of the first sample in each set. Bars above each plot represent the approximate time patients spent in the same room (black bars) or in the hospital (grey bars). Red symbols indicate approximate dates of bloodstream infection with the specified organism. Hypothesized direction of transmission progresses from the top to the bottom patient. Fractions of the bar with >99.999% popANI strains in each panel are indicated with solid colors, and different strains are indicated with hashed colors. All taxa except *H. hathewayi* or *A. muciniphila* are shown at the genus level for clarity. **a)** Putative case of *H. hathewayi* transmission from 11639 to 11662. **b)** Putative case of *A. muciniphila* transmission from 11742 and 11647.

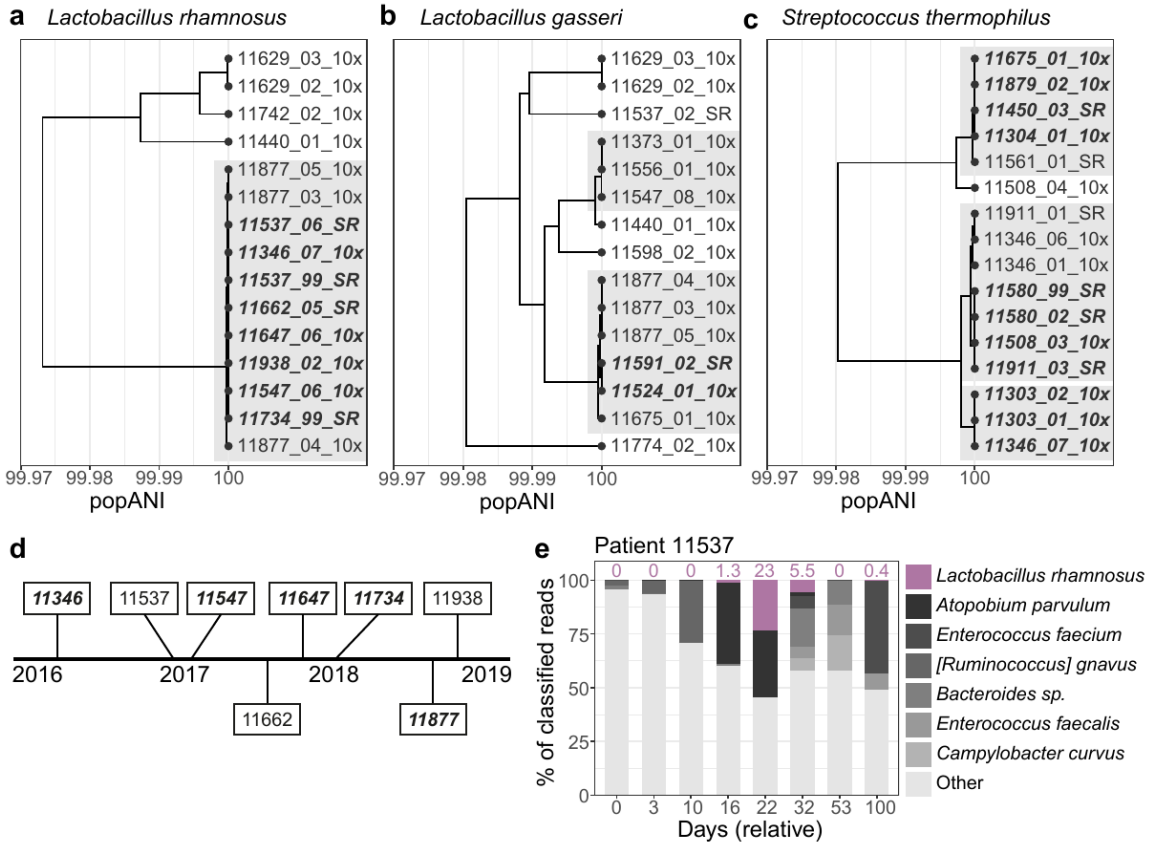


Figure 4.7: **Lactobacillus and Streptococcus strains are acquired after HCT and identical between many patients.** PopulationANI based tree of **(a)** *Lactobacillus rhamnosus*, **(b)** *Lactobacillus gasseri*, **(c)** *Streptococcus thermophilus* strains present in patient samples. Clades containing samples from different patients with ≥99.999% popANI are highlighted with a grey background. Clades with 100% popANI between all pairs are additionally bolded and italicized. **d)** Timeline of approximate date of samples containing a *L. rhamnosus* strain in the transmission cluster in (a). Patients who were discharged from the hospital after HCT and prior to acquiring *L. rhamnosus* are bolded and italicized. **e)** Microbiome composition of patient 11537. *L. rhamnosus* abundance at each time point is indicated above the bar. This patient received HCT on relative day 3.

4.8 Supplementary Figures

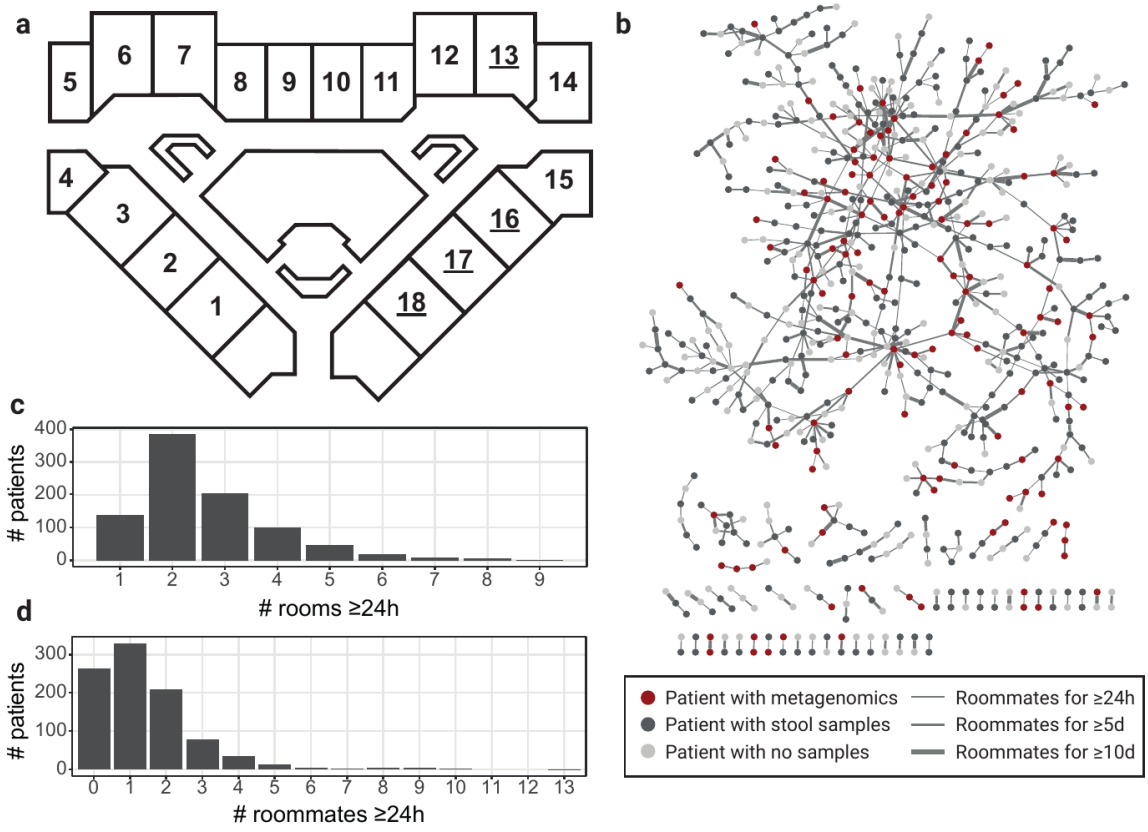


Figure 4.8: **Analysis of hospital geography.** **a)** Layout of rooms in the HCT ward. Room numbers are indicated and double occupancy rooms are underlined. **b)** Network view of patients who were roommates for at least 24 hours. Each node represents a single patient, colored according to if they have a banked stool sample or metagenomic sequencing data present. Edges are drawn between patients who were roommates, and edge width represents the length of overlap in the same room. **c)** Histogram of the number of rooms patients occupied for at least 24 hours. **d)** Histogram of the number of unique roommates patients had for at least 24 hours.

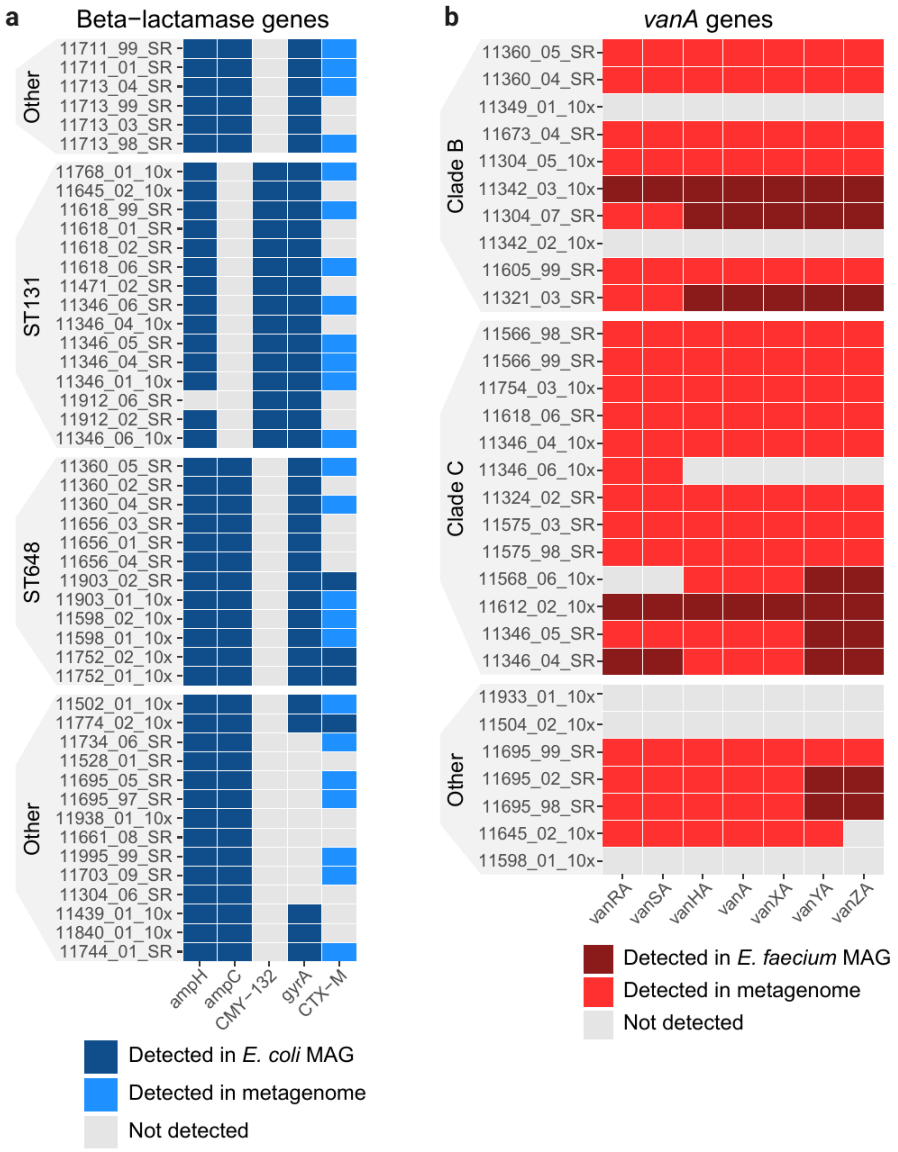


Figure 4.9: **Antibiotic resistance genes detected in HCT patient microbiome samples.** In each panel, samples are rows and resistance genes are columns. Samples are ordered and clades are highlighted corresponding to the respective figure in the main text. Cells are colored whether the gene was detected in the respective MAG from the sample, or just in the metagenome (indicating it may be on a plasmid). **a)** Beta-lactamase genes detected in *E. coli* samples from Figure 2. The gyrA gene was detected with the CARD protein variant model, which requires a genetic variant conveying resistance in addition to the presence of the gene. **b)** Vancomycin resistance genes of the vanA operon detected in *E. faecium* in samples from Figure 3.

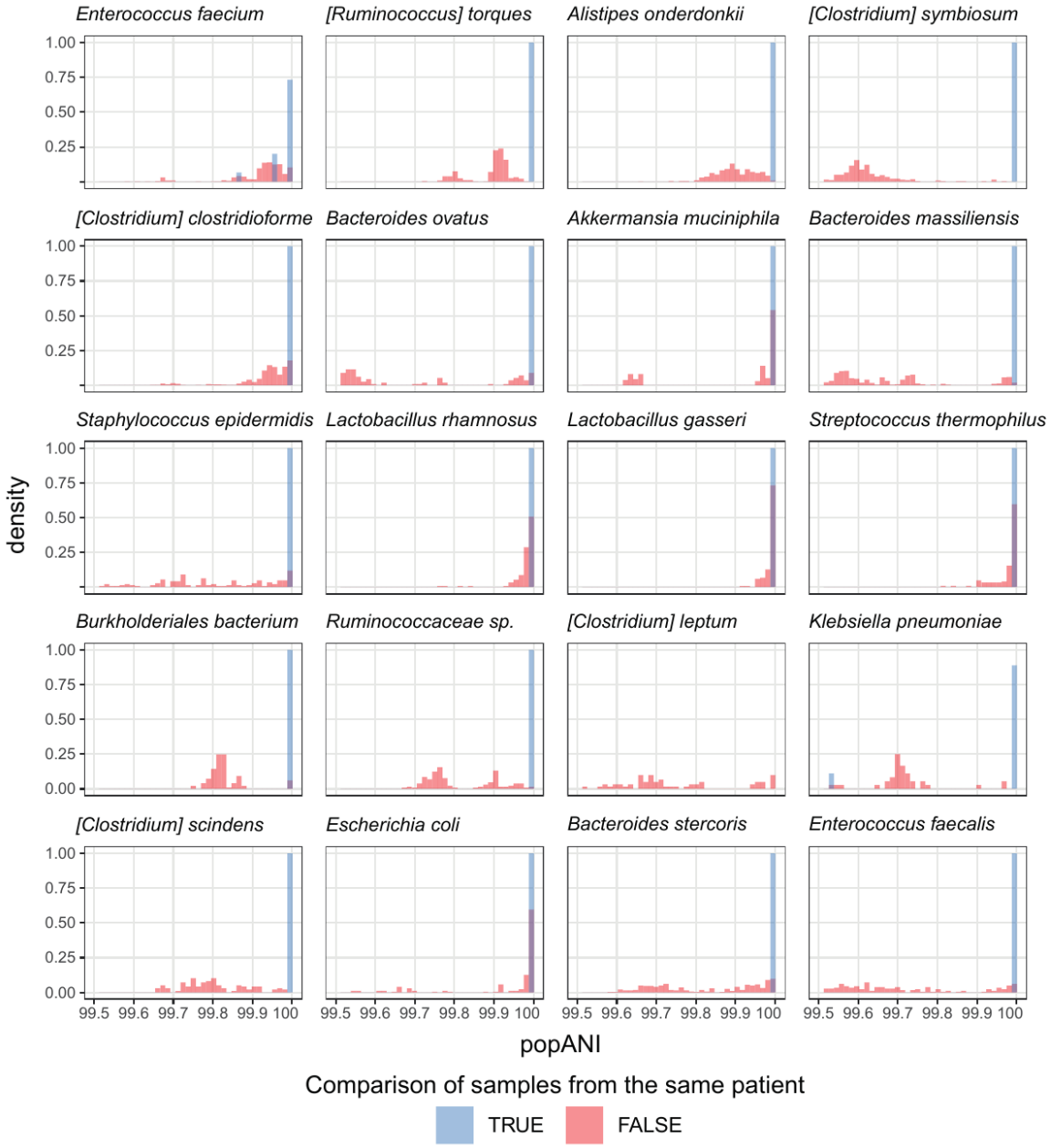


Figure 4.10: **Distribution of popANI values comparing samples from the same or different patients.** Distributions are split by species and the most common 25 species are shown. While in many cases the two distributions overlap, very rarely did popANI values comparing samples from different patients exceed the 99.999% transmission threshold. Comparisons with <99.5% popANI are omitted from the figure for clarity.

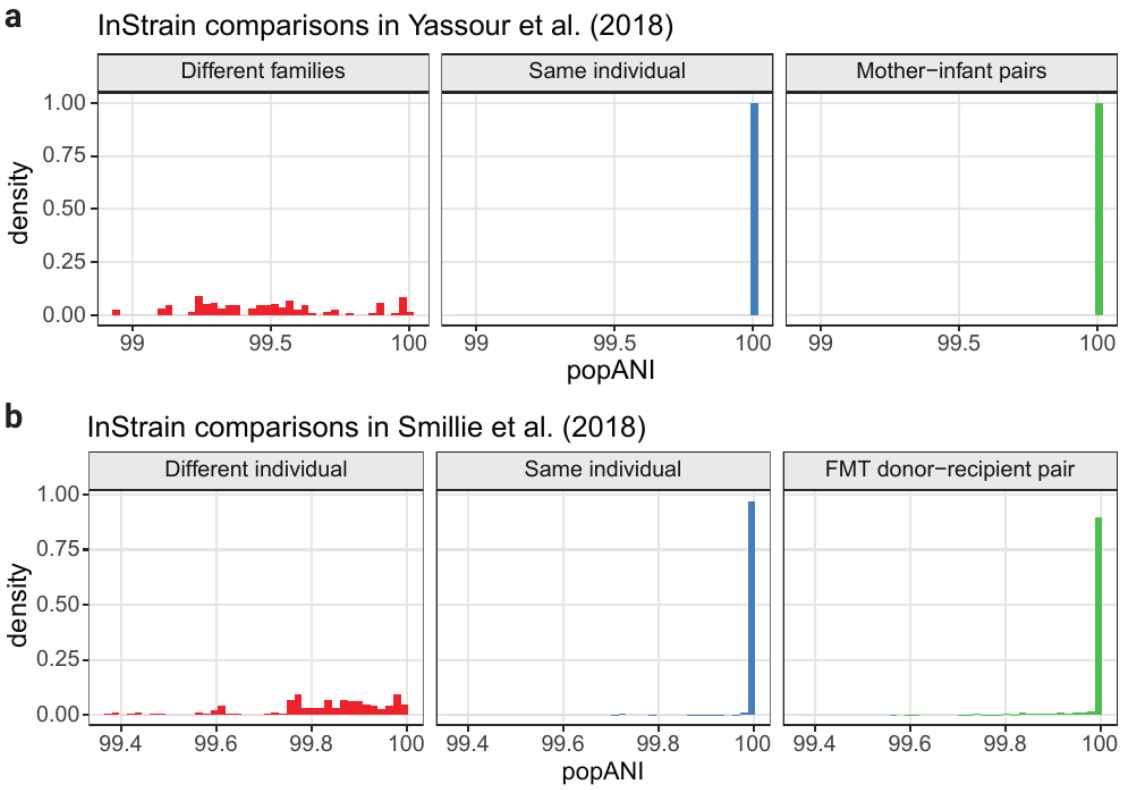


Figure 4.11: **InStrain analysis of the five most common species in external datasets where transmission is expected to occur.** Distributions of popANI values are separated based on the individuals the samples came from, with putative transmission events contained in the far right panel. **a)** Metagenomic sequencing datasets from mother-infant pairs18. The maximum popANI value obtained when comparing samples from different families was 99.995%. **b)** Metagenomic sequencing datasets from fecal microbiota transplantation donors and recipients. The maximum popANI value obtained comparing samples from individuals not related by FMT was 99.998%

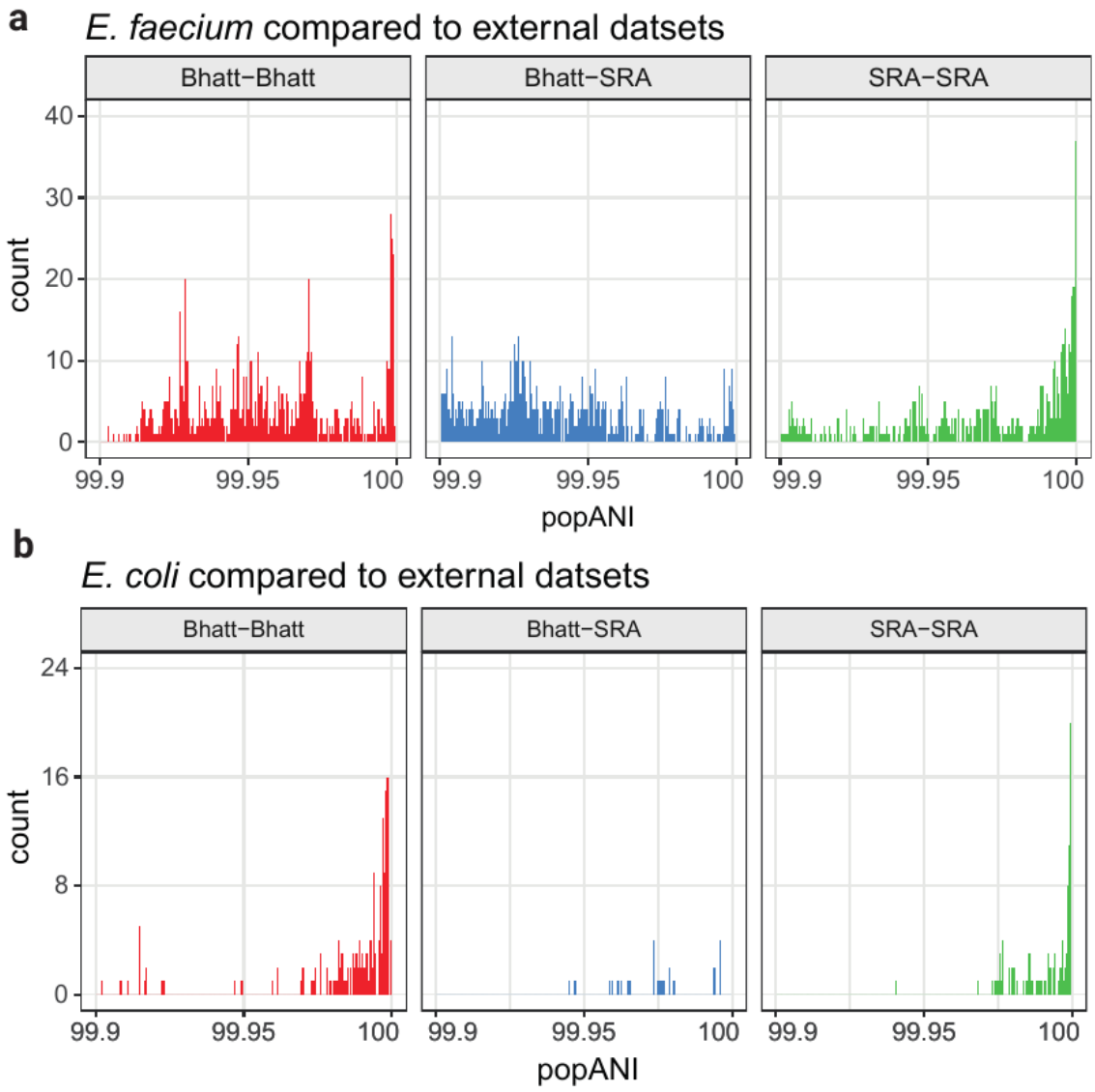


Figure 4.12: ***Enterococcus faecium*** **(a) and** ***Escherichia coli*** **(b) strains compared to external datasets.** Including hospitalized adult and pediatric HCT patients, hospitalized infants and vancomycin-resistant *E. faecium* isolates3,69-73. Panels are separated according to whether comparisons were made within the data in this manuscript (Bhatt-Bhatt), between our data and external data (Bhatt-SRA) or within external data (SRA-SRA).

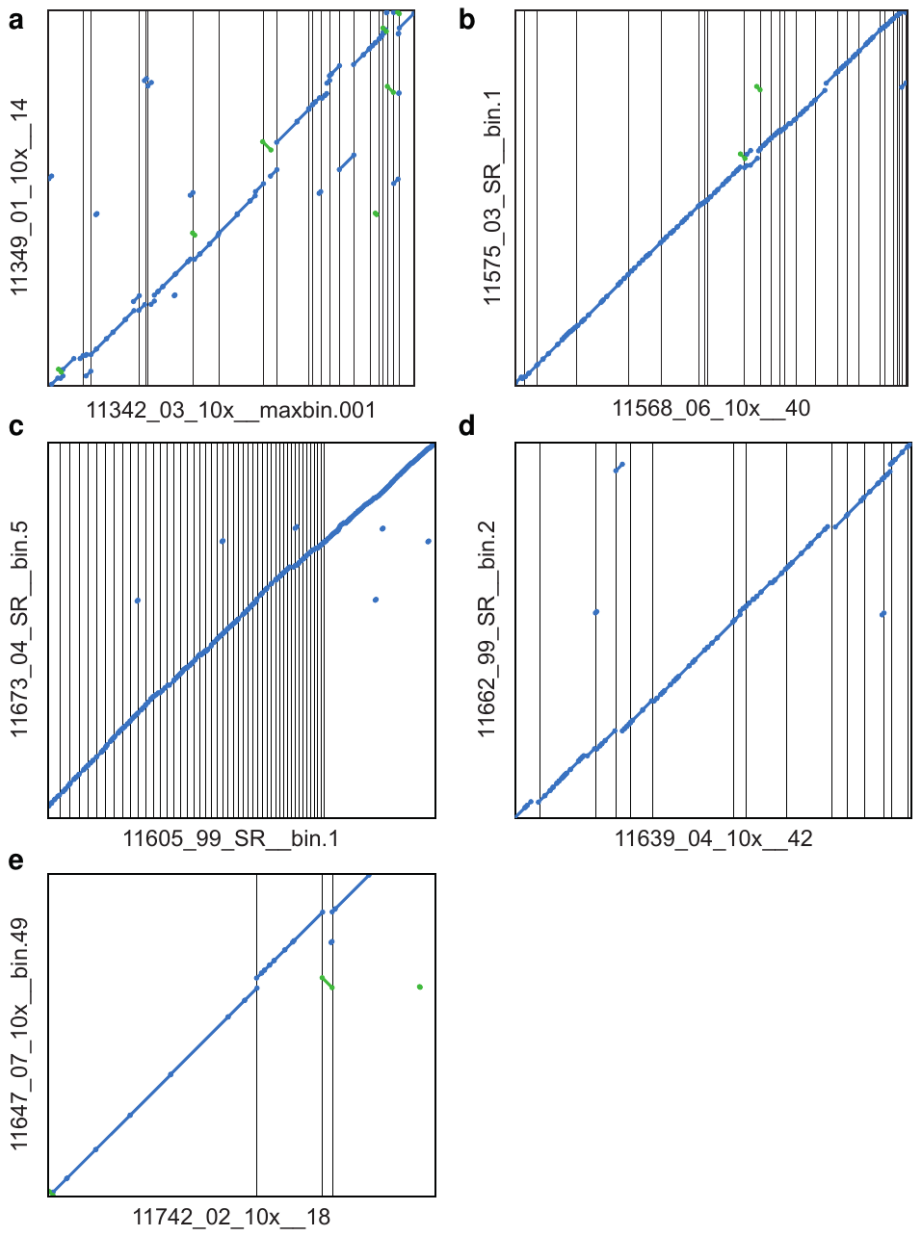


Figure 4.13: **Dotplots showing pairwise alignment of MAGs in cases of putative transmission of the given species.** Blue lines along the diagonal indicate 1-1 homology between the two sequences. Green lines indicate inversions that are likely the result of assembly or binning errors. **a)** *E. faecium* MAGs from patients 11342 and 11349, corresponding to figure 4a. **b)** *E. faecium* MAGs from patients 11575 and 11568, corresponding to figure 4b. **c)** *E. faecium* MAGs from patients 11605 and 11673, corresponding to figure 4c. **d)** *H. hathewayi* MAGs from patients 11639 and 11662, corresponding to figure 5a. **e)** *A. muciniphila* MAGs from patients 11742 and 11647 corresponding to figure 5b.

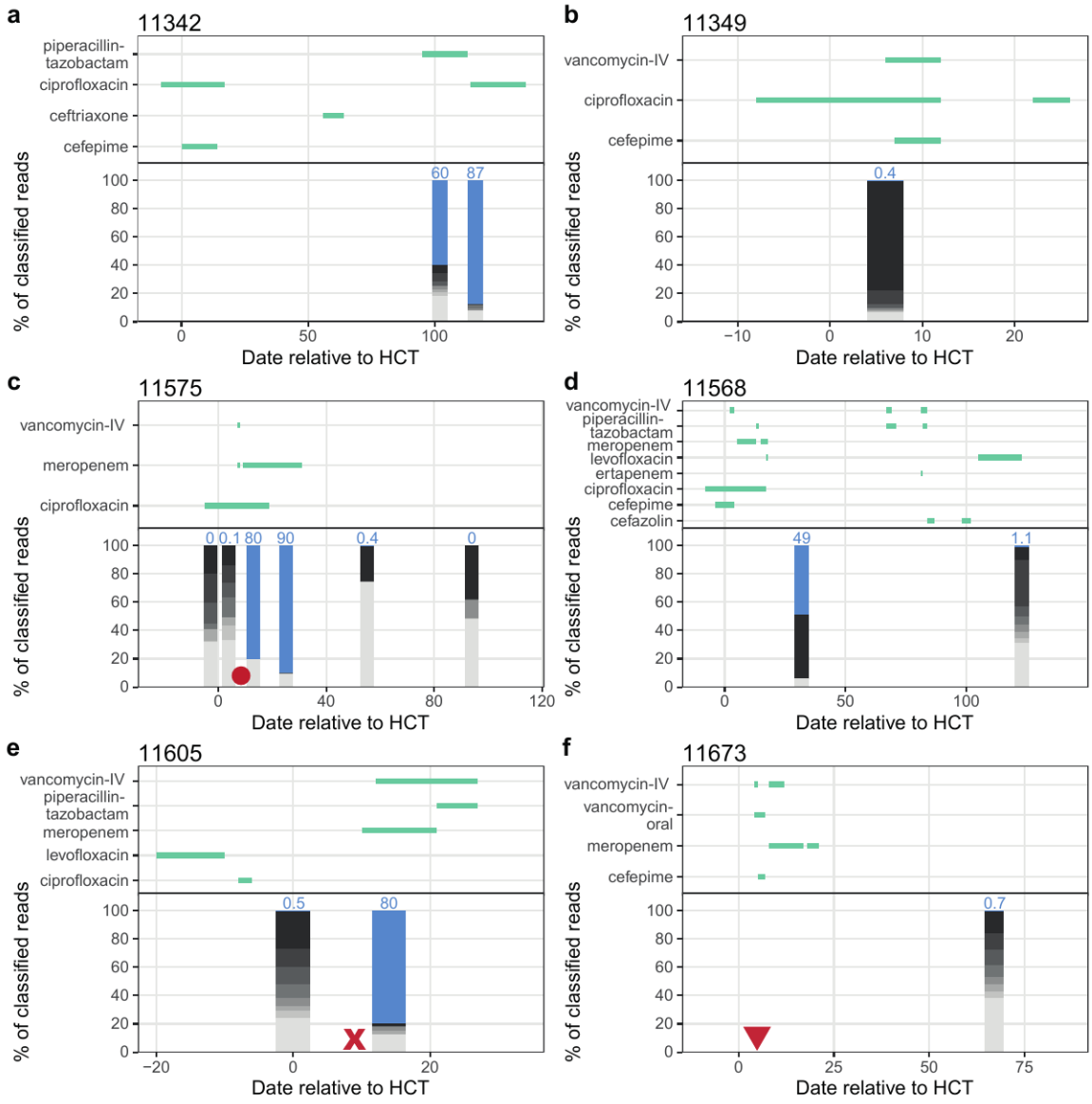


Figure 4.14: **Antibiotic prescription and taxonomic composition of patients with nearly identical *Enterococcus faecium* strains.** *E. faecium* abundance is shown in blue and indicated with text. Other taxa are shown in grey. All dates are relative to HCT for the particular patient. Approximate dates of BSI are shown with red symbols. Circle: *Klebsiella pneumoniae*, X: *Enterococcus faecium*, triangle: *Escherichia coli*.

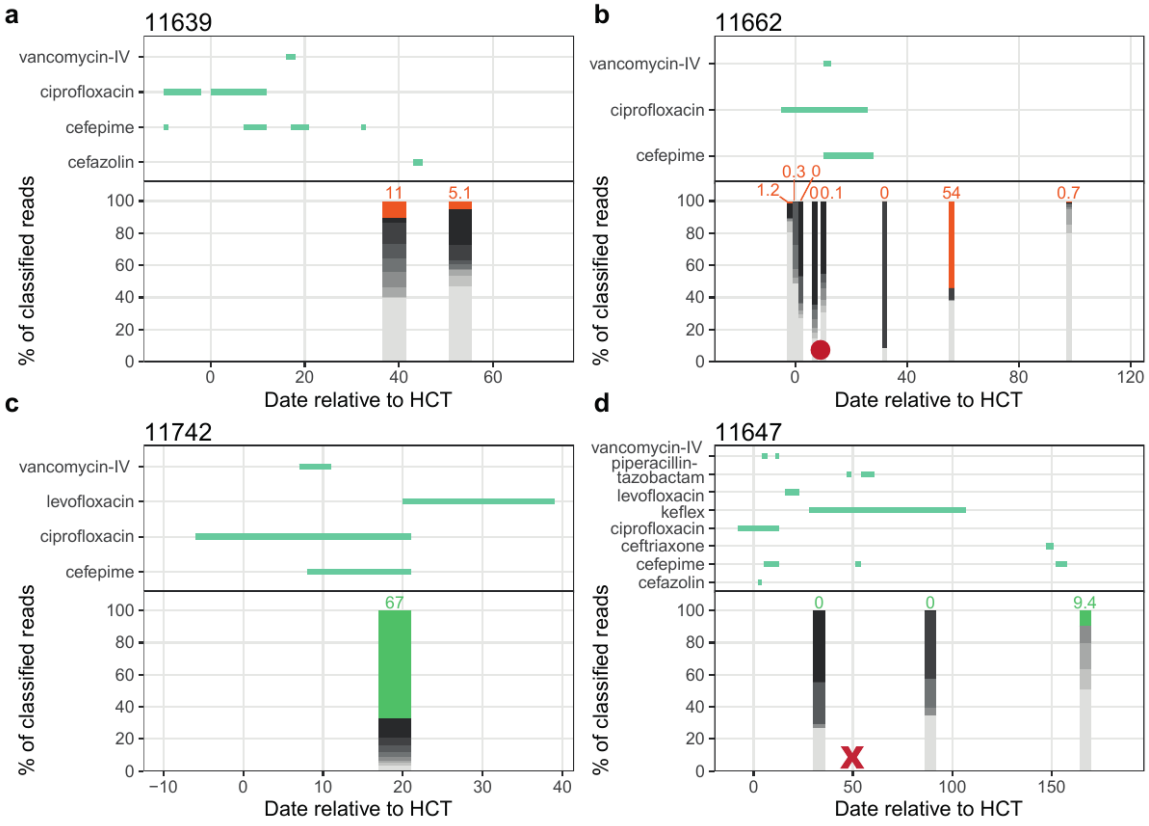


Figure 4.15: **Antibiotic prescription and taxonomic composition of patients with nearly identical *Hungatella hathewayi* or *Akkermansia muciniphila* strains.** Other taxa are shown in grey. All dates are relative to HCT for the particular patient. Approximate dates of BSI are shown with red symbols. Circle: Streptococcus mitis, X: *Klebsiella pneumoniae*.

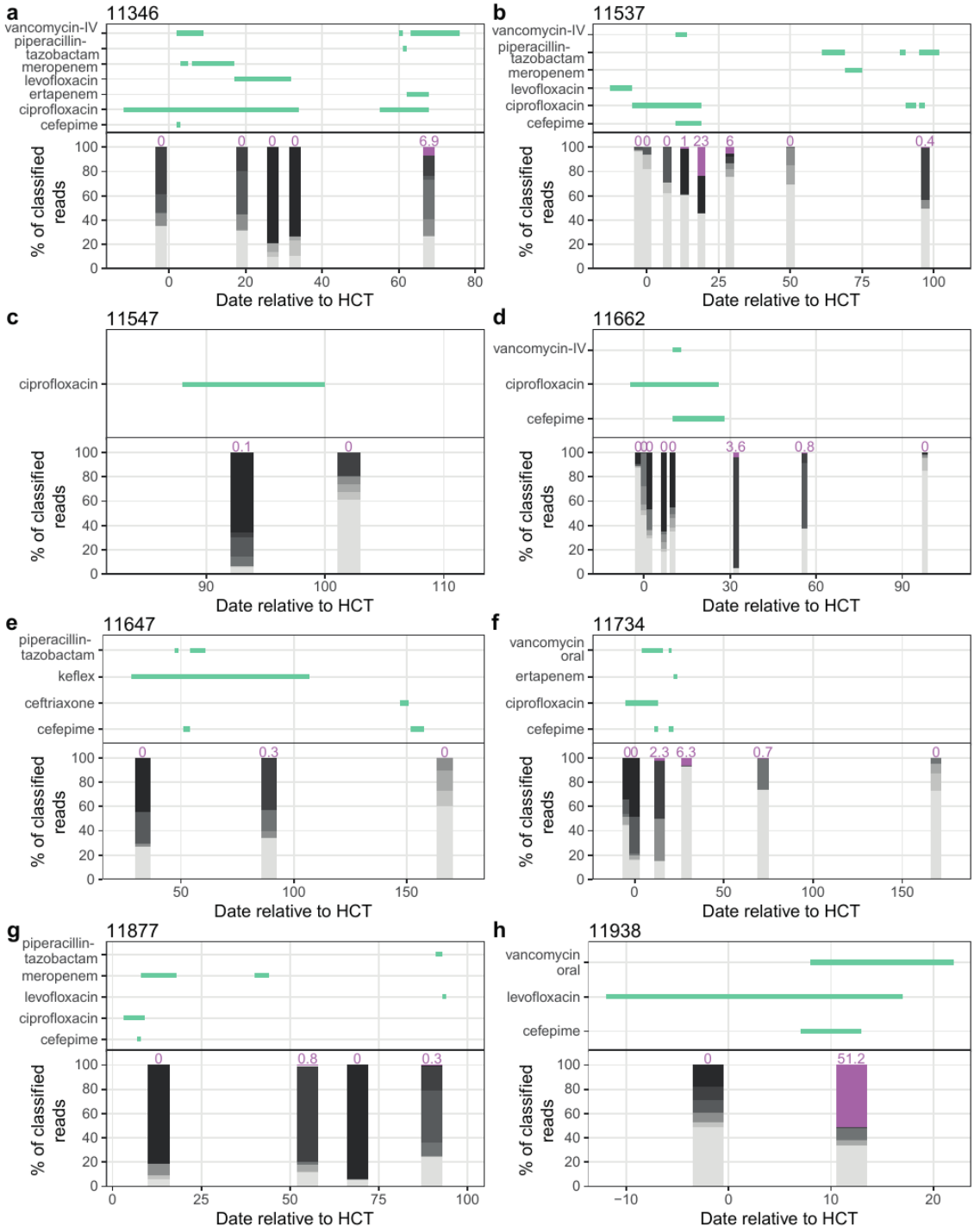


Figure 4.16: **Antibiotic prescription and taxonomic composition of patients with nearly identical *Lactobacillus rhamnosus* strains.** *L. rhamnosus* abundance is shown in purple and indicated with text. Other taxa are shown in grey. All dates are relative to HCT for the particular patient.

4.9 Tables

Attribute	**n**	**%**
Total sequenced patients	149	100%
AGE		
≤ 30	9	6%
31-40	17	11%
41-50	24	16%
51-60	38	26%
61-70	55	37%
≥ 71	6	4%
SEX		
Sex M	87	58%
DIAGNOSIS		
ALL: Acute lymphocytic leukemia	21	14%
AML: Acute myelogenous leukemia	42	28%
CML: Chronic myeloid leukemia	6	4%
HL: Hodgkin lymphoma	4	3%
MDS: Myelodysplastic syndrome	42	28%
NHL: Non-Hodgkin lymphoma	23	15%
OTHER: Other malignancy	11	7%
GRAFT		
Allo	132	89%
Auto	17	11%
GVHD		
Accute GVHD yes	88	59%
Chronic GVHD yes	29	19%
Bloodstream Infection (BSI) Genera, within 0-180 days after HCT		
Any BSI	54	36%
Bacillus	1	1%
Enterobacter	2	1%
Enterococcus	7	5%
Escherichia	8	5%
Gemella	1	1%
Klebsiella	6	4%
Pseudomonas	2	1%
Rothia	4	3%
Staphylococcus	14	9%
Streptococcus	9	6%

Table 4.1: Aggregated characteristics of patients with samples investigated in this study

	all patients	patients with at least one sequenced sample
Number of patients spent ≥24h on ward	923	149
Days spent as inpatient on BMT ward		
Mean	21.9	37.6
Median	18	30.8
SD	18.2	21.8
Min	1	8.7
Max	175.4	137.7
Number of rooms occupied ≥24h		
Mean	2.6	3.5
Median	2	3
SD	1.3	1.7
Min	1	1
Max	9	9
Number of patients overlapped ≥24h		
Mean	55.9	80.8
Median	48	72
SD	30.7	38.6
Min	8	25
Max	246	198
Number of patients roommates ≥24h		
Mean	1.5	2.5
Median	1	2
SD	1.5	2.4
Min	0	0
Max	13	13

Table 4.2: Aggregated statistics of temporal geographic data for all patients on the ward during the study period

Attribute	**n**		
Total stools sequenced	401		
Total sequencing datasets	405		
short read (SR)	312		
linked read (LR)	93		
Sequenced with SR and LR	4		
Samples sequenced per patient	median	range	SD
	2	1 - 13	2.4
Reads after processing (millions)	median	range	SD
SR	7.6	0.01 - 28.8	4.4
LR	104	0.9 - 323.6	40
s Assembly N50 (kilobase-pair)	median	range	SD
SR	17.2	0.7 - 163.6	24.8
LR	147.6	9.5 - 956.3	168.5
Binned genomes	n	%	
SR	2859		
High Quality	103	4%	
Medium Quality	2124	74%	
Low quality	632	22%	
LR	1900		
High Q Bowers	518	27%	
High Q Nayfach	950	50%	
Low quality	432	23%	

Table 4.3: Aggregated statistics of sequencing datasets and metagenome-assembled genomes (MAGs) generated in this study

Chapter 5

Building bioinformatics workflows for scalability and reproducibility

Computational analysis of large-scale genomics experiments is no small feat. With terabytes of sequencing data, thousands of samples and hundreds of measured covariates, the data that goes into today's genomics publications is increasingly complicated. However, modern tools for workflow management, metadata tracking and distributed computing make these tasks much easier. With a little foresight and planning ahead of time, these tools enable computational scientists to spend more time on interesting analysis and less time on data processing. Here, I'll share a few principles I've learned from both conducting large bioinformatics and computational biology experiments and developing data processing pipelines for the lab and the research community as a whole.

5.1 Use a workflow management system

The easiest way to create a scalable and reproducible data analysis pipeline is to organize all code under a workflow management system. Modern workflow managers include Snakemake, Nextflow, Workflow Description Language (WDL) and others [191]. These tools provide a way to define steps of a bioinformatics analysis and can integrate scripts or tools written in many different languages. The four main benefits of a workflow manager, as defined in [191], are data provenance, portability, scalability and re-entrancy (the ability to resume a workflow without re-executing completed steps).

I used Snakemake and Nextflow during my thesis research, so my comments will be limited to those two. I found Snakemake to be easier to learn, teach to other students and rapidly prototype workflows. Snakemake pipelines are written in python, so many people are immediately familiar with the syntax. Under the hood, Snakemake models a workflow as a directed acyclic graph, where files are nodes and steps to convert input files to output files are edges. In addition to being easier

to understand, this model allows for pre-computing the total number of steps to be conducted in a workflow and easier re-entry into workflows.

Nextflow may have a higher barrier to entry but is more extensible. Nextflow pipelines are written in groovy and use a more complicated "channel and process" model for defining workflow steps. While not as easy to learn or rapidly prototype workflows, Nextflow is quicker to deploy on cloud computing architecture and has a much larger set of community-developed workflows for common bioinformatics tasks.

5.2 Use established and validated pipelines when possible

Bioinformatics is no different than other fields of science, you should "stand on the shoulders of giants." Unless you're developing a new method or novel analysis, someone has probably already done the computational task you're thinking about. For example, the nf-core community for Nextflow [47] has a variety of pipelines that are well-developed, tested and maintained. An equivalent community exists for Snakemake (`https://github.com/snakemake-workflows`) but does not offer as many workflows. I do take pride in the workflows I write, and want to "own" as many of them as possible. However, using established pipelines, like nf-core, has likely saved me hundreds of hours of work and allows me to focus on higher-impact parts of my science. Plus, you can always fork the github repository and add features or make the pipeline your own.

In addition to these established communities, many laboratories have github pages with established workflows that have been validated through use in publications. For Example, our bhattlab_workflwows [161] and kraken2_classification [160] are used within the lab and by collaborators around the world. While using established pipelines can certainly speed up development of a bioinformatics workflow, it's important to keep in mind that developers can fall behind on updating software, taking advantage of new features or incorporating new tools. Most developers are underpaid grad students or unpaid volunteers, and using these pipelines should come with a warning "some assembly may be required."

5.3 Treat metadata as a first-class citizen

Metadata is just as, if not more important, than the sequencing data from an experiment. An error in the metadata for an experiment can be catastrophic. Luckily there are several tools to help solve metadata issues. For example, databases like project RedCap [63] can be used to collect survey data from users, as well as maintain databases of clinical metadata, sample locations, and similar covariates. Compared to traditional programmatic databases, a RedCap database is easier to set up and maintain for scientists without a strong computer science background.

To keep metadata and data harmonized, I rely on a package *CMapR* and the *gctx* file format

developed by my previous group at the Broad Institute [46]. A *gctx* file is a two dimensional data matrix that always has associated row and column metadata. Convenient subsetting and filtering operations also subset the associated metadata. That way, metadata is never the wrong size or out of date. As a practical example, I use a *gctx* object for read count and relative abundance matrices when doing microbiome classification with *Kraken2* [189]. The row metadata contains the NCBI taxonomic identifier and the complete classification lineage, while the column metadata contains sample names and any associated group or clinical data.

5.4 Leverage high-performance computing or cloud infrastructure

The compute, memory and storage requirements of large-scale genomics experiments typically exceed that of laptop or desktop workstations. Most computational labs are familiar with university-provided supercomputing services, like SCG or Sherlock at Stanford. These services typically charge a membership fee plus a per-cpu-hour compute charge, and are convenient to access with included support services. However, supercomputing infrastructure may not scale large enough for the needs of a particular project or may be delayed because resources are shared across multiple labs. In these cases, cloud computing offers additional benefits.

Cloud computing platforms, like Amazon Web Services (AWS) or Google Cloud Platform (GCP) offer inexpensive access to compute that can scale theoretically infinitely. Users only pay for the compute, storage and bandwidth that they use. With minimal effort, users can spin up a virtual machine with up to 96 cores for a low hourly cost. With a more complex setup, a workflow manager can automatically initialize and deploy compute jobs for each step in a bioinformatics workflow. Note that these systems are more complicated than a traditional managed supercomputer environment, where it's likely there are several university employees managing resources and data access. It's likely best to have one lab member with more of a computer science or software engineering background "in charge" of these resources for the rest of the lab. It's also important to keep a close eye on costs. While compute may be cheap, data storage and download is relatively expensive and can easily rack up charges with the user being unaware. Most cloud computing providers offer grants for academic research that are easy to obtain.

5.5 Best practices for developing bioinformatics workflows

While developing bioinformatics pipelines for our lab and collaborators, I've settled upon a few key rules.

- Workflows need to be easy to install and use. This is perhaps the most important requirement - more so than additional functionality or publication in a journal. Some of the most widely

used and reliable bioinformatics software remains unpublished, including snippy for bacterial SNP calling and core genome alignment (`https://github.com/tseemann/snippy`)

- Package software with conda and docker containers so users can easily install the software, or use the container with no installation required.
- Maintenance and responding to bugs/issues is a challenging and thankless task, especially when students leave the lab. Maintaining software should remain a priority, and the lab should set an expectation that, for example, students will maintain software for two years after publication, even if they graduate before that date.
- Provide a thorough readme and simple test dataset to ensure installation and usage works as expected.
- Outputs, both data table and figures, should be easily interpretable to the users. More detailed and verbose output should be made available if needed.

The best examples of microbiome-related software that I've encountered so far continue to be the Biobakery suite of tools from the Huttenhower lab [16] and Matt Olm's tool *dRep* [121]. These tools are widely used - you frequently see references to them in talks and publications and most people in the field know the core idea behind the tools.

Chapter 6

Future directions and conclusions

6.1 Future directions of the mother-infant crAss-like phage transmission work

While a few future directions were listed in the main text of this work, I would like to expand on some areas that have continued to interest me in the time since publication.

6.1.1 Extension to other phages

CrAss-like phages remained undiscovered for years because researchers didn't deeply investigate unclassified metagenomic sequencing reads. In my mother-infant work, I only looked at phages that had been previously discovered (crAss-like phages and other phages present in NCBI Genbank). This choice allowed me to conduct a focused analysis and eliminated the need for methods to discover novel phages, which were not very mature at the time.

More recent tools and databases of phage sequences have changed this paradigm. For example, virFinder [138], VirSorter [142], and VIBRANT [77] are all methods to identify likely phage contigs in assembled metagenomic sequencing data. Recent publications with large databases of novel phage families [27, 150] have added to the collection of phage genomes available for comparison. There are likely to be other phages that are similarly transmitted from mothers to infants; using these tools in a revised analysis should uncover new and interesting patterns of transmission.

6.1.2 Discovering bacterial hosts of novel phages

The setting of mother-infant or Fecal Microbiota Transplantation (FMT) donor-recipient microbiome transmission is also appropriate to discover the host of crAss-like and other phages. If a phage is transmitted, engrafts and persists in the recipient's microbiome, the bacterial host must also be

present. The host could be present in the recipient's microbiome prior, or it could be transmitted alongside the phage. A significant association between the presence or abundance of a phage and the host should emerge in a large enough sample collection. I attempted to identify the elusive host for crAss-like phages by comparing bacterial species abundance in crAss-like phage positive and negative infant samples. However, the limited sample numbers in the mother-infant datasets I examined precluded any significant associations. An investigation that quantifies bacterial and phage transmission together in a larger sample collection may be able to uncover signals of phage-host relationships.

6.1.3 Quantifying bacterial strain diversity upon transmission

I quantified the strain diversity of crAss-like pages upon transmission and demonstrated that the phage population experiences a bottleneck and decrease of strain diversity. However, the original authors of mother-infant transmission datasets [195, 13] did not examine strain diversity at the single nucleotide level. Now that we have the ability to generate complete bacterial metagenome-assembled genomes (MAGs) and examine variants and nucleotide diversity at the strain level, it would be interesting to repeat the same experiments for the transmitted bacterial strains. I expect the most abundant strain within a species to be the most frequently transmitted from mothers to infants, with a similar decrease in diversity upon transmission.

6.2 Future directions of the HCT patient transmission work

6.2.1 Validation, replication and extension of this work

While I found evidence for transmission of commensal and pathogenic bacteria between the gut microbiomes of adults recovering from HCT, the sample set I investigated was not large enough to quantify the rate of transmission of different organisms. The results from my work need be replicated in a larger cohort, potentially at a different hospital. The same hypotheses about transmission could also be tested in a pediatric cohort, because more transmission may be expected in younger patients. Although mother-infant transmission is most common in the first few years of life, the gut microbiome continues to develop and acquire new strains into adolescence. Alternatively, transmission between the patient and family members could be investigated in the setting of allogeneic transplants at home [62].

The bacterial strains I hypothesized to be transmitted between individuals may persist in the hospital reservoir. Performing environmental sampling of surfaces in the patient's room, including beds, sinks and toilets, could help identify reservoirs for the potentially transmitted strains. Additionally, swabbing the hands and analyzing stool samples from healthcare workers and hospital visitors could help determine if these individuals were sources for the new strains colonizing patient

microbiomes. I only analyzed transmission of bacterial species in my work, although I did test for transmission of crAss-like phages with null findings. Expanding the results of this work to phages and fungi may reveal other interesting trends in this cohort.

6.2.2 HCT patients often acquire new bacterial strains. Where do they come from?

I frequently observed that HCT patient microbiomes return to a similar state after perturbation with antibiotics. For example, some bacterial species are present upon hospital admission, become undetectable during periods of antibiotic use, and reappear later when the microbiome recovers diversity. Often the strains in early and late samples have significantly different genomes, indicating that a new strain may have colonized the patient's microbiome. It's also possible that a strain at very low abundance initially survived and proliferated after antibiotic use ceased. Given the high depth of sequencing in many of these patient samples, I believe that strain persistence is insufficient to explain all cases of species re-emergence.

Species re-emergence often occurs in bacterial species that are not frequently detected within HCT patient microbiomes. For example, *Hungatella hathewayi* is present in less than 10% of patients, but appears both in pre and post HCT samples from patient 11662 (Figure 4.6), after becoming undetectable in the middle four samples. Comparing early and late genomes in this case revealed ¡95% ANI, below the 95% threshold commonly used to define different species. Earlier presence of a species may prime the microbiome to return to a state where the species is present. This "priority effect" [52, 163, 165] from prior colonization may assist the species to re-colonize the gut in a more favorable scenario, given the proper exposure. I don't believe there's enough patients with dense time course sampling in the data generated from our study to fully investigate this question. Samples from our study could be combed with the HCT patient microbiome data from Memorial Sloan Kettering Cancer Center [194], which contains more dense sampling of fewer patients. Future research may shed light on how priority effects influence the microbiome following HCT, antibiotic treatment and community recovery.

6.2.3 Are roommates at risk for colonization with pathogenic bacteria?

I initially hypothesized that patients who were roommates in the hospital would be more likely to transmit pathogenic strains. However, I found relatively limited transmission of bacteria between patients who were roommates, including a single case involving *Enterococcus faecium* and two cases involving commensal bacteria. I did not identify any bacterial strains shared between more than two patients, indicating that there are not common, hospital-acquired bacterial strains colonizing multiple patients. Finally, I did not find an association between patients who were roommates and the incidence of bloodstream infections. Therefore, I do not believe exposure to a roommate

following HCT is a significant risk factor for colonization with pathogenic bacteria or for acquiring a bloodstream infection. Roommate exposure may still be a consideration for communicable illnesses, such as COVID-19.

It was recently shown that autologous FMT following HCT can improve microbiome diversity and reduce the incidence of graft versus host disease (GVHD) [36, 172]. FMT is a crude therapy where the entire microbiome of a donor is given to the host. Doctors are not able to control which microbes are present in a donor sample, nor which microbes engraft and persist in the recipient's microbiome. If an uncontrolled and drastic therapy like FMT is beneficial to HCT patients, perhaps more gentle exposure to new strains via other individuals in the room could also prove to be beneficial. However, the best person to have in a room may be the patient's family member or spouse at home [62], not another HCT patient with an antibiotic-perturbed microbiome potentially filled with pathogenic strains.

6.2.4 Future strain-specific investigations

The HCT patient microbiome is an excellent model to study strain-specific effects, as natural experiments involving selection and succession are conducted on a daily basis in each patient. My investigation of transmission only scratched the surface of the potential strain-specific research involving these patients. Examining strain diversity during periods of antibiotic treatment or microbiome recovery may reveal new principles about community assembly in the human gut [187, 184]. The deep linked read metagenomic sequencing could also be used to phase strain variants [141], which would help identify the relative proportions and functional capabilities of each bacterial strain. Finally, I am interested in testing why certain bacterial strains or species engraft into a recipient, while others do not. However, separating the effects of transmission and engraftment will remain challenging.

6.2.5 Managing the microbiome in the clinic

While there were no direct clinical outcomes from this observational study, the conclusions do highlight some relevant points for clinicians. Building on previous work showing that HCT patients can acquire infections from their own microbiome [168, 76], this research raises the possibility that the infectious agent can be sourced from another patient in the hospital. In cases where transmission of bacterial strains was likely, the recipient patient had a microbiome that was extremely perturbed by antibiotic exposure. Therefore, reducing antibiotic treatment whenever possible may further reduce the risk of microbiome transmission. In addition, clinicians may desire to increase gut microbiome diversity, and therefore colonization resistance, by one of the many methods available: probiotics [29], prebiotics [8] and FMT [36, 172]. However, clinical trials are necessary to demonstrate the safety and efficacy of these methods.

6.2.6 Are MAGs derived from the samples of interest necessary?

I put considerable effort into generating high-quality de-novo MAGs from HCT patient samples, using deep linked-read sequencing sequencing and extensive computational assembly pipelines. However, throughout the course of the computational experiments, I started to wonder if de-novo MAGs were truly necessary. There are now large databases of microbial genomes that should cover much of the diversity present in a human microbiome, such as The Genome Taxonomy Database (GTDB) [130] and the Unified Human Gastrointestinal Genome (UHGG) [5]. This is especially true for microbiome samples from western individuals and HCT patients, which often lack microbial diversity. An alternative workflow to the de-novo MAG workflow used in the HCT transmission work follows:

1. Sequence all samples with short-read Illumina sequencing (2x150bp)
2. Map reads from all samples against GTDB
3. Select reference strains with sufficient coverage in samples from multiple patients
4. Analyze SNPs with inStrain

Eliminating the need for de-novo MAG generation will also decreases the required sequencing depth. A few quality checks can ensure the diversity in patient samples is well-represented by the MAG database, including:

- Examine the proportion of reads from each sample that mapped stringently to the MAG database. Samples with a low mapping fraction may have diverse strains that are underrepresented in the database.
- Re-sequencing with a long-read technology and MAG assembly might be required for samples that are poorly represented, but this effort could be targeted where it's most needed.

As a sanity check in the HCT patient transmission work, I replicated all of the putative transmission findings using a public reference genome rather than the sample-derived MAG. Therefore, I believe the same results could have been obtained using a MAG database instead of of the sample-derived MAGs. Alternatively, MAGs could still be created from patient samples, and a database combining GTDB and patient-derived MAGs could be used for short read mapping. This hybrid approach would fill in areas missing in one of the genome collections.

6.3 Conclusions

Throughout my PhD research, I studied transmission of human gut microbiota under two extremes: infants as the microbiome developed, and HCT patients as the microbiome recovered from antibiotic treatment, chemotherapy and radiation. These two settings are some of the most likely to result in

microbiome transmission, outside of the drastic therapy of FMT. First, I characterized transmission of crAss-like phages from the microbiome of a mother to the microbiome of her infant, discovering that transmission was common during the first year of life. Tracking strain-specific variants in the phage population revealed that strain diversity was reduced upon mother-infant transmission, as would be expected given a population bottleneck. This work highlighted that mother-infant transmission of crAss-like phages is likely to be a part of normal infant development, and posed the question that crAss-like phages may have been transmitted along the maternal line for millennia, much like mitochondrial DNA.

In adult patients recovering from HCT, I found that microbiome transmission of bacteria is likely to be rare and limited to cases where patients are roommates and have microbiomes recovering from antibiotic exposure. In the case of *Enterococcus faecium*, individuals who were not roommates appeared to share nearly identical strains, indicating that these strains may be acquired through unsampled intermediates or from the hospital environment. Transmission of the commensal microbes *Hungatella hathewayi* and *Akkermansia muciniphila* was limited to cases where the two patients were roommates and the recipient patient was recovering from antibiotic treatment initiated for a bloodstream infection. The time course sampling, deep metagenomic sequencing, and careful analysis in these cases gives me confidence that patient-patient transmission is a true phenomenon. Overall, these results point to the resilience of the adult microbiome. Even under the extreme pressures from HCT and antibiotic treatment, microbiome transmission remained rare, and the patients did not appear to acquire common pathogenic strains into their gut.

The developing infant and HCT patient microbiome are excellent settings to study strain specific effects, as natural experiments of colonization and succession are conducted daily on the microbial communities. I hope that future strain specific microbiome research teaches us more about the overarching ecological principles shaping the microbial communities living on and within us.

Bibliography

[1] J. Abranches. "Biology of Oral Streptococci." In: *Microbiol. Spectr.* (2018).

[2] Varun Aggarwala et al. "Precise quantification of bacterial strains after fecal microbiota transplantation delineates long-term engraftment and explains outcomes". In: *Nature Microbiology* 6.10 (Oct. 2021), pp. 1309–1318. ISSN: 2058-5276. DOI: 10.1038/s41564-021-00966-0. URL: http://www.nature.com/articles/s41564-021-00966-0 (visited on 10/03/2021).

[3] M. O. Ahmed. "Vancomycin-Resistant Enterococci: A Review of Antimicrobial Resistance Mechanisms and Perspectives of Human and Animal Health." In: *Microb. Drug Resist.* (2017).

[4] B. P. Alcock. "CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database." In: *Nucleic Acids Res.* (2019).

[5] A. Almeida. "A unified catalog of 204,938 reference genomes from the human gut microbiome." In: *Nat. Biotechnol* (2020).

[6] J. Alneberg et al. "Binning metagenomic contigs by coverage and composition". In: *Nat Methods* 11 (2014). DOI: 10.1038/nmeth.3103. URL: https://doi.org/10.1038/nmeth.3103.

[7] T. M. Andermann. "The Microbiome and Hematopoietic Cell Transplantation: Past, Present, and Future. Biol." In: *Blood Marrow Transplant.* (2018). DOI: doi:10.1038/nbt.4266.

[8] Tessa M. Andermann et al. "A Fructo-Oligosaccharide Prebiotic Is Well Tolerated in Adults Undergoing Allogeneic Hematopoietic Stem Cell Transplantation: A Phase I Dose-Escalation Trial". In: *Transplantation and Cellular Therapy* (July 16, 2021). ISSN: 2666-6367. DOI: 10.1016/j.jtct.2021.07.009. URL: https://www.sciencedirect.com/science/article/pii/S2666636721010745 (visited on 10/07/2021).

[9] "Andrews S. Fastqc a quality control tool for high throughput sequence data. http://www.bioinformatics.babra
Accessed 10 Nov 2017." In: (). URL: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[10] "Asnicar, F. et al. Studying vertical microbiome transmission from mothers to infants by strain-level metagenomic profiling. mSystems2, e00164–16 (2017)." In: ().

[11] K. Atarashi. "T reg induction by a rationally selected mixture of Clostridia strains from the human microbiota." In: *Nature* (2013).

[12] "Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data." In: (). URL: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[13] F. Bäckhed. "Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life." In: *Cell Host Microbe* (2015).

[14] J. Barroso-Batista. "The First Steps of Adaptation of Escherichia coli to the Gut Are Dominated by Soft Sweeps." In: *PLOS Genet.* (2014).

[15] Erez N. Baruch et al. "Fecal microbiota transplant promotes response in immunotherapy-refractory melanoma patients". In: *Science* 371.6529 (Feb. 5, 2021), pp. 602–609. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.abb5920. URL: http://science.sciencemag.org/content/371/6529/602 (visited on 02/08/2021).

[16] Francesco Beghini et al. "Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3". In: *bioRxiv* (Nov. 21, 2020), p. 2020.11.19.388223. DOI: 10.1101/2020.11.19.388223. URL: https://www.biorxiv.org/content/10.1101/2020.11.19.388223v1 (visited on 06/13/2021).

[17] S. Benler. "A diversity-generating retroelement encoded by a globally ubiquitous Bacteroides phage". In: *Microbiome* 6 (2018). DOI: 10.1186/s40168-018-0573-6. URL: https://doi.org/10.1186/s40168-018-0573-6.

[18] D. Bertrand. "Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes." In: *Nat. Biotechnol.* (2019). DOI: doi:10.17504/protocols.io.n7hdhj6.

[19] A. Bishara et al. "High-quality genome sequences of uncultured microbes by assembly of read clouds". In: *Nat Biotechnol* 36 (2018). DOI: 10.1038/nbt.4266. URL: https://doi.org/10.1038/nbt.4266.

[20] R. M. Bowers et al. "Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea". In: *Nat Biotechnol* 35 (2017). DOI: 10.1038/nbt.3893. URL: https://doi.org/10.1038/nbt.3893.

[21] I. L. Brito. "Transmission of human-associated microbiota along family and social networks." In: *Nat. Microbiol.* (2019).

[22] B. Brooks. "Strain-resolved analysis of hospital rooms and infants reveals overlap between the human and room microbiome." In: *Nat. Commun.* (2017). DOI: doi:10.1038/s41587-020-0422-6.

[23] "Brown, B. P. et al. crAssphage abundance and genomic selective pressure correlate with altered bacterial abundance in the fecal microbiota of South African mother-infant dyads. https://doi.org/10.1101/582015." In: ().

[24] H. P. Browne. "Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation." In: *Nature* (2016).

[25] B. Buchfink, C. Xie, and D. H. Huson. "Fast and sensitive protein alignment using DIAMOND". In: *Nat Methods* 12 (2015). DOI: 10.1038/nmeth.3176. URL: https://doi.org/10.1038/nmeth.3176.

[26] K. Cadwell. "Virus-plus-susceptibility gene interaction determines Crohn's disease gene Atg16L1 phenotypes in intestine". In: *Cell* 141 (2010). DOI: 10.1016/j.cell.2010.05.009. URL: https://doi.org/10.1016/j.cell.2010.05.009.

[27] Luis F. Camarillo-Guerrero et al. "Massive expansion of human gut bacteriophage diversity". In: *Cell* 184.4 (Feb. 18, 2021), 1098–1109.e9. ISSN: 0092-8674. DOI: 10.1016/j.cell.2021.01.029. URL: https://www.sciencedirect.com/science/article/pii/S0092867421000726 (visited on 10/06/2021).

[28] M. Cervantes-Echeverría. "Whole-genome of Mexican-crAssphage isolated from the human gut microbiome". In: *BMC Res. Notes* 11 (2018). DOI: 10.1186/s13104-018-4010-5. URL: https://doi.org/10.1186/s13104-018-4010-5.

[29] Children's Oncology Group. *The Effectiveness of Lactobacillus Plantarum (LBP, IND# 17339) in Preventing Acute Graft-Versus-Host Disease (GvHD) in Children Undergoing Alternative Hematopoietic Progenitor Cell Transplantation (HCT)*. Clinical trial registration NCT03057054. clinicaltrials.gov, Aug. 23, 2021. URL: https://clinicaltrials.gov/ct2/show/NCT03057054 (visited on 10/06/2021).

[30] O. Cinek. "Quantitative CrAssphage real-time PCR assay derived from data of multiple geographically distant populations". In: *J. Med. Virol.* 90 (2018). DOI: 10.1002/jmv.25012. URL: https://doi.org/10.1002/jmv.25012.

[31] P. Cingolani. "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff". In: *Fly* 6 (2012). DOI: 10.4161/fly.19695. URL: https://doi.org/10.4161/fly.19695.

[32] F. D'Amico. "Gut resistome plasticity in pediatric patients undergoing hematopoietic stem cell transplantation." In: *Sci. Rep.* (2019).

[33] C. E. Dandoy. "Bacterial bloodstream infections in the allogeneic hematopoietic cell transplant patient: new considerations for a persistent nemesis." In: *Bone Marrow Transplant.* (2017). DOI: doi:10.1038/s41587-019-0191-2.

[34] P. Danecek. "The variant call format and VCFtools". In: *Bioinformatics* 27 (2011). DOI: 10.1093/bioinformatics/btr330. URL: https://doi.org/10.1093/bioinformatics/btr330.

[35] Diwakar Davar et al. "Fecal microbiota transplant overcomes resistance to anti–PD-1 therapy in melanoma patients". In: *Science* 371.6529 (Feb. 5, 2021), pp. 595–602. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.abf3363. URL: http://science.sciencemag.org/content/371/6529/595 (visited on 02/05/2021).

[36] Zachariah DeFilipp et al. "Third-party fecal microbiota transplantation following allo-HCT reconstitutes microbiome diversity". In: *Blood Advances* 2.7 (Apr. 10, 2018), pp. 745–753. ISSN: 2473-9529, 2473-9537. DOI: 10.1182/bloodadvances.2018017731. URL: http://www.bloodadvances.org/content/2/7/745 (visited on 01/29/2019).

[37] A. L. Delcher et al. "Alignment of whole genomes". In: *Nucleic Acids Res* 27 (1999). DOI: 10.1093/nar/27.11.2369. URL: https://doi.org/10.1093/nar/27.11.2369.

[38] Centers for Disease Control and Prevention (U.S.) "Diseases and Organisms in Healthcare Setting". In: (2019). DOI: doi:10.1038/s41587-020-00797-0. URL: https://www.cdc.gov/hai/organisms/organisms.html.

[39] Y. Doi et al. "Community-associated extended-spectrum -lactamase-producing Escherichia coli infection in the United States". In: *Clin Infect Dis* 56 (2013). DOI: 10.1093/cid/cis942. URL: https://doi.org/10.1093/cid/cis942.

[40] M. G. Dominguez-Bello. "Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns". In: *Proc. Natl Acad. Sci. USA.* 107 (2010). DOI: 10.1073/pnas.1002601107. URL: https://doi.org/10.1073/pnas.1002601107.

[41] L. A. Draper. "Long-term colonisation with donor bacteriophages following successful faecal microbial transplantation". In: *Microbiome* 6 (2018). DOI: 10.1186/s40168-018-0598-x. URL: https://doi.org/10.1186/s40168-018-0598-x.

[42] K. Dubin. "Enterococci and Their Interactions with the Intestinal Microbiome." In: *Bugs Drugs* (2018).

[43] K. A. Dubin. "Diversification and Evolution of Vancomycin-Resistant Enterococcus faecium during Intestinal Domination." In: *Infect. Immun.* (2019).

[44] B. E. Dutilh. "A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes". In: *Nat. Commun.* 5 (2014). DOI: 10.1038/ncomms5498. URL: https://doi.org/10.1038/ncomms5498.

[45] "Edwards, R. A. et al. Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. Nat. Microbiol. 4, 1727–1736 (2019)." In: ().

[46] Oana M. Enache et al. "The GCTx format and cmap{Py, R, M, J} packages: resources for optimized storage and integrated traversal of annotated dense matrices". In: *Bioinformatics (Oxford, England)* 35.8 (Apr. 15, 2019), pp. 1427–1429. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bty784.

[47] Philip A. Ewels et al. "The nf-core framework for community-curated bioinformatics pipelines". In: *Nature Biotechnology* 38.3 (Mar. 2020), pp. 276–278. ISSN: 1546-1696. DOI: 10.1038/s41587-020-0439-x. URL: http://www.nature.com/articles/s41587-020-0439-x (visited on 10/03/2021).

[48] C. Ewers. "CTX-M-15-D-ST648 Escherichia coli from companion animals and horses: another pandemic clone combining multiresistance and extraintestinal virulence?" In: *J. Antimicrob. Chemother.* (2014).

[49] A. D. Fernandes. "Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis". In: *Microbiome* 2 (2014). DOI: 10.1186/2049-2618-2-15. URL: https://doi.org/10.1186/2049-2618-2-15.

[50] P. Ferretti. "Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome". In: *Cell Host Microbe* 24 (2018). DOI: 10.1016/j.chom.2018.06.005. URL: https://doi.org/10.1016/j.chom.2018.06.005.

[51] S. C. Forster. "A human gut bacterial genome and culture collection for improved metagenomic analyses." In: *Nat. Biotechnol.* (2019).

[52] T. Fukami. "Historical Contingency in Community Assembly: Integrating Niches, Species Pools, and Priority Effects." In: *Annu. Rev. Ecol. Evol. Syst.* (2015).

[53] "Garrison, E. & G., M. Haplotype-based variant detection from short-read sequencing. Preprint at https://arxiv.org/abs/1207.3907 (2012)." In: (). URL: https://arxiv.org/abs/1207.3907.

[54] Nandita R. Garud et al. "Evolutionary dynamics of bacteria in the gut microbiome within and across hosts". In: *PLOS Biology* 17.1 (Jan. 23, 2019), e3000102. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.3000102. URL: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000102 (visited on 11/08/2019).

[55] M. G. K. Ghequire. "Different ancestries of R tailocins in rhizospheric pseudomonas isolates". In: *Genome Biol. Evol.* 7 (2015). DOI: 10.1093/gbe/evv184. URL: https://doi.org/10.1093/gbe/evv184.

[56] A. Giraud. "Dissecting the Genetic Components of Adaptation of Escherichia coli to the Mouse Gut." In: *PLOS Genet.* (2008).

[57] E. Goz et al. "Evidence of translation efficiency adaptation of the coding regions of the bacteriophage lambda". In: *DNA Res.* 24 (2017). DOI: 10.1093/dnares/dsx005. URL: https://doi.org/10.1093/dnares/dsx005.

[58] Z. Gu et al. "Circlize implements and enhances circular visualization in R". In: *Bioinformatics.* 30 (2014). DOI: 10.1093/bioinformatics/btu393. URL: https://doi.org/10.1093/bioinformatics/btu393.

[59] E. Guerin et al. "Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut". In: *Cell Host & Microbe* 24 (2018). DOI: 10.1016/j.chom.2018.10.002. URL: https://doi.org/10.1016/j.chom.2018.10.002.

[60] A. Gurevich et al. "QUAST: quality assessment tool for genome assemblies". In: *Bioinformatics* 29 (2013). DOI: 10.1093/bioinformatics/btt086. URL: https://doi.org/10.1093/bioinformatics/btt086.

[61] E. A. Gurnee. "Gut Colonization of Healthy Children and Their Mothers With Pathogenic Ciprofloxacin-Resistant Escherichia coli." In: *J. Infect. Dis.* (2015).

[62] Gonzalo Gutiérrez-García et al. "A reproducible and safe at-home allogeneic haematopoietic cell transplant program: first experience in Central and Southern Europe". In: *Bone Marrow Transplantation* 55.5 (May 2020), pp. 965–973. ISSN: 1476-5365. DOI: 10.1038/s41409-019-0768-x. URL: http://www.nature.com/articles/s41409-019-0768-x (visited on 10/07/2021).

[63] Paul A. Harris et al. "The REDCap consortium: Building an international community of software platform partners". In: *Journal of Biomedical Informatics* 95 (July 2019), p. 103208. ISSN: 1532-0480. DOI: 10.1016/j.jbi.2019.103208.

[64] Jan-Hendrik Hehemann et al. "Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota". In: *Nature* 464.7290 (Apr. 2010), pp. 908–912. ISSN: 1476-4687. DOI: 10.1038/nature08937. URL: http://www.nature.com/articles/nature08937 (visited on 10/02/2021).

[65] T. A. Henderson. "AmpC and AmpH, proteins related to the class C beta-lactamases, bind penicillin and contribute to the normal morphology of Escherichia coli." In: *J. Bacteriol.* (1997).

[66] "Honap, T. P. et al. Biogeographic study of human gut associated crAssphage suggests impacts from industrialization and recent expansion. Preprint at https://www.biorxiv.org/content/10.1101/384677v2 (2019)." In: (). URL: https://www.biorxiv.org/content/10.1101/384677v2.

[67] B. P. Howden. "Genomic Insights to Control the Emergence of Vancomycin-Resistant Enterococci". In: *mBio* (2013).

[68] H. E. Jakobsson. "Decreased gut microbiota diversity, delayed Bacteroidetes colonisation and reduced Th1 responses in infants delivered by caesarean section". In: *Gut* 63 (2014). DOI: 10.1136/gutjnl-2012-303249. URL: https://doi.org/10.1136/gutjnl-2012-303249.

[69] R. R. Jenq. "Intestinal Blautia Is Associated with Reduced Death from Graft-versus-Host Disease." In: *Biol. Blood Marrow Transplant.* (2015).

[70] B. Jia et al. "CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database". In: *Nucleic Acids Res* 45 (2017). DOI: 10.1093/nar/gkw1004. URL: https://doi.org/10.1093/nar/gkw1004.

[71] E. Jiménez. "Metagenomic analysis of milk of healthy and mastitis-suffering women". In: *J. Hum. Lact.* 31 (2015). DOI: 10.1177/0890334415585078. URL: https://doi.org/10.1177/0890334415585078.

[72] D. D. Kang. "MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies." In: *PeerJ* (2019). DOI: doi:10.1038/s41564-018-0171-1.

[73] D. D. Kang et al. "MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities". In: *PeerJ* 3 (2015). DOI: 10.7717/peerj.1165. URL: https://doi.org/10.7717/peerj.1165.

[74] J. B. Kang. "Intestinal microbiota domination under extreme selective pressures characterized by metagenomic read cloud sequencing and assembly." In: *BMC Bioinformatics* (2019).

[75] S. Kaur. "Hungatella effluvii gen. nov., sp. nov., an obligately anaerobic bacterium isolated from an effluent treatment plant, and reclassification of Clostridium hathewayi as Hungatella hathewayi gen. nov., comb. nov." In: *Int. J. Syst. Evol. Microbiol.* (2014).

[76] M. S. Kelly. "Gut Colonization Preceding Mucosal Barrier Injury Bloodstream Infection in Pediatric Hematopoietic Stem Cell Transplantation Recipients." In: *Biol. Blood Marrow Transplant.* (2019).

[77] Kristopher Kieft, Zhichao Zhou, and Karthik Anantharaman. "VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences". In: *Microbiome* 8.1 (Dec. 2020), pp. 1–23. ISSN: 2049-2618. DOI: 10.1186/s40168-020-00867-0. URL: http://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-020-00867-0 (visited on 10/07/2021).

[78] H. Kittana. "Commensal Escherichia coli Strains Can Promote Intestinal Inflammation via Differential Interleukin-6 Production." In: *Front. Immunol.* (2018).

[79] K. Korpela. "Selective maternal seeding and environment shape the human gut microbiome". In: *Genome Res.* 28 (2018). DOI: 10.1101/gr.233940.117. URL: https://doi.org/10.1101/gr.233940.117.

[80] J. Köster and S. Rahmann. "Snakemake—a scalable bioinformatics workflow engine." In: *Bioinformatics* (2012).

[81] "Krueger, F. Trim Galore! Available at: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore." In: (). URL: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore.

[82] J. Kumar et al. "An improved methodology to overcome key issues in human fecal metagenomic DNA extraction". In: *Genomics Proteomics Bioinformatics* 14 (2016). DOI: 10.1016/j.gpb.2016.06.002. URL: https://doi.org/10.1016/j.gpb.2016.06.002.

[83] S. Kurtz. "Versatile and open software for comparing large genomes." In: *Genome Biol.* (2004). DOI: doi:10.1093/nar/gkz935.

[84] L. et al L. Livornese Jr. "Hospital-acquired Infection with Vancomycin-resistant Enterococcus faecium Transmitted by Electronic Thermometers." In: *Ann. Intern. Med.* (1992).

[85] Emily R Leeming et al. "Effect of Diet on the Gut Microbiota: Rethinking Intervention Duration". In: *Nutrients* 11.12 (Nov. 22, 2019), p. 2862. ISSN: 2072-6643. DOI: 10.3390/nu11122862. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6950569/ (visited on 09/28/2021).

[86] D. Li et al. "MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph". In: *Bioinformatics* 31 (2015). DOI: 10.1093/bioinformatics/btv033. URL: https://doi.org/10.1093/bioinformatics/btv033.

[87] H. Li and R. Durbin. "Fast and accurate short read alignment with Burrows-Wheeler transform". In: *Bioinformatics* 25 (2009). DOI: 10.1093/bioinformatics/btp324. URL: https://doi.org/10.1093/bioinformatics/btp324.

[88] H. Li and R. Durbin. "Fast and accurate short read alignment with burrows-wheeler transform". In: *Bioinformatics* 26 (2010). DOI: 10.1093/bioinformatics/btp698. URL: https://doi.org/10.1093/bioinformatics/btp698.

[89] H. Li et al. "The sequence alignment/map format and SAMtools". In: *Bioinformatics* 25 (2009). DOI: 10.1093/bioinformatics/btp352. URL: https://doi.org/10.1093/bioinformatics/btp352.

[90] S. S. Li. "Durable coexistence of donor and recipient strains after fecal microbiota transplantation." In: *Science* (2016).

[91] Y. Liang et al. "Development and application of a real-time polymerase chain reaction assay for detection of a novel gut bacteriophage (crAssphage)". In: *J. Med. Virol.* 90 (2018). DOI: 10.1002/jmv.24974. URL: https://doi.org/10.1002/jmv.24974.

[92] Y. Y. Liang et al. "crAssphage is not associated with diarrhoea and has high genetic diversity". In: *Epidemiol. Infect.* 144 (2016). DOI: 10.1017/S095026881600176X. URL: https://doi.org/10.1017/S095026881600176X.

[93] E. S. Lim. "Early life dynamics of the human gut virome and bacterial microbiome in infants". In: *Nat. Med.* 21 (2015). DOI: 10.1038/nm.3950. URL: https://doi.org/10.1038/nm.3950.

[94] H. H. Lin and Y. C. Liao. "Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes". In: *Sci Rep* 6 (2016). DOI: 10.1038/srep24175. URL: https://doi.org/10.1038/srep24175.

[95] A. J. Linscott. "Fatal septicemia due to Clostridium hathewayi and Campylobacter hominis." In: *Anaerobe* (2005).

[96] Y. Liu. "The perturbation of infant gut microbiota caused by cesarean delivery is partially restored by exclusive breastfeeding". In: *Front. Microbiol.* 10 (2019). DOI: 10.3389/fmicb.2019.00598. URL: https://doi.org/10.3389/fmicb.2019.00598.

[97] J. Lloyd-Price. "Strains, functions and dynamics in the expanded Human Microbiome Project". In: *Nature* (2017). DOI: doi:10.1128/microbiolspec.BAD-0014-2016.

[98] J. Lloyd-Price, G. Abu-Ali, and C. Huttenhower. "The healthy human microbiome". In: *Genome Med* 8 (2016). DOI: 10.1186/s13073-016-0307-y. URL: https://doi.org/10.1186/s13073-016-0307-y.

[99] C. A. Lozupone. "Diversity, stability and resilience of the human gut microbiota." In: *Nature* (2012). DOI: doi:10.1038/s41564-019-0409-6.

[100] J. Lu et al. "Bracken: estimating species abundance in metagenomics data". In: *PeerJ Comput. Sci.* 3 (2017). DOI: 10.7717/peerj-cs.104. URL: https://doi.org/10.7717/peerj-cs.104.

[101] T. Madigan. "Extensive Household Outbreak of Urinary Tract Infection and Intestinal Colonization due to Extended-Spectrum -Lactamase–Producing Escherichia coli Sequence Type 131." In: *Clin. Infect. Dis.* (2015).

[102] G. Marçais. "MUMmer4: a fast and versatile genome alignment system". In: *PLoS Comput. Biol.* 14 (2018). DOI: 10.1371/journal.pcbi.1005944. URL: https://doi.org/10.1371/journal.pcbi.1005944.

[103] Angela Marcobal et al. "Consumption of Human Milk Oligosaccharides by Gut-related Microbes". In: *Journal of agricultural and food chemistry* 58.9 (May 12, 2010), pp. 5334–5340. ISSN: 0021-8561. DOI: 10.1021/jf9044205. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2866150/ (visited on 10/08/2021).

[104] M. Martin. "Cutadapt removes adapter sequences from high-throughput sequencing reads". In: *EMBnet J* 17 (2011). DOI: 10.14806/ej.17.1.200. URL: https://doi.org/10.14806/ej.17.1.200.

[105] N. D. Mathewson. "Gut microbiome–derived metabolites modulate intestinal epithelial cell damage and mitigate graft-versus-host disease." In: *Nat. Immunol.* (2016).

[106] A. McCann. "Viromes of one year old infants reveal the impact of birth mode on microbiome diversity". In: *PeerJ* 6 (2018). DOI: 10.7717/peerj.4694. URL: https://doi.org/10.7717/peerj.4694.

[107] "Minimize index hopping in multiplexed runs. Illumina Available at: https://www.illumina.com/science/education/min index-hopping.html." In: (). URL: https://www.illumina.com/science/education/minimizing-index-hopping.html.

[108] Andrew H. Moeller et al. "Transmission modes of the mammalian gut microbiota". In: *Science* 362.6413 (Oct. 26, 2018), pp. 453–457. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aat7164. URL: http://science.sciencemag.org/content/362/6413/453 (visited on 10/27/2018).

[109] Livia H. Morais, Henry L. Schreiber, and Sarkis K. Mazmanian. "The gut microbiota–brain axis in behaviour and brain disorders". In: *Nature Reviews Microbiology* 19.4 (Apr. 2021), pp. 241–255. ISSN: 1740-1534. DOI: 10.1038/s41579-020-00460-0. URL: http://www.nature.com/articles/s41579-020-00460-0 (visited on 10/02/2021).

[110] E. L. Moss. "Complete, closed bacterial genomes from microbiomes using nanopore sequencing." In: *Nat. Biotechnol.* (2020).

[111] E. L. Moss. "Long-term taxonomic and functional divergence from donor bacterial strains following fecal microbiota transplantation in immunocompromised patients". In: *PLoS One* 12 (2017). DOI: 10.1371/journal.pone.0182585. URL: https://doi.org/10.1371/journal.pone.0182585.

[112] A. Müller. "Distribution of virulence factors in ESBL-producing Escherichia coli isolated from the environment, livestock, food and humans." In: *Sci. Total Environ.* (2016).

[113] M. Nattestad. "MariaNattestad/dot." In: (2020).

[114] noauthor. "Babraham Bioinformatics - Trim Galore!" In: (). URL: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

[115] noauthor. "Minimizing Index Hopping." In: (). URL: https://www.illumina.com/techniques/sequencing/ngs-library-prep/multiplexing/index-hopping.html.

[116] P. Nordmann, L. Dortet, and L. Poirel. "Carbapenem resistance in Enterobacteriaceae: here is the storm!" In: *Trends Mol Med* 18 (2012). DOI: 10.1016/j.molmed.2012.03.003. URL: https://doi.org/10.1016/j.molmed.2012.03.003.

[117] S. Nurk. "Assembling single-cell genomes and mini-metagenomes from chimeric MDA products". In: *J. Comput. Biol.* 20 (2013). DOI: 10.1089/cmb.2013.0084. URL: https://doi.org/10.1089/cmb.2013.0084.

[118] S. Nurk et al. "metaSPAdes: a new versatile metagenomic assembler". In: *Genome Res* 27 (2017). DOI: 10.1101/gr.213959.116. URL: https://doi.org/10.1101/gr.213959.116.

[119] J. Oksanen. "vegan: Community Ecology Package." In: (2020).

[120] "Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre R, McGlinn D, et al. vegan: Community Ecology Package. R package version 2.5–3. 2018. https://CRAN.R-project.org/package=vegan. Accessed 1 Aug 2018." In: (). URL: https://cran.r-project.org/package=vegan.

[121] M. R. Olm. "dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication". In: *ISME J* (2017).

[122] M. R. Olm. "inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains." In: *Nat. Biotechnol.* (2021).

[123] M. R. Olm. "Necrotizing enterocolitis is preceded by increased gut bacterial replication, Klebsiella, and fimbriae-encoding bacteria." In: *Sci. Adv.* (2019).

[124] Matthew R. Olm et al. "Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries". In: *mSystems* 5.1 (Feb. 25, 2020). ISSN: 2379-5077. DOI: 10.1128/mSystems.00731-19. URL: https://msystems.asm.org/content/5/1/e00731-19 (visited on 01/16/2020).

[125] Noora Ottman et al. "The function of our microbiota: who is out there and what do they do?" In: *Frontiers in Cellular and Infection Microbiology* 2 (2012), p. 104. ISSN: 2235-2988. DOI: 10.3389/fcimb.2012.00104. URL: https://www.frontiersin.org/article/10.3389/fcimb.2012.00104 (visited on 10/03/2021).

[126] M. D. Paepe. "Trade-Off between Bile Resistance and Nutritional Competence Drives Escherichia coli Diversification in the Mouse Gut." In: *PLOS Genet.* (2011).

[127] J. Palarea-Albaladejo and J. A. Martín-Fernández. "zCompositions — R package for multivariate imputation of left-censored data under a compositional approach". In: *Chemometrics Intell. Lab. Syst.* 143 (2015). DOI: 10.1016/j.chemolab.2015.02.019. URL: https://doi.org/10.1016/j.chemolab.2015.02.019.

[128] K. L. Palmer. "Comparative Genomics of Enterococci: Variation in Enterococcus faecalis, Clade Structure in E. faecium, and Defining Characteristics of E. gallinarum and E. casseliflavus." In: *mBio* (2012).

[129] D. H. Parks et al. "CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes". In: *Genome Res* 25 (2015). DOI: 10.1101/gr.186072.114. URL: https://doi.org/10.1101/gr.186072.114.

[130] Donovan H Parks et al. "GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy". In: *Nucleic Acids Research* (gkab776 Sept. 14, 2021). ISSN: 0305-1048. DOI: 10.1093/nar/gkab776. URL: https://doi.org/10.1093/nar/gkab776 (visited on 10/07/2021).

[131] E. Paulshus. "Repeated Isolation of Extended-Spectrum--Lactamase-Positive Escherichia coli Sequence Types 648 and 131 from Community Wastewater Indicates that Sewage Systems Are Important Sources of Emerging Clones of Antibiotic-Resistant Bacteria". In: *Antimicrob. Agents Chemother.* (2019).

[132] J. U. Peled. "Microbiota as Predictor of Mortality in Allogeneic Hematopoietic-Cell Transplantation." In: *N. Engl. J. Med.* (2020).

[133] J. Pereira-Marques. "Impact of Host DNA and Sequencing Depth on the Taxonomic Resolution of Whole Metagenome Sequencing for Microbiome Analysis." In: *Front. Microbiol.* (2019).

[134] V. Pittet. "Genome Sequence of Lactobacillus rhamnosus ATCC 8530." In: *J. Bacteriol.* (2012).

[135] Ramawatar. "DNA size selection (¿3-4kb) and purification of DNA using an improved home-made SPRI beads solution." In: (2018).

[136] A. Rashidi et al. "Pre-transplant recovery of microbiome diversity without recovery of the original microbiome". In: *Bone Marrow Transplant* 29 (2018).

[137] K. E. Raven. "Complex Routes of Nosocomial Vancomycin-Resistant Enterococcus faecium Transmission Revealed by Genome Sequencing." In: *Clin. Infect. Dis.* (2017).

[138] Jie Ren et al. "VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data". In: *Microbiome* 5.1 (July 6, 2017), p. 69. ISSN: 2049-2618. DOI: 10.1186/s40168-017-0283-5. URL: https://doi.org/10.1186/s40168-017-0283-5 (visited on 12/10/2019).

[139] J. Reunanen. "Akkermansia muciniphila Adheres to Enterocytes and Strengthens the Integrity of the Epithelial Cell Layer." In: *Appl Env. Microbiol* (2015).

[140] A. Reyes. "Viruses in the faecal microbiota of monozygotic twins and their mothers". In: *Nature* 466 (2010). DOI: 10.1038/nature09199. URL: https://doi.org/10.1038/nature09199.

[141] Morteza Roodgar et al. "Longitudinal linked-read sequencing reveals ecological and evolutionary responses of a human gut microbiome during antibiotic treatment". In: *Genome Research* 31.8 (Aug. 1, 2021), pp. 1433–1446. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.265058.120. URL: http://genome.cshlp.org/content/31/8/1433 (visited on 10/07/2021).

[142] Simon Roux et al. "VirSorter: mining viral signal from microbial genomic data". In: *PeerJ* 3 (May 28, 2015), e985. ISSN: 2167-8359. DOI: 10.7717/peerj.985. URL: https://peerj.com/articles/985 (visited on 10/06/2021).

[143] K. Schaufler. "Genomic and Functional Analysis of Emerging Virulent and Multidrug-Resistant Escherichia coli Lineage Sequence Type 648." In: *Antimicrob. Agents Chemother.* (2019).

[144] J. Schluter. "The gut microbiota is associated with immune cell dynamics in humans." In: *Nature* (2020).

[145] D. Scholl et al. "Bacteriophage K1-5 encodes two different tail fiber proteins, allowing it to infect and replicate on both K1 and K5 strains of Escherichia coli". In: *J. Virol.* 75 (2001). DOI: 10.1128/JVI.75.6.2509-2515.2001. URL: https://doi.org/10.1128/JVI.75.6.2509-2515.2001.

[146] I. See. "Mucosal Barrier Injury Laboratory-Confirmed Bloodstream Infection: Results from a Field Test of a New National Healthcare Safety Network Definition". In: *Infect. Control Hosp. Epidemiol.* (2013). DOI: doi:10.1038/s41591-019-0709-7.

[147] "Seeman, T. snippy: fast bacterial variant calling from NGS reads. (2015)." In: ().

[148] T. Seemann. "Prokka: rapid prokaryotic genome annotation". In: *Bioinformatics.* 30 (2014). DOI: 10.1093/bioinformatics/btu153. URL: https://doi.org/10.1093/bioinformatics/btu153.

[149] J. Seishima. "Gut-derived Enterococcus faecium from ulcerative colitis patients promotes colitis in a genetically susceptible mouse host." In: *Genome Biol.* (2019).

[150] Shiraz A. Shah et al. "Manual resolution of virome dark matter uncovers hundreds of viral families in the infant gut". In: *bioRxiv* (July 3, 2021), p. 2021.07.02.450849. DOI: 10.1101/2021.07.02.450849. URL: https://www.biorxiv.org/content/10.1101/2021.07.02.450849v1 (visited on 07/07/2021).

[151] Y. Shao. "Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth". In: *Nature* 574 (2019). DOI: 10.1038/s41586-019-1560-1. URL: https://doi.org/10.1038/s41586-019-1560-1.

[152] W. Shen et al. "SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation". In: *PLoS One* 11 (2016). DOI: 10.1371/journal.pone.0163962. URL: https://doi.org/10.1371/journal.pone.0163962.

[153] A. N. Shkoporov. "The human gut virome is highly diverse, stable, and individual specific". In: *Cell Host Microbe* 26 (2019). DOI: 10.1016/j.chom.2019.09.009. URL: https://doi.org/10.1016/j.chom.2019.09.009.

[154] A. N. Shkoporov. "CrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis". In: *Nat. Commun.* 9 (2018). DOI: 10.1038/s41467-018-07225-7. URL: https://doi.org/10.1038/s41467-018-07225-7.

[155] Y. Shono. "Increased GVHD-related mortality with broad-spectrum antibiotic use after allogeneic hematopoietic stem cell transplantation in human patients and mice." In: *Sci. Transl. Med.* (2016).

[156] Y. Shono and M. R. M. van den Brink. "Gut microbiota injury in allogeneic haematopoietic stem cell transplantation." In: *Nat. Rev. Cancer* (2018).

[157] C. M. K. Sieber et al. "Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy". In: *Nat Microbiol* 3 (2018). DOI: 10.1038/s41564-018-0171-1. URL: https://doi.org/10.1038/s41564-018-0171-1.

[158] T. R. Simms-Waldrip. "Antibiotic-Induced Depletion of Anti-inflammatory Clostridia Is Associated with the Development of Graft-versus-Host Disease in Pediatric Stem Cell Transplantation Patients." In: *Biol. Blood Marrow Transplant.* (2017).

[159] B. A. Siranosian. "Acquisition, transmission and strain diversity of human gut-colonizing crAss-like phages." In: *Nat. Commun.* (2020).

[160] Ben Siranosian and Eli Moss. *bhattlab/kraken2_classification: A mostly finished pipeline.* Aug. 19, 2021. DOI: 10.5281/zenodo.5219057. URL: https://zenodo.org/record/5219057 (visited on 09/25/2021).

[161] Ben Siranosian et al. *bhattlab/bhattlab_workflows: v1.0.1.* Oct. 4, 2021. DOI: 10.5281/zenodo.5546646. URL: https://zenodo.org/record/5546646 (visited on 10/04/2021).

[162] C. S. Smillie. "Strain tracking reveals the determinants of bacterial engraftment in the human gut following fecal microbiota transplantation". In: *Cell Host Microbe* 23 (2018). DOI: 10.1016/j.chom.2018.01.003. URL: https://doi.org/10.1016/j.chom.2018.01.003.

[163] Chuliang Song, Tadashi Fukami, and Serguei Saavedra. *Untangling the complexity of priority effects in multispecies communities.* Aug. 4, 2021, p. 2021.03.29.437541. DOI: 10.1101/2021.03.29.437541. URL: https://www.biorxiv.org/content/10.1101/2021.03.29.437541v2 (visited on 08/12/2021).

[164] Se Jin Song et al. "Cohabiting family members share microbiota with one another and with their dogs". In: *eLife* 2 (Apr. 16, 2013). Ed. by Detlef Weigel. Publisher: eLife Sciences Publications, Ltd, e00458. ISSN: 2050-084X. DOI: 10.7554/eLife.00458. URL: https://doi.org/10.7554/eLife.00458 (visited on 10/08/2021).

[165] Daniel Sprockett, Tadashi Fukami, and David A. Relman. "Role of priority effects in the early-life assembly of the gut microbiota". In: *Nature Reviews Gastroenterology & Hepatology* 15.4 (Apr. 2018), pp. 197–205. ISSN: 1759-5053. DOI: 10.1038/nrgastro.2017.173. URL: http://www.nature.com/articles/nrgastro.2017.173 (visited on 06/19/2020).

[166] E. Stachler et al. "Correlation of crAssphage qPCR markers with culturable and molecular indicators of human fecal pollution in an impacted urban watershed". In: *Environ. Sci. Technol.* 52 (2018). DOI: 10.1021/acs.est.8b00638. URL: https://doi.org/10.1021/acs.est.8b00638.

[167] J. Suez. "Diversity, stability and resilience of the human gut microbiota." In: *Cell* (2018).

[168] F. B. Tamburini et al. "Precision identification of diverse bloodstream pathogens in the gut microbiome". In: *Nat Med* 24 (2018). DOI: 10.1038/s41591-018-0202-8. URL: https://doi.org/10.1038/s41591-018-0202-8.

[169] A. Tan, G. R. Abecasis, and H. M. Kang. "Unified representation of genetic variants". In: *Bioinformatics* 31 (2015). DOI: 10.1093/bioinformatics/btv112. URL: https://doi.org/10.1093/bioinformatics/btv112.

[170] Y. Taur et al. "Intestinal domination and the risk of bacteremia in patients undergoing allogeneic hematopoietic stem cell transplantation". In: *Clin Infect Dis* 55 (2012). DOI: 10.1093/cid/cis580. URL: https://doi.org/10.1093/cid/cis580.

[171] Y. Taur et al. "The effects of intestinal tract bacterial diversity on mortality following allogeneic hematopoietic stem cell transplantation". In: *Blood* 124 (2014). DOI: 10.1182/blood-2014-02-554725. URL: https://doi.org/10.1182/blood-2014-02-554725.

[172] Ying Taur et al. "Reconstitution of the gut microbiota of antibiotic-treated patients by autologous fecal microbiota transplant". In: *Science Translational Medicine* 10.460 (Sept. 26, 2018), eaap9489. ISSN: 1946-6234, 1946-6242. DOI: 10.1126/scitranslmed.aap9489. URL: http://stm.sciencemag.org/content/10/460/eaap9489 (visited on 09/27/2018).

[173] R Development Core Team. "R: A Language and Environment for Statistical Computing." In: (2012).

[174] A. P. Tedim. "Complete Genome Sequences of Isolates of Enterococcus faecium Sequence Type 117, a Globally Disseminated Multidrug-Resistant Clone." In: *Genome Announc.* (2017).

[175] O. Tenaillon. "The population genetics of commensal Escherichia coli." In: *Nat. Rev. Microbiol.* (2010).

[176] Pamela Thomson, Daniel A. Medina, and Daniel Garrido. "Human milk oligosaccharides and infant gut bifidobacteria: Molecular strategies for their utilization". In: *Food Microbiology* 75 (Oct. 2018), pp. 37–46. ISSN: 1095-9998. DOI: 10.1016/j.fm.2017.09.001.

[177] M. Touchon, J. A. Moura de Sousa, and E. P. Rocha. "Embracing the enemy: the diversification of microbial gene repertoires by phage-mediated horizontal gene transfer". In: *Curr. Opin. Microbiol.* 38 (2017). DOI: 10.1016/j.mib.2017.04.010. URL: https://doi.org/10.1016/j.mib.2017.04.010.

[178] D. T. Truong. "Microbial strain-level population structure and genetic diversity from metagenomes." In: *Genome Res.* (2017).

[179] Y.-C. Tsai. "Resolving the Complexity of Human Skin Metagenomes Using Single-Molecule Sequencing." In: *mBio* (2016).

[180] Peter J. Turnbaugh and Jeffrey I. Gordon. "The core gut microbiome, energy balance and obesity". In: *The Journal of Physiology* 587.17 (2009), pp. 4153–4158. ISSN: 1469-7793. DOI: 10.1113/jphysiol.2009.174136. URL: https://onlinelibrary.wiley.com/doi/abs/10.1113/jphysiol.2009.174136 (visited on 10/02/2021).

[181] C. Ubeda. "Vancomycin-resistant Enterococcus domination of intestinal microbiota is enabled by antibiotic treatment in mice and precedes bloodstream invasion in humans." In: *J. Clin. Invest.* (2010).

[182] T. Van Rossum. "Diversity within species: interpreting strains in microbiomes." In: *Nat. Rev. Microbiol.* (2020).

[183] "Warnes, G. R. et al. gplots: Various R Programming Tools for Plotting Data." In: ().

[184] A. R. Watson. "Adaptive ecological processes and metabolic independence drive microbial colonization and resilience in the human gut." In: *bioRxiv* (2021). DOI: doi:10.1038/s41409-018-0414-z.

[185] D. Weber. "Microbiota Disruption Induced by Early Use of Broad-Spectrum Antibiotics Is an Independent Risk Factor of Outcome after Allogeneic Stem Cell Transplantation." In: *Biol. Blood Marrow Transplant.* (2017).

[186] G. R. Whitmer. "The pandemic Escherichia coli sequence type 131 strain is acquired even in the absence of antibiotic exposure." In: *PLOS Pathog.* (2019).

[187] Richard Wolff, William R. Shoemaker, and Nandita R. Garud. *Ecological Stability Emerges at the Level of Strains in the Human Gut Microbiome*. Oct. 1, 2021, p. 2021.09.30.462616. DOI: 10.1101/2021.09.30.462616. URL: https://www.biorxiv.org/content/10.1101/2021.09.30.462616v1 (visited on 10/03/2021).

[188] P. C. Y. Woo. "Bacteremia Due to Clostridium hathewayi in a Patient with Acute Appendicitis." In: *J. Clin. Microbiol.* (2004).

[189] D. E. Wood. "Improved metagenomic analysis with Kraken 2". In: *Genome Biol.* (2019).

[190] D. E. Wood and S. L. Salzberg. "Kraken: ultrafast metagenomic sequence classification using exact alignments". In: *Genome Biol* 15 (2014). DOI: 10.1186/gb-2014-15-3-r46. URL: https://doi.org/10.1186/gb-2014-15-3-r46.

[191] Laura Wratten, Andreas Wilm, and Jonathan Göke. "Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers". In: *Nature Methods* (Sept. 23, 2021), pp. 1–8. ISSN: 1548-7105. DOI: 10.1038/s41592-021-01254-9. URL: http://www.nature.com/articles/s41592-021-01254-9 (visited on 10/03/2021).

[192] Y. W. Wu, B. A. Simmons, and S. W. Singer. "MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets". In: *Bioinformatics* 32 (2016). DOI: 10.1093/bioinformatics/btv638. URL: https://doi.org/10.1093/bioinformatics/btv638.

[193] Eitan Yaffe and David A. Relman. "Tracking microbial evolution in the human gut using Hi-C reveals extensive horizontal gene transfer, persistence and adaptation". In: *Nature Microbiology* 5.2 (Feb. 2020), pp. 343–353. ISSN: 2058-5276. DOI: 10.1038/s41564-019-0625-0. URL: http://www.nature.com/articles/s41564-019-0625-0 (visited on 10/03/2021).

[194] Jinyuan Yan et al. *A compilation of fecal microbiome shotgun metagenomics from hospitalized patients undergoing hematopoietic cell transplantation*. Aug. 25, 2021, p. 2021.08.23.457365. DOI: 10.1101/2021.08.23.457365. URL: https://www.biorxiv.org/content/10.1101/2021.08.23.457365v2 (visited on 09/21/2021).

[195] M. Yassour. "Strain-level analysis of mother-to-child bacterial transmission during the first few months of life". In: *Cell Host Microbe* 24 (2018). DOI: 10.1016/j.chom.2018.06.007. URL: https://doi.org/10.1016/j.chom.2018.06.007.

[196] Idan Yelin et al. "Genomic and epidemiological evidence of bacterial transmission from probiotic capsule to blood in ICU patients". In: *Nature Medicine* 25.11 (Nov. 2019). Number: 11 Publisher: Nature Publishing Group, pp. 1728–1732. ISSN: 1546-170X. DOI: 10.1038/s41591-019-0626-9. URL: http://www.nature.com/articles/s41591-019-0626-9 (visited on 10/26/2020).

[197] S. H. Yi. "Prevalence of probiotic use among inpatients: A descriptive study of 145 U.S. hospitals." In: *Am. J. Infect. Control* (2016).

[198] J.-A. H. Young. "Infections after Transplantation of Bone Marrow or Peripheral Blood Stem Cells from Unrelated Donors". In: *Biol. Blood Marrow Transplant.* (2016). DOI: doi:10.1038/s41591-018-0202-8.

[199] N. Yutin. "Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut". In: *Nat. Microbiol.* 3 (2018). DOI: 10.1038/s41564-017-0053-y. URL: https://doi.org/10.1038/s41564-017-0053-y.

[200] B. Zhai. "High-resolution mycobiota analysis reveals dynamic intestinal translocation preceding invasive candidiasis." In: *Nat. Med.* (2020).

[201] Yan Zhang et al. "Effects of probiotic type, dose and treatment duration on irritable bowel syndrome diagnosed by Rome III criteria: a meta-analysis". In: *BMC Gastroenterology* 16.1 (Dec. 2016). Number: 1 Publisher: BioMed Central, pp. 1–11. ISSN: 1471-230X. DOI: 10.1186/s12876-016-0470-z. URL: http://bmcgastroenterol.biomedcentral.com/articles/10.1186/s12876-016-0470-z (visited on 10/08/2021).

[202] G. Zhao. "Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children". In: *Proc. Natl Acad. Sci. USA* 114 (2017). DOI: 10.1073/pnas.1706359114. URL: https://doi.org/10.1073/pnas.1706359114.

[203] G. X. Zheng et al. "Haplotyping germline and cancer genomes with high-throughput linked-read sequencing". In: *Nat Biotechnol* 34 (2016). DOI: 10.1038/nbt.3432. URL: https://doi.org/10.1038/nbt.3432.

[204] N. Zmora. "Personalized Gut Mucosal Colonization Resistance to Empiric Probiotics Is Associated with Unique Host and Microbiome Features." In: *Cell* (2018). DOI: doi:10.1101/2021.03.02.433653.