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

Benjamin A. Siranosian¹, Erin F. Brooks², Tessa Andermann³, Andrew R. Rezvani⁴, Niaz Banaei^{5,6,7}, Hua Tang¹, Ami S. Bhatt^{*1,2,4}

1. Department of Genetics, Stanford University, Stanford, CA, USA
2. Department of Medicine, Division of Hematology, Stanford University, Stanford, CA, USA.
3. Division of Infectious Diseases, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina
4. Department of Medicine, Division of Blood and Marrow Transplantation and Cellular Therapy, Stanford University School of Medicine, Stanford, CA
5. Department of Medicine, Division of Infectious Diseases and Geographic Medicine, Stanford University, Stanford, CA, USA.
6. Clinical Microbiology Laboratory, Stanford University Medical Center, Stanford, CA, USA
7. Department of Pathology, Stanford University, Stanford, CA, USA.

* to whom correspondence should be addressed: asbhatt@stanford.edu

22 Abstract

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

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

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

47 recovery of microbiome diversity after HCT, and indicates that patients in this setting may acquire new
48 microbes by sharing space with others.
49

50 Introduction

51 Patients undergoing hematopoietic cell transplantation (HCT), a potentially curative treatment for
 52 a range of hematologic malignancies and disorders, are at increased risk for bloodstream infections
 53 (BSIs) and associated morbidity and mortality¹. While the bacterial pathogens that cause BSIs in HCT
 54 patients are well understood, their routes of transmission are often unclear. Determining these
 55 transmission pathways involves identifying two critical elements: the source of the infection, i.e., how the
 56 pathogen was introduced into the patient's bloodstream, and the origins of the particular pathogen
 57 causing the BSI.

58 The most common ways bacterial pathogens can be introduced into an HCT patient's
 59 bloodstream include contaminated central intravenous lines and translocation of intestinal microbiota
 60 across a damaged epithelium². Indeed, research from our group and others has shown that strains of
 61 bacteria isolated from the blood of HCT patients with BSIs may be indistinguishable from the strains in
 62 the intestinal microbiota of these patients prior to infection³⁻⁵. In addition, HCT patients with a
 63 microbiome dominated by a single bacterial taxon, such as *Enterococcus* or *Streptococcus*, are at
 64 increased risk for not only BSI^{6,7}, but also graft-versus-host disease^{8,9} and death¹⁰⁻¹³.

65 Identifying the source of the BSI is only the first step. To fully understand the transmission
 66 pathways of bacterial pathogens in hospital settings, it is also essential to determine the origin of the
 67 pathogen that caused the BSI. For gut-based pathogens, there are three possibilities. First, they may
 68 exist in the HCT patient's microbiome upon admission to the hospital. Second, hospital environments
 69 and equipment may serve as unintentional reservoirs of pathogens, thereby infecting multiple patients
 70 through exposure¹⁴. Lastly, a pathogen could originate from the microbiome of other patients, healthcare
 71 workers or hospital visitors and be transmitted via shared spaces. In cases where traditional
 72 epidemiological links cannot be found, this patient-patient transmission of gut microbes may be the
 73 "missing link" that explains the persistence of BSIs in hospital environments¹⁵.

74 Transmission of gut bacteria and phages between individuals is known to occur in specific cases,
 75 such as from mothers to developing infants¹⁶⁻¹⁸. By contrast, adults have a microbiome that is relatively

resistant to colonization with new organisms even after perturbation by antibiotics^{19–21}. While adults living in the same household or in close-knit communities may have more similar microbes than those outside the group²², 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^{23–25}, a drastic reshaping of the gut microbiota often used in response to *Clostridioides 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^{10,26–28}, 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^{29,30} while retaining high per-base accuracy. Both of these technologies also capture information about strain diversity, genetic variation within the population of a species^{31,32}, 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

102 organisms, demonstrating that transmission may depend more on niche availability than pathogenicity or
 103 antibiotic resistance capacity. We find that pathogens colonizing HCT patient microbiomes are present in
 104 the first sample in a time course roughly 60-70% of the time in our cohort. This suggests that in most
 105 cases, prior colonization, rather than direct transmission from other patients or the hospital environment,
 106 is responsible for pathogenic organisms in the gut microbiomes of this patient population. Even though
 107 patients were frequently placed into double occupancy hospital rooms with a shared bathroom, we
 108 observe relatively few putative transmission events. This implies that sharing a room with another patient
 109 may not place a patient recovering from HCT at a greatly increased risk of acquiring pathogens in their
 110 gut microbiome.

111

112 Results

113 Sample characteristics and patient geography

114 We collected weekly stool samples (see methods) from adult patients undergoing hematopoietic
 115 cell transplantation (HCT) at Stanford University Medical Center from 2015-2019. At the time of the
 116 study, our biobank contained over 2000 stool samples from over 900 patients. Samples from October
 117 2015 to November 2018 were considered for this study. Relevant patient health, medication,
 118 demographic, hospital admission and room occupancy data were extracted from electronic health
 119 records (Table 1, Table S1). All patients stayed in a single ward of the hospital during treatment, which
 120 contained 14 single-occupancy and four double-occupancy rooms, the latter of which included shared
 121 bathrooms (Figure S1a). Patients spent a median of 18 days on the ward and were frequently moved
 122 between rooms: 42% of patients spent at least one day in three or more rooms during the course of
 123 treatment (Table 2, Figure S1c). 73% of patients shared a room with a roommate for ≥ 24 hours. Over the
 124 course of their hospital stays, many patients had several roommates, though never more than one at a
 125 time (Figure S1d). Patients with multi-resistant Gram-negative infections were always placed into single
 126 rooms with contact precautions.

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

134 Metagenomic sequencing, assembly and binning

135 For an overview of the steps used in the generation and processing of sequence data, see Figure
 136 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 metaSPAdes³³ for short-read data, and MEGAHIT³⁴ followed by Athena²⁹ for linked-read data. Short-read assemblies had a median N50 of 17.2 kb $\pm$ 24.8 kb, while linked-read assemblies had a median N50 of 147.6 $\pm$ 165.8 kb. We binned metagenome-assembled genomes (MAGs) using Metabat2³⁵, Maxbin³⁶ and CONCOCT³⁷ and aggregated across results from each tool using DASTool³⁸. MAG completeness and contamination was evaluated using CheckM³⁹ and MAG quality was determined by previously established standards⁴⁰. 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 collection⁴¹, 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

163 diversity of strains present in the microbiomes of our patients, we clustered all medium- and high-quality
164 MAGs at 95% and 99% identity thresholds (roughly “species” and “strain” level, see methods) using
165 dRep⁴², yielding 1615 unique genomes representative of the microbial diversity in this sample set.

166

Figure 1: Overview of the methods, data generated, and clinical features of this sample set.

a) Overview of the experimental 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 level⁴⁰, 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

177 by ($\geq 30\%$ relative abundance) hospital acquired infection (HAI) organisms, as identified by Kraken2- and
 178 Bracken-based classification.

179 f) Percentage of bloodstream infections (BSIs) identified with each organism or group in HCT patients
 180 and hospital-wide.

181

Classification of abundant Healthcare-associated Infection organisms

We performed taxonomic classification of sequencing reads with Kraken2⁴³ and abundance estimation with Bracken⁴⁴ using a custom database of bacterial, fungal, archaeal and viral genomes in NCBI Genbank (see methods) (Table S4, S5). A median of 33% $\pm$ 15% SD reads were classified to the species level with Kraken2 (72% $\pm$ 15% SD at the genus level), which was improved to 96% $\pm$ 7% SD using Bracken (97% $\pm$ 8% SD at the genus level). Organisms that cause healthcare-associated Infections (HAI) were identified from the CDC list of pathogens⁴⁵. 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 $\geq$ 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 Project⁴⁶ 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 ($\geq$ 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^{2,47} compared to the gut microbiome (individual species in the group only reach 1% relative abundance in 8/149 patients, 5%).

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).

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, $R^2 = 0.18, 0.14$, $p=8.9e-11, 2.6e-8$ 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).

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

259 2e,f), indicating that the broad taxonomic composition of patient microbiomes may not converge under
 260 temporal and geographic overlap. However, there may be isolated strains that are shared between
 261 patients. We used the strain diversity-aware, SNP-based method inStrain⁴⁸ to conduct a sensitive
 262 analysis of bacterial strains shared between patients. InStrain compares alignments of short reads from
 263 multiple samples to the same reference genome and reports two metrics: Consensus ANI (conANI) and
 264 PopulationANI (popANI). ConANI counts a SNP when two samples differ in the consensus allele at a
 265 position in the reference genome, similar to many conventional SNP calling methods. PopANI counts a
 266 SNP only if both samples share no alleles. For example, if A/T alleles were found at frequencies of
 267 90/10% and 10/90% in two samples, a consensus SNP would be called because the consensus base is
 268 different. A population SNP would not be called because both samples share an A and T allele.

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

Figure 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.

290 Taxonomic similarity at the species level (1 - Bray-Curtis dissimilarity) between samples from different
 291 patients, evaluated against days of hospital overlap (e) or hours of roommate overlap (f) prior to the
 292 earlier sample.
 293 Maximum inStrain popANI achieved by comparing all strains in all samples from two patients, evaluated
 294 against hours of roommate (g) or days of hospital overlap (j) prior to the earlier sample.
 295 In all panels, trend lines are calculated as the best-fit linear regression between the X and Y variables. R
 296 and p-values are the Pearson correlation coefficient and correlation p-value, respectively.

297

298

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

301 *E. coli* and *E. faecium* are common commensal colonizers of human microbiomes^{49–51}. These
 302 species can also be pathogenic and contribute to inflammation, dysbiosis and infection in the host^{52,53}.
 303 The specific strain of these species is key in determining the balance between a healthy and diseased
 304 state in the microbiome. We compared patient-derived *E. coli* and *E. faecium* MAGs with several
 305 reference genomes (Table S6) to identify the closest strains or sequence types.

306 ***Escherichia coli***

307 The 95% identity, or species-level, cluster of *E. coli* MAGs contained 47 genomes from 26
 308 patients (Figure 3). Within this alignment average nucleotide identity (ANI) based tree, we observed two
 309 clades of genomes where MAGs from multiple patients had >99.9% ANI. These clades were investigated
 310 further as they may represent common sequence types. The first clade contained 15 MAGs from 7
 311 patients; these MAGs had >99.9% ANI to pathogenic *E. coli* sequence type (ST) 131 clade C2 genomes,
 312 including EC958⁵⁴ and JJ1886⁵⁵. ST131 is an extraintestinal, pathogenic, multidrug resistant *E. coli* strain
 313 which frequently causes urinary tract infections⁵⁶. *E. coli* ST131 often carries extended-spectrum β -
 314 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 exposure⁵⁷, and there are reports of this pathogen causing urinary tract infections in multiple individuals within a household⁵⁸.

The second clade contained 12 MAGs from 5 patients with >99.8% ANI to the pathogenic ST648 representative IMT16316⁵⁹. ST648 is also an ESBL-producing *E. coli* strain, but it is not as widespread as ST131. Both STs have been isolated from wastewater⁶⁰ and ST648 has been isolated from the gut of humans⁶¹ and other mammals⁶². 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 resistance⁶³ (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.

Escherichia coli

Figure 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.

338 ***Enterococcus faecium***

339 The species-level cluster of *E. faecium* MAGs contained 30 genomes from 20 patients (Figure
340 4a). All MAGs were $\geq 99\%$ identical, suggesting a single ST is present in most patients⁵⁰. These genomes
341 matched closest to *E. faecium* ST117, a well-described vancomycin-resistant strain that frequently
342 causes bloodstream infections⁶⁴. Notably, five other MAGs had $\sim 94\%$ ANI (below the 95% clustering
343 threshold, and therefore not shown in Figure 3a) to the ST117 clade and $>99\%$ ANI to commensal *E.*
344 *faecium* strains, including strains Com15 and Com12⁵⁰. To understand the vancomycin resistance
345 capabilities of these strains, we searched for the seven genes in the *vanA* operon⁶⁵. Only 2/30 samples
346 had the full operon present within the *E. faecium* MAG (Figure S2b). When we looked in the entire
347 metagenome, 22/30 samples had the full operon present, and the *vanA* genes were usually detected on
348 a contig that was not assigned to any MAG. *Van* genes are often carried on mobile genetic elements or
349 plasmids in *E. faecium*, which frequently do not assemble and bin well due to repetitive sequences.

350 Interestingly, no *van* genes are present in sample 11342_02, but the full operon is present in the
351 *E. faecium* MAG in sample 11342_03, collected 14 days later from the same patient. In sample
352 11342_03, the *vanA* operon appears on a contig 8.7 kb long with 143x coverage, much lower than the
353 $\sim 2000x$ coverage of the rest of the *E. faecium* MAG. Mapping reads against this contig revealed
354 scattered coverage in 11342_02, leading us to believe the operon is present, but not at high enough
355 coverage to be assembled in the earlier sample. We attempted to culture vancomycin-resistant
356 *Enterococcus* (VRE) from these samples (see methods), and positively identified *E. faecium* by MALDI
357 coupled to time-of-flight mass spectrometry in samples 11342_02 and 11342_03. Taken together, these
358 results suggest that there were at least two strains of *E. faecium* in the gut microbiota of patient 11342.
359 The low relative coverage of the *vanA* operon indicates that the VRE strain may have been a small
360 fraction of the total *E. faecium* population. Patient 11342 was never prescribed vancomycin (Figure S7a),
361 so the VRE strain may have not had an advantage in this environment.

362 *VanA* genes were also not detected in sample 11349_01, where we observed a nearly identical
363 *E. faecium* strain to patient 11342 after the patients shared a room for 11 days. When we attempted to

364 culture VRE from 11349_01, a bacterium grew poorly on plates containing vancomycin and was
365 identified as *Klebsiella pneumoniae*. Therefore, we believe the *E. faecium* strain in the microbiome of this
366 patient was vancomycin sensitive. If transmission from patient 11342 was responsible for colonization of
367 patient 11349, the vancomycin-sensitive strain may have been transmitted.

368

Figure 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.

382 Nearly identical strains indicative of putative patient-patient *Enterococcus faecium*, 383 but not *Escherichia coli* transmission

384 We used the results from the inStrain comparison to search for nearly identical bacterial strains,
385 which may be indicative of transmission between patients. To determine a threshold for putative
386 transmission, we examined comparisons in several “positive control” datasets where we expect to find
387 identical strains, either as the result of persistence or transmission: time course samples from the same
388 HCT patient, stool samples from mother-infant pairs¹⁸ and samples from fecal microbiota transplantation
389 donors and recipients²³. We often observed 100% popANI in these “positive control” comparisons,
390 indicating that there were no SNPs that could differentiate the strain populations in the two samples
391 (Figure S3, S4). Due to expected noise and errors in sequencing data, we set the lower bound for
392 transmission in our HCT cohort at 99.999% popANI, equivalent to 30 population SNPs in a 3 megabase
393 (Mb) genome. The same threshold was used to identify identical strains by the authors of inStrain⁴⁸.

394 ***Escherichia coli***

395 While *E. coli* genomes in samples collected from the same patient over time were always more
396 similar than the putative transmission threshold, in no case did we observe a pair of samples from
397 different patients with $\geq 99.999\%$ popANI (Table S8). This result suggests that all *E. coli* strains observed
398 are patient-specific, and argues that there are not common strains circulating in the hospital environment
399 or passing between patients. Alternatively, patient-patient transmission or acquisition of common
400 environmental strains is either notably rare or rapid genetic drift after a patient acquires a new strain is
401 reducing popANI levels below the threshold. Deeper metagenomic sequencing or isolation and
402 sequencing of *E. coli* strains may allow us to detect transmission in previously missed cases.

403 ***Enterococcus faecium***

404 We performed the same analysis in *E. faecium* and observed four examples where two patients
405 shared a strain with $\geq 99.999\%$ popANI (Figure 4b,c). In one case, the two patients were roommates and
406 direct transmission appears to be the most likely route. In the other three cases, epidemiological links
407 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 $\geq 5x$ 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”⁶⁶ (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 $\geq 5x$ 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 $\geq 5x$ 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 patients⁶⁷, 113 stool samples from pediatric HCT patients^{3,68,69}, 732 stool samples from hospitalized infants⁷⁰ and 58 vancomycin-resistant *E. faecium* isolates⁷¹. 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 $\geq 99.999\%$, 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 Center⁶⁷ 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).

477

478

479

480

481

Figure 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

482 same room (black bars) or in the hospital (grey bars). Red symbols indicate approximate dates of
 483 bloodstream infection with the specified organism. Hypothesized direction of transmission progresses
 484 from the top to the bottom patient. Fractions of the bar with >99.999% popANI strains in each panel are
 485 indicated with solid colors, and different strains are indicated with hashed colors. All taxa except *E.*
 486 *faecium* are shown at the genus level for clarity.

487 a) Case 1: Putative transmission from patient 11342 to 11349.
 488 b) Case 2: Putative transmission from patient 11575 to 11568.
 489 c) Case 3: Putative transmission from patient 11605 to 11673.
 490

491 Putative transmission of commensal bacteria

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

496 *Hungatella hathewayi*

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

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

516 *Flavonifractor plautii* in later samples. This strain was different from the strain in earlier samples from this
517 patient, as well as all other *Flavonifractor* strains in our sample collection.

518 *H. hathewayi* is known to form spores and is able to persist outside a host for days⁷². Although
519 these patients were only roommates for a single day, 11662 remained in the same room for 4 days after
520 11639 was discharged, increasing the chance that a *H. hathewayi* spore could be transmitted from a
521 surface in the shared room or bathroom. The question remains as to why transmission of *H. hathewayi* is
522 not more common, given it is found at $\geq 1\%$ relative abundance in 31 patients. Perhaps the earlier
523 colonization of the microbiome of 11662 with a different *H. hathewayi* strain was key - the microbiome in
524 this patient was “primed” to receive a new strain of the same species, despite the significant perturbation
525 this patient experienced.

526 Notably, *H. hathewayi* was recently reclassified from *Clostridium hathewayi*⁷³, and was previously
527 shown to induce regulatory T-cells and suppress inflammation⁷⁴. Although the interaction of this microbe
528 in HCT is not known, it may be interesting to investigate further given that the microbe may be
529 transmitted between individuals and may contribute to inflammation suppression that may be relevant in
530 diseases such as graft-vs-host disease. However, *H. hathewayi* may not be entirely beneficial or
531 harmless and has been reported to cause BSI and sepsis in rare cases^{75,76}.

532 ***Akkermansia muciniphila***

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

541 In contrast to *H. hathewayi*, *A. muciniphila* is not known to form spores, which may reduce the
 542 chance of this microbe being transmitted. However, it is an aerotolerant anaerobe that may survive in
 543 oxygen for short periods of time⁷⁷. The microbiome domination of 11742 with *A. muciniphila* and the
 544 relatively long overlap period of nine days in the same room may provide a greater “infectious dose”
 545 (abundance * exposure time) to the recipient patient.

546

Figure 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.

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 $\geq 99.999\%$, 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 hospitals⁷⁸. 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

587 HCT ward. We also verified that no prescriptions were written for probiotics by examining electronic
588 health records. A majority of patients were discharged from the hospital between HCT and acquiring the
589 *L. rhamnosus* strain, which may have provided them with the opportunity to consume a probiotic
590 supplement or dairy product. Contact with a family member or other individual who had the strain in their
591 microbiome could also be responsible for colonization of the HCT patient.

592 If this *L. rhamnosus* strain is a commonly used probiotic supplement or is found in commonly
593 consumed dairy products, it may be found in other gut microbiome sequencing datasets. Comparing
594 MAGs from this cluster against all Genbank genomes revealed a maximum alignment-based ANI of
595 99.95% to *L. rhamnosus* ATCC 8530⁷⁹. Instrain-based comparisons against this reference had a
596 maximum popANI of 99.98%, below the putative transmission threshold. We then searched against all
597 genomes in the Unified Human Gastrointestinal Genome collection⁴¹ and identified two genomes that
598 were nearly identical to the strain found in HCT patients. These genomes were originally from the Human
599 Gastrointestinal Bacteria Culture Collection⁸⁰ (accessions ERR2221226 and ERR1203919, belonging to
600 the same isolate per a personal communication with the authors). Assembled isolate and patient-derived
601 genomes had $\geq 99.99\%$ ANI; inStrain-based SNP comparisons had $\geq 99.999\%$ popANI. This suggests
602 that a *L. rhamnosus* strain that is nearly identical to the genomes in our HCT patients has been isolated
603 from human stool in the past.

604

605

Figure 7: *Lactobacillus* and *Streptococcus* strains are acquired after HCT and identical between many patients.

Population ANI based tree of (a) *Lactobacillus rhamnosus*, (b) *Lactobacillus gasseri*, (c) *Streptococcus thermophilus* strains present in patient samples. Clades containing samples from different patients with $\geq 99.999\%$ 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.

619 Discussion

620 Our investigation using high-resolution metagenomic sequencing attempts to quantify if and when
621 patient-patient microbiome transmission is involved in the spread of pathogenic organisms. We first
622 found that hospitalized HCT patients frequently harbor HAI organisms in their gut microbiome, validating
623 previous studies which used culture-based approaches or 16S rRNA sequencing^{6,26,27}. MAGs created
624 from patient samples had high identity to several globally disseminated and antibiotic resistant sequence
625 types, including *Escherichia coli* ST131 and ST648. Interestingly, whereas ST131 is a well-recognized
626 multi-drug resistant pathogen, ST648 is nearly as prevalent as ST131 in our sample collection, and may
627 thus be an emerging pathogen in this patient population.

628 Despite the high degree of temporal and spatial overlap between patients in our study, we found
629 no association between these factors and the taxonomic similarity of patient microbiomes, and only a
630 weak association between hospital overlap and maximum strain identity. These findings suggest that
631 patient-patient transmission is not driving microbiome composition, but individual strains may still be
632 shared between patients. We did not identify any pairs of patients harboring *E. coli* strains with popANI
633 values above the 99.999% popANI transmission threshold. Taken together with the observation that *E.*
634 *coli* is commonly detected in the patient's microbiome upon admission, this finding argues that patients
635 usually enter the hospital with an "individual-specific" *E. coli* strain and do not frequently transmit it to
636 others. An exclusion principle may be at play, where an *E. coli* niche can only be filled by a single strain
637 and new strains are unlikely to engraft when the niche is already occupied. In contrast to *E. coli*, we
638 observed four pairs of patients with *E. faecium* strains that were more similar than 99.999% popANI. In
639 one case, the two patients spent 11 days sharing a room and bathroom prior to observing the shared
640 strain. Direct links between patients were less clear or non-existent in the other three cases, and
641 transmission through unsampled intermediates or acquisition from an environmental source may have
642 been responsible. We also found evidence for an *E. faecium* strain in a published dataset from HCT
643 patient microbiomes⁶⁷ 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 effect⁸¹ 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 days⁷². 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

670 matches the observation that abundance of probiotic species in the gut often declines after
671 supplementation ends²⁰.

672 The healthy adult gut microbiome is relatively resistant to perturbation and colonization with new
673 strains or species¹⁹. In contrast, mother-to-infant transmission of bacteria and phages is common and
674 well-described^{16–18}. Patients in our study often shared spaces, were exposed to dramatic “niche clearing”
675 therapies and were often immunosuppressed. We frequently observed patients acquiring new organisms
676 into their gut microbiome during their hospital stay, especially following BSI. Still, we found that patient-
677 patient transmission of gut microbes is relatively rare. This suggests that age, rather than perturbation or
678 microbial exposure, may play the largest role in microbiome transmission. There are also several
679 alternative explanations for the relative lack of transmission between patients. First, the adult gut
680 microbiome may remain densely colonized even when dramatically perturbed by antibiotics and
681 chemotherapy, and thus resistant to invasion with new strains. Strains we observed in later samples may
682 have existed at low very levels in earlier samples from the individual, therefore evading detection.
683 Second, it is possible that the microbiome of healthcare workers, hospital visitors, or other staff serve as
684 the source of newly colonizing strains. Third, it is possible that the built environment, equipment used in
685 the care of these individuals, and other environmental sources such as food and personal items harbored
686 the microbes that were later transmitted. As we did not sample these other potential sources, it is difficult
687 to know the extent to which they contributed to the collective reservoir of potentially transmitted
688 organisms. Fourth, genetic drift and adaptive evolution may rapidly act on newly acquired microbes, as
689 has been described for *E. faecium* in the human gut⁸² and *E. coli* in the mouse gut^{83–85}. Rapid drift or
690 evolution would move the genomes of transmitted microbes below the popANI transmission threshold.
691 Finally, transmitted organisms may be killed by antibiotics before they can establish a community within
692 the host and thus be detected with metagenomic sequencing.

693 Our findings have important implications for hospital management and infection prevention.
694 55/149 patients (37%) in our study experienced BSIs, which is comparable to the rate of BSI in other
695 transplant centers⁸⁶. 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 patients⁶⁷. 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 originated⁸⁷. Our work is also entirely based on metagenomic sequencing data, which has its own challenges and sources of bias, including “barcode swapping”⁶⁶, 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 DNA⁸⁸, 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^{89–91} 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

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

733

Methods

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:

8. Only entries for antibiotics, antifungals, antivirals, antibacterials, and *Pneumocystis jirovecii* pneumonia prophylaxis were retained.
9. Medications with a missing start or end date were excluded.
10. Medications with a frequency of “PRN” (*pro re nata*, or *as needed*) or a prescription status of “Canceled” were excluded.
11. Medications with a difference between start and end date of less than one day were excluded.

12. Medications prescribed to be taken by eyes, ears, topical application, or “swish and spit” were excluded.

13. 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.

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 approach⁹² 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).

Sequence data processing

TrimGalore version 0.5.0⁹³ was used to perform quality and adapter trimming with the flags “–clip_R1 15–clip_R2 15–length 60”. SeqKit version 0.9.1⁹⁴ 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-r1188⁹⁵ and only unmapped reads were retained. Quality metrics were verified with FastQC version 0.11.8⁹⁶. Bioinformatics workflows were implemented with Snakemake⁹³.

Short-read classification with Kraken2

We classified all short-read data with a Kraken2⁴² 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 Bracken⁴⁰ 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 vegan⁹⁷ version 2.5-7.

Assembly and binning

Short-read sequencing samples were assembled using SPAdes version 3.14.0³³ using the ‘--meta’ flag. Linked-read sequencing samples were assembled with Megahit version 1.2.9³⁴ to generate seed contigs, which were then assembled with the barcode-aware assembler Athena²⁹. Metagenome-assembled genomes (MAGs) were binned with Metabat2 version 2.15³⁵, Maxbin version 2.2.7³⁶ and CONCOCT version 1.1.0³⁷ and aggregated using DASTool version 1.1.1³⁸. MAG completeness and contamination was evaluated using CheckM version 1.0.13³⁹ and MAG quality was evaluated by the standards set in⁴⁰. 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.

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 2.6.2⁵³ 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 1.3.11⁴⁸ using default parameters. A Snakemake workflow for dRep and inStrain analysis is available at https://github.com/bhattlab/bhattlab_workflows.

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 program⁹⁹. 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 4.0.3¹⁰⁰. Heatmaps were created using pairwise popANI values from inStrain, transformed into a distance matrix, and hierarchically clustered using the 'ward.D2' method.

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 pairs¹⁸ and fecal microbiota transplantation donors and recipients²³

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.

Pairwise MAG comparison

MAGs were aligned with the mummer program using default settings⁹⁹ and filtered for 1-1 alignments. Dotplots were visualized with the “Dot” program¹⁰¹ filtering for non-repetitive alignments ≥ 1 kb.

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)¹⁰² with default parameters. Genes were counted if they met the “strict” or “perfect” threshold from RGI. ARGs were

863 annotated both if they occurred on a contig in the MAG of interest, or anywhere in the metagenomic
864 assembly.

865 Isolation, culture and identification of VRE organisms

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

875

876

877 Acknowledgements

878 We would like to acknowledge the contributions of several members of the Bhatt lab for their assistance
879 in the experimental design, laboratory work and manuscript preparation, including C. Severyn, D.
880 Maghini, S. Vance, A. Natarajan, P. West, A. Behr, M. Chakraborty and S. Zliniti. We would also like to
881 thank B. Wu at Stanford Health Care for providing information on the geographic layout of the HCT ward,
882 K. Chen for providing information relevant to restrictions on patients consuming probiotics, A. Sidow and
883 Z. Weng for their instruction and assistance in preparing linked-read sequencing libraries, K. Siranosian
884 for edits to the manuscript, and S. Madison for providing information on HCT patient isolation practices.
885 We would like to thank the Stanford Clinical Microbiology Lab for assistance in identifying isolated and
886 cultured bacterial strains. This work was supported by NIH P01 CA049605 (PI: R. Negrin; Cell Therapy
887 for Leukemia and Lymphoma), and NIH R01 AI143757, CDC BAA 75D30118C02921, and a Damon
888 Runyon Clinical Investigator Award to A.S.B. B.A.S. is supported by a NIST JIMB training grant.
889 Computing costs were supported via a NIH S10 Shared Instrumentation Grant 1S10OD02014101. This
890 work utilized computing resources provided by the Stanford Genetics Bioinformatics Service Center.
891 Finally, we would like to acknowledge the patients and nurses on the Blood and Marrow Transplantation
892 service for their enthusiastic participation in this project.

893 Author Contributions

894 B.A.S., A.S.B., H.T., N.B., T.A. designed the study. B.A.S. performed laboratory and computational work,
895 conducted the analysis and wrote the manuscript. E.B. performed laboratory work and contributed to
896 writing the manuscript. T.A. performed laboratory work. A.R.R. and T.A. Designed the stool sample
897 collection effort. N.B. assisted with analysis of BSI data. H.T. assisted with computational analysis.
898 A.S.B. oversaw the project and contributed to writing of the manuscript. All authors contributed to editing
899 the manuscript.

900 Competing Interests statement

901 The authors have no competing interests to declare.

902 Data availability

903 Raw sequence data for this manuscript, when not previously published, have been uploaded to NCBI
 904 SRA under project number PRJNA707487, which will be released under publication. MAGs generated in
 905 this study are available in a tar.gz archive (8.4 Gb) from Google Cloud Platform:
 906 https://storage.googleapis.com/gbsec-gcp-lab-bhatt_public/transmission/MAGs.tar.gz. Workflows for data
 907 processing, Kraken2 Classification, MAG binning and inStrain comparisons can be found at
 908 https://github.com/bhattlab/bhattlab_workflows. Days of roommate and hospital overlap between patients
 909 is provided in Tables S9 and S10.

910 Table titles

- 911 1. Aggregated characteristics of patients with samples investigated in this study.
- 912 2. Aggregated statistics of temporal geographic data for all patients on the ward during the study
913 period.
- 914 3. Aggregated statistics of sequencing datasets and metagenome-assembled genomes (MAGs)
915 generated in this study.

916 Supplemental table titles

- 917 1. Clinical metadata of patients with samples investigated in this study.
- 918 2. Sequencing datasets analyzed in this study.
- 919 3. Statistics on MAGs generated in this study.
- 920 4. Kraken2 classification results for all samples at the species level. Species with less than 0.01%
921 abundance have been removed.
- 922 5. Kraken2 classification results for all samples at the genus level. Genera with less than 0.01%
923 abundance have been removed.
- 924 6. Source and publication for each reference genome used in Figures 2 and 3.

7. Pairwise alignment-based ANI for *Escherichia coli* and *Enterococcus faecium* genomes analyzed in Figures 2 and 3. MAGs are identified by a concatenation of the seq_id and Bin columns in Table S3.
8. InStrain results for all comparisons made between samples in this manuscript. Filtered to remove comparisons <99.99% popANI, <0.5 percent_compared and potential barcode swapping results.
9. Matrix of number of days patients overlapped in the hospital.
10. Matrix of number of days patients overlapped as roommates.
11. Antibiotic prescription data for each patient, limited to $\mp$ 100 days relative to HCT.

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”⁶⁶ 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 genome¹⁰³. 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

962 to be assigned to improper samples because of barcode swapping, we estimate the rate in our linked-
963 read data to be 0.1-0.2%.

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

969 We also attempted to measure the degree of barcode swapping in dual-indexed lanes of short
970 read Illumina sequencing. Using the uniquely identifiable p-crAssphage genome as a marker for
971 swapping, we observed roughly 0.5% of sequencing reads swapped between samples on the same lane
972 that shared one index sequence. Samples on the lane that shared no sequencing indices often had p-
973 crAssphage below 1e-5%. Simple relative abundance metrics cannot distinguish between barcode
974 swapping and a true difference in abundance between samples. However, even with the 0.5% rate of
975 swapping, we regularly observed >5x coverage of the p-crAssphage genome in what we believe to be
976 the truly negative samples, and the resulting inStrain comparisons revealed sufficient paired genome
977 coverage and 100% popANI. We never observed identical p-crAssphage genomes between samples
978 from different patients sequenced with unique dual indices or on different lanes.

979 An example of p-crAssphage relative abundance measured with Kraken2 in a single lane of
980 sequencing is shown in the figure below. Here, we believe the sample 11713_98 in blue is a “true
981 positive” for p-crAssphage. While measured abundance in the other samples is likely some combination
982 of true abundance and barcode swapping, the separation between samples that share one or zero index
983 sequences is clear. For short-read sequencing samples, we know which pairs of samples share one of
984 two index sequences and have the possibility of being impacted by swapping. We cannot estimate the
985 impact of barcode swapping like was done for linked-read datasets. We simply eliminated all
986 comparisons where two samples had the possibility of barcode swapping, and all comparisons that could
987 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.

References

1. Young, J.-A. H. et al. Infections after Transplantation of Bone Marrow or Peripheral Blood Stem Cells from Unrelated Donors. *Biol. Blood Marrow Transplant.* 22, 359–370 (2016).
2. See, I. et al. Mucosal Barrier Injury Laboratory-Confirmed Bloodstream Infection: Results from a Field Test of a New National Healthcare Safety Network Definition. *Infect. Control Hosp. Epidemiol.* 34, 769–776 (2013).
3. Kelly, M. S. et al. Gut Colonization Preceding Mucosal Barrier Injury Bloodstream Infection in Pediatric Hematopoietic Stem Cell Transplantation Recipients. *Biol. Blood Marrow Transplant.* 25, 2274–2280 (2019).
4. Tamburini, F. B. et al. Precision identification of diverse bloodstream pathogens in the gut microbiome. *Nat. Med.* 1 (2018) doi:10.1038/s41591-018-0202-8.
5. Zhai, B. et al. High-resolution mycobiota analysis reveals dynamic intestinal translocation preceding invasive candidiasis. *Nat. Med.* 1–6 (2020) doi:10.1038/s41591-019-0709-7.
6. Taur, Y. et al. Intestinal Domination and the Risk of Bacteremia in Patients Undergoing Allogeneic Hematopoietic Stem Cell Transplantation. *Clin. Infect. Dis.* 55, 905–914 (2012).
7. Ubeda, C. et al. Vancomycin-resistant *Enterococcus* domination of intestinal microbiota is enabled by antibiotic treatment in mice and precedes bloodstream invasion in humans. *J. Clin. Invest.* 120, 4332–4341 (2010).
8. Jenq, R. R. et al. Intestinal *Blautia* Is Associated with Reduced Death from Graft-versus-Host Disease. *Biol. Blood Marrow Transplant.* 21, 1373–1383 (2015).
9. Mathewson, N. D. et al. Gut microbiome-derived metabolites modulate intestinal epithelial cell damage and mitigate graft-versus-host disease. *Nat. Immunol.* 17, 505–513 (2016).
10. Peled, J. U. et al. Microbiota as Predictor of Mortality in Allogeneic Hematopoietic-Cell Transplantation. *N. Engl. J. Med.* 382, 822–834 (2020).
11. Shono, Y. et al. Increased GVHD-related mortality with broad-spectrum antibiotic use after allogeneic hematopoietic stem cell transplantation in human patients and mice. *Sci. Transl. Med.* 8, 339ra71 (2016).
12. Taur, Y. et al. The effects of intestinal tract bacterial diversity on mortality following allogeneic hematopoietic stem cell transplantation. *Blood* 124, 1174–1182 (2014).
13. Weber, D. et al. Microbiota Disruption Induced by Early Use of Broad-Spectrum Antibiotics Is an Independent Risk Factor of Outcome after Allogeneic Stem Cell Transplantation. *Biol. Blood Marrow Transplant.* 23, 845–852 (2017).
14. L. Livornese Jr, L. et al. Hospital-acquired Infection with Vancomycin-resistant *Enterococcus faecium* Transmitted by Electronic Thermometers. *Ann. Intern. Med.* (1992).
15. Raven, K. E. et al. Complex Routes of Nosocomial Vancomycin-Resistant *Enterococcus faecium* Transmission Revealed by Genome Sequencing. *Clin. Infect. Dis.* 64, 886–893 (2017).
16. Bäckhed, F. et al. Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host Microbe* 17, 690–703 (2015).

- 1036 17. Siranosian, B. A., Tamburini, F. B., Sherlock, G. & Bhatt, A. S. Acquisition, transmission and strain
1037 diversity of human gut-colonizing crAss-like phages. *Nat. Commun.* 11, 1–11 (2020).
- 1038 18. Yassour, M. et al. Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First
1039 Few Months of Life. *Cell Host Microbe* 24, 146–154.e4 (2018).
- 1040 19. Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity, stability and
1041 resilience of the human gut microbiota. *Nature* 489, 220–230 (2012).
- 1042 20. Suez, J. et al. Post-Antibiotic Gut Mucosal Microbiome Reconstitution Is Impaired by Probiotics and
1043 Improved by Autologous FMT. *Cell* 174, 1406–1423.e16 (2018).
- 1044 21. Zmora, N. et al. Personalized Gut Mucosal Colonization Resistance to Empiric Probiotics Is
1045 Associated with Unique Host and Microbiome Features. *Cell* 174, 1388–1405.e21 (2018).
- 1046 22. Brito, I. L. et al. Transmission of human-associated microbiota along family and social networks.
1047 *Nat. Microbiol.* 1 (2019) doi:10.1038/s41564-019-0409-6.
- 1048 23. Smillie, C. S. et al. Strain Tracking Reveals the Determinants of Bacterial Engraftment in the
1049 Human Gut Following Fecal Microbiota Transplantation. *Cell Host Microbe* 23, 229–240.e5 (2018).
- 1050 24. Watson, A. R. et al. Adaptive ecological processes and metabolic independence drive microbial
1051 colonization and resilience in the human gut. *bioRxiv* 2021.03.02.433653 (2021)
1052 doi:10.1101/2021.03.02.433653.
- 1053 25. Li, S. S. et al. Durable coexistence of donor and recipient strains after fecal microbiota
1054 transplantation. *Science* 352, 586–589 (2016).
- 1055 26. Andermann, T. M. et al. The Microbiome and Hematopoietic Cell Transplantation: Past, Present,
1056 and Future. *Biol. Blood Marrow Transplant.* 24, 1322–1340 (2018).
- 1057 27. Rashidi, A. et al. Pre-transplant recovery of microbiome diversity without recovery of the original
1058 microbiome. *Bone Marrow Transplant.* 1 (2018) doi:10.1038/s41409-018-0414-z.
- 1059 28. Shono, Y. & van den Brink, M. R. M. Gut microbiota injury in allogeneic haematopoietic stem cell
1060 transplantation. *Nat. Rev. Cancer* 18, 283–295 (2018).
- 1061 29. Bishara, A. et al. High-quality genome sequences of uncultured microbes by assembly of read
1062 clouds. *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4266.
- 1063 30. Kang, J. B. et al. Intestinal microbiota domination under extreme selective pressures characterized
1064 by metagenomic read cloud sequencing and assembly. *BMC Bioinformatics* 20, 585 (2019).
- 1065 31. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population
1066 structure and genetic diversity from metagenomes. *Genome Res.* 27, 626–638 (2017).
- 1067 32. Van Rossum, T., Ferretti, P., Maistrenko, O. M. & Bork, P. Diversity within species: interpreting
1068 strains in microbiomes. *Nat. Rev. Microbiol.* 1–16 (2020) doi:10.1038/s41579-020-0368-1.
- 1069 33. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile
1070 metagenomic assembler. *Genome Res.* 27, 824–834 (2017).
- 1071 34. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution
1072 for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinforma. Oxf. Engl.* 31,
1073 1674–1676 (2015).

1074 35. Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome
1075 reconstruction from metagenome assemblies. *PeerJ* 7, e7359 (2019).

1076 36. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover
1077 genomes from multiple metagenomic datasets. *Bioinformatics* 32, 605–607 (2016).

1078 37. Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. *Nat. Methods* 11,
1079 1144–1146 (2014).

1080 38. Sieber, C. M. K. et al. Recovery of genomes from metagenomes via a dereplication, aggregation
1081 and scoring strategy. *Nat. Microbiol.* 1 (2018) doi:10.1038/s41564-018-0171-1.

1082 39. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing
1083 the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*
1084 gr.186072.114 (2015) doi:10.1101/gr.186072.114.

1085 40. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a
1086 metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* 35, 725–731
1087 (2017).

1088 41. Almeida, A. et al. A unified catalog of 204,938 reference genomes from the human gut microbiome.
1089 *Nat. Biotechnol.* 1–10 (2020) doi:10.1038/s41587-020-0603-3.

1090 42. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic
1091 comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME*
1092 *J.* 11, 2864–2868 (2017).

1093 43. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.*
1094 20, 1–13 (2019).

1095 44. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in
1096 metagenomics data. *PeerJ Comput. Sci.* 3, e104 (2017).

1097 45. Centers for Disease Control and Prevention (U.S.). Diseases and Organisms in Healthcare
1098 Settings | HAI | CDC. <https://www.cdc.gov/hai/organisms/organisms.html> (2019).

1099 46. Lloyd-Price, J. et al. Strains, functions and dynamics in the expanded Human Microbiome Project.
1100 *Nature* 550, 61–66 (2017).

1101 47. Abranches, J. et al. Biology of Oral Streptococci. *Microbiol. Spectr.* 6, (2018).

1102 48. Olm, M. R. et al. inStrain profiles population microdiversity from metagenomic data and sensitively
1103 detects shared microbial strains. *Nat. Biotechnol.* (2021) doi:10.1038/s41587-020-00797-0.

1104 49. Dubin, K. & Pamer, E. G. Enterococci and Their Interactions with the Intestinal Microbiome. *Bugs*
1105 *Drugs* 309–330 (2018) doi:10.1128/microbiolspec.BAD-0014-2016.

1106 50. Palmer, K. L. et al. Comparative Genomics of Enterococci: Variation in *Enterococcus faecalis*,
1107 Clade Structure in *E. faecium*, and Defining Characteristics of *E. gallinarum* and *E. casseliflavus*. *mBio* 3,
1108 (2012).

1109 51. Tenaillon, O., Skurnik, D., Picard, B. & Denamur, E. The population genetics of commensal
1110 *Escherichia coli*. *Nat. Rev. Microbiol.* 8, 207–217 (2010).

- 1111 52. Kittana, H. et al. Commensal *Escherichia coli* Strains Can Promote Intestinal Inflammation via
1112 Differential Interleukin-6 Production. *Front. Immunol.* 9, (2018).
- 1113 53. Seishima, J. et al. Gut-derived *Enterococcus faecium* from ulcerative colitis patients promotes
1114 colitis in a genetically susceptible mouse host. *Genome Biol.* 20, 252 (2019).
- 1115 54. Forde, B. M. et al. The Complete Genome Sequence of *Escherichia coli* EC958: A High Quality
1116 Reference Sequence for the Globally Disseminated Multidrug Resistant *E. coli* O25b:H4-ST131 Clone.
1117 *PLOS ONE* 9, e104400 (2014).
- 1118 55. Andersen, P. S. et al. Complete Genome Sequence of the Epidemic and Highly Virulent CTX-M-15-
1119 Producing H30-Rx Subclone of *Escherichia coli* ST131. *Genome Announc.* 1, (2013).
- 1120 56. Gurnee, E. A. et al. Gut Colonization of Healthy Children and Their Mothers With Pathogenic
1121 Ciprofloxacin-Resistant *Escherichia coli*. *J. Infect. Dis.* 212, 1862–1868 (2015).
- 1122 57. Whitmer, G. R., Moorthy, G. & Arshad, M. The pandemic *Escherichia coli* sequence type 131 strain
1123 is acquired even in the absence of antibiotic exposure. *PLOS Pathog.* 15, e1008162 (2019).
- 1124 58. Madigan, T. et al. Extensive Household Outbreak of Urinary Tract Infection and Intestinal
1125 Colonization due to Extended-Spectrum β -Lactamase–Producing *Escherichia coli* Sequence Type 131.
1126 *Clin. Infect. Dis.* 61, e5–e12 (2015).
- 1127 59. Schaufler, K. et al. Genomic and Functional Analysis of Emerging Virulent and Multidrug-Resistant
1128 *Escherichia coli* Lineage Sequence Type 648. *Antimicrob. Agents Chemother.* 63, (2019).
- 1129 60. Paulshus, E. et al. Repeated Isolation of Extended-Spectrum- β -Lactamase-Positive *Escherichia*
1130 *coli* Sequence Types 648 and 131 from Community Wastewater Indicates that Sewage Systems Are
1131 Important Sources of Emerging Clones of Antibiotic-Resistant Bacteria. *Antimicrob. Agents Chemother.*
1132 63, (2019).
- 1133 61. Müller, A., Stephan, R. & Nüesch-Inderbinen, M. Distribution of virulence factors in ESBL-producing
1134 *Escherichia coli* isolated from the environment, livestock, food and humans. *Sci. Total Environ.* 541, 667–
1135 672 (2016).
- 1136 62. Ewers, C. et al. CTX-M-15-D-ST648 *Escherichia coli* from companion animals and horses: another
1137 pandemic clone combining multiresistance and extraintestinal virulence? *J. Antimicrob. Chemother.* 69,
1138 1224–1230 (2014).
- 1139 63. Henderson, T. A., Young, K. D., Denome, S. A. & Elf, P. K. AmpC and AmpH, proteins related to
1140 the class C beta-lactamases, bind penicillin and contribute to the normal morphology of *Escherichia coli*.
1141 *J. Bacteriol.* 179, 6112–6121 (1997).
- 1142 64. Tedim, A. P. et al. Complete Genome Sequences of Isolates of *Enterococcus faecium* Sequence
1143 Type 117, a Globally Disseminated Multidrug-Resistant Clone. *Genome Announc.* 5, (2017).
- 1144 65. Ahmed, M. O. & Baptiste, K. E. Vancomycin-Resistant *Enterococci*: A Review of Antimicrobial
1145 Resistance Mechanisms and Perspectives of Human and Animal Health. *Microb. Drug Resist.* 24, 590–
1146 606 (2017).
- 1147 66. Minimizing Index Hopping. [https://www.illumina.com/techniques/sequencing/ngs-library-](https://www.illumina.com/techniques/sequencing/ngs-library-prep/multiplexing/index-hopping.html)
1148 [prep/multiplexing/index-hopping.html](https://www.illumina.com/techniques/sequencing/ngs-library-prep/multiplexing/index-hopping.html).
- 1149 67. Schluter, J. et al. The gut microbiota is associated with immune cell dynamics in humans. *Nature*
1150 588, 303–307 (2020).

1151 68. D'Amico, F. et al. Gut resistome plasticity in pediatric patients undergoing hematopoietic stem cell
1152 transplantation. *Sci. Rep.* 9, 5649 (2019).

1153 69. Simms-Waldrup, T. R. et al. Antibiotic-Induced Depletion of Anti-inflammatory Clostridia Is
1154 Associated with the Development of Graft-versus-Host Disease in Pediatric Stem Cell Transplantation
1155 Patients. *Biol. Blood Marrow Transplant.* 23, 820–829 (2017).

1156 70. Olm, M. R. et al. Necrotizing enterocolitis is preceded by increased gut bacterial replication,
1157 *Klebsiella*, and fimbriae-encoding bacteria. *Sci. Adv.* 5, eaax5727 (2019).

1158 71. Howden, B. P. et al. Genomic Insights to Control the Emergence of Vancomycin-Resistant
1159 Enterococci. *mBio* 4, (2013).

1160 72. Browne, H. P. et al. Culturing of 'unculturable' human microbiota reveals novel taxa and extensive
1161 sporulation. *Nature* 533, 543–546 (2016).

1162 73. Kaur, S., Yawar, M., Kumar, P. A. & Suresh, K. *Hungatella effluvii* gen. nov., sp. nov., an obligately
1163 anaerobic bacterium isolated from an effluent treatment plant, and reclassification of *Clostridium*
1164 *hathewayi* as *Hungatella hathewayi* gen. nov., comb. nov. *Int. J. Syst. Evol. Microbiol.* 64, 710–718
1165 (2014).

1166 74. Atarashi, K. et al. T reg induction by a rationally selected mixture of Clostridia strains from the
1167 human microbiota. *Nature* 500, 232–236 (2013).

1168 75. Linscott, A. J. et al. Fatal septicemia due to *Clostridium hathewayi* and *Campylobacter hominis*.
1169 *Anaerobe* 11, 97–98 (2005).

1170 76. Woo, P. C. Y. et al. Bacteremia Due to *Clostridium hathewayi* in a Patient with Acute Appendicitis.
1171 *J. Clin. Microbiol.* 42, 5947–5949 (2004).

1172 77. Reunanen, J. et al. *Akkermansia muciniphila* Adheres to Enterocytes and Strengthens the Integrity
1173 of the Epithelial Cell Layer. *Appl. Environ. Microbiol.* 81, 3655–3662 (2015).

1174 78. Yi, S. H., Jernigan, J. A. & McDonald, L. C. Prevalence of probiotic use among inpatients: A
1175 descriptive study of 145 U.S. hospitals. *Am. J. Infect. Control* 44, 548–553 (2016).

1176 79. Pittet, V., Ewen, E., Bushell, B. R. & Ziola, B. Genome Sequence of *Lactobacillus rhamnosus*
1177 ATCC 8530. *J. Bacteriol.* 194, 726–726 (2012).

1178 80. Forster, S. C. et al. A human gut bacterial genome and culture collection for improved
1179 metagenomic analyses. *Nat. Biotechnol.* 37, 186–192 (2019).

1180 81. Fukami, T. Historical Contingency in Community Assembly: Integrating Niches, Species Pools, and
1181 Priority Effects. *Annu. Rev. Ecol. Evol. Syst.* 46, 1–23 (2015).

1182 82. Dubin, K. A. et al. Diversification and Evolution of Vancomycin-Resistant *Enterococcus faecium*
1183 during Intestinal Domination. *Infect. Immun.* 87, (2019).

1184 83. Giraud, A. et al. Dissecting the Genetic Components of Adaptation of *Escherichia coli* to the Mouse
1185 Gut. *PLOS Genet.* 4, e2 (2008).

1186 84. Paepe, M. D. et al. Trade-Off between Bile Resistance and Nutritional Competence Drives
1187 *Escherichia coli* Diversification in the Mouse Gut. *PLOS Genet.* 7, e1002107 (2011).

1188 85. Barroso-Batista, J. et al. The First Steps of Adaptation of Escherichia coli to the Gut Are Dominated
1189 by Soft Sweeps. *PLOS Genet.* 10, e1004182 (2014).

1190 86. Dandoy, C. E., Ardura, M. I., Papanicolaou, G. A. & Auletta, J. J. Bacterial bloodstream infections in
1191 the allogeneic hematopoietic cell transplant patient: new considerations for a persistent nemesis. *Bone*
1192 *Marrow Transplant.* 52, 1091–1106 (2017).

1193 87. Brooks, B. et al. Strain-resolved analysis of hospital rooms and infants reveals overlap between the
1194 human and room microbiome. *Nat. Commun.* 8, 1–7 (2017).

1195 88. Pereira-Marques, J. et al. Impact of Host DNA and Sequencing Depth on the Taxonomic
1196 Resolution of Whole Metagenome Sequencing for Microbiome Analysis. *Front. Microbiol.* 10, (2019).

1197 89. Bertrand, D. et al. Hybrid metagenomic assembly enables high-resolution analysis of resistance
1198 determinants and mobile elements in human microbiomes. *Nat. Biotechnol.* 1 (2019)
1199 doi:10.1038/s41587-019-0191-2.

1200 90. Moss, E. L., Maghini, D. G. & Bhatt, A. S. Complete, closed bacterial genomes from microbiomes
1201 using nanopore sequencing. *Nat. Biotechnol.* 1–7 (2020) doi:10.1038/s41587-020-0422-6.

1202 91. Tsai, Y.-C. et al. Resolving the Complexity of Human Skin Metagenomes Using Single-Molecule
1203 Sequencing. *mBio* 7, (2016).

1204 92. Ramawatar. DNA size selection (>3-4kb) and purification of DNA using an improved homemade
1205 SPRI beads solution. (2018) doi:10.17504/protocols.io.n7hdhj6.

1206 93. Babraham Bioinformatics - Trim Galore!
1207 https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

1208 94. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File
1209 Manipulation. *PLOS ONE* 11, e0163962 (2016).

1210 95. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform.
1211 *Bioinformatics* (2009).

1212 96. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data.
1213 <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.

1214 97. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*
1215 28, 2520–2522 (2012).

1216 98. Oksanen, J. et al. *vegan: Community Ecology Package.* (2020).

1217 99. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome Biol.* 5, R12
1218 (2004).

1219 100. R Development Core Team. *R: A Language and Environment for Statistical Computing.* (2012).

1220 101. Nattestad, M. *MariaNattestad/dot.* (2020).

1221 102. Alcock, B. P. et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic
1222 resistance database. *Nucleic Acids Res.* (2019) doi:10.1093/nar/gkz935.

1223 103. Dutilh, B. E. et al. A highly abundant bacteriophage discovered in the unknown sequences of
1224 human faecal metagenomes. *Nat. Commun.* 5, 4498 (2014).

Table 1

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		
<i>Any BSI</i>	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 2

	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 3

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
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%

Figure S1: 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 S2: Antibiotic resistance genes detected. 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 S3: 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 S4: 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 recipients⁶⁷. The maximum popANI value obtained comparing samples from individuals not related by FMT was 99.998%.

Figure S5: *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* isolates^{3,69-73}. 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 S6: 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 S7: 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 S7: 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.