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

Benjamin A. Siranosian[1], Erin F. Brooks[2], Tessa Andermann[3], Andrew R. Rezvani[4], Niaz Banaei[5,6,7], Hua Tang[1], 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

Abstract

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

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

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

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 mortality[1]. 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 epithelium[2]. 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 infection[3–5]. 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 death[10–13].

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 exposure[14]. 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 environments[15].

74 Transmission of gut bacteria and phages between individuals is known to occur in specific cases,
75 such as from mothers to developing infants[16–18]. 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 group[22], 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 *Clostridiodes difficile* infection. Transmission of gut microbiota is thought to occur by a fecal-oral route, which could happen in the hospital environment by exposure to contaminated surfaces or equipment, sharing a room or bathroom, contaminated hands of healthcare workers or other sources. The perturbed microbiomes of HCT patients, often lacking key species to provide colonization resistance, may be primed to acquire new species from these sources.

Previous studies of the microbiome in HCT patients have often used 16S rRNA sequencing[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

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

Results

Sample characteristics and patient geography

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

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

Metagenomic sequencing, assembly and binning

For an overview of the steps used in the generation and processing of sequence data, see Figure 1a. 328 stool samples from 94 HCT patients were subject to short-read metagenomic sequencing as part

137 of previous projects (for references and SRA IDs of these samples, see Table S2). 96 additional samples
138 from 62 patients were selected for linked-read sequencing to span periods of roommate overlap between
139 patients. Samples were subjected to bead beating-based DNA extraction and bead-based DNA size
140 selection for fragments ≥2 kb (see methods). We prepared linked-read sequencing libraries with the 10X
141 Genomics Chromium platform from 89 samples with sufficient DNA concentration. Samples were
142 sequenced to a median of 116 million (M) (± 37 M SD) read pairs on an Illumina HiSeq4000. In total, 401
143 stool samples from 149 patients were sequenced (Table 3), with a median of 2 and maximum of 13
144 samples per patient (Figure 1b).

145 We processed all existing short-read data and newly generated linked-read data by first trimming
146 and then removing low quality reads, PCR duplicates (short-read data only) and reads that aligned
147 against the human genome (see methods). After quality control, newly sequenced linked-read samples
148 had a median 104 M (± 40 M SD) read pairs, while short-read data had a median 7.6 M (± 4.4 M SD)
149 read pairs. Metagenomic assembly was conducted using metaSPAdes[33] for short-read data, and
150 MEGAHIT[34] followed by Athena[29] for linked-read data. Short-read assemblies had a median N50 of 17.2
151 kb ± 24.8 kb, while linked-read assemblies had a median N50 of 147.6 ± 165.8 kb. We binned
152 metagenome-assembled genomes (MAGs) using Metabat2[35], Maxbin[36] and CONCOCT[37] and
153 aggregated across results from each tool using DASTool[38]. MAG completeness and contamination was
154 evaluated using CheckM[39] and MAG quality was determined by previously established standards[40]. The
155 vast majority of short-read and linked-read MAGs were at least medium quality, and 27% of linked-read
156 MAGs contained the 5S, 16S and 23S rRNA genes and at least 18 tRNAs to be considered high-quality
157 (Figure 1c, Table S3). Linked-read MAGs had higher quality than the 4,644 species-level genomes in the
158 Unified Human Gastrointestinal Genome collection[41], where 573 genomes (12.3%) are high-quality, and
159 only 38 (6.6%) of those came from metagenomes rather than isolates. Sequencing dataset type (short-
160 read vs linked-read) did not have a linear relationship with MAG length (linear regression, $p > 0.9$); the
161 increase in quality was mainly due to the inclusion of ribosomal and transfer RNA genes in the linked-
162 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 dRep[42], yielding 1615 unique genomes representative of the microbial diversity in this sample set.
166

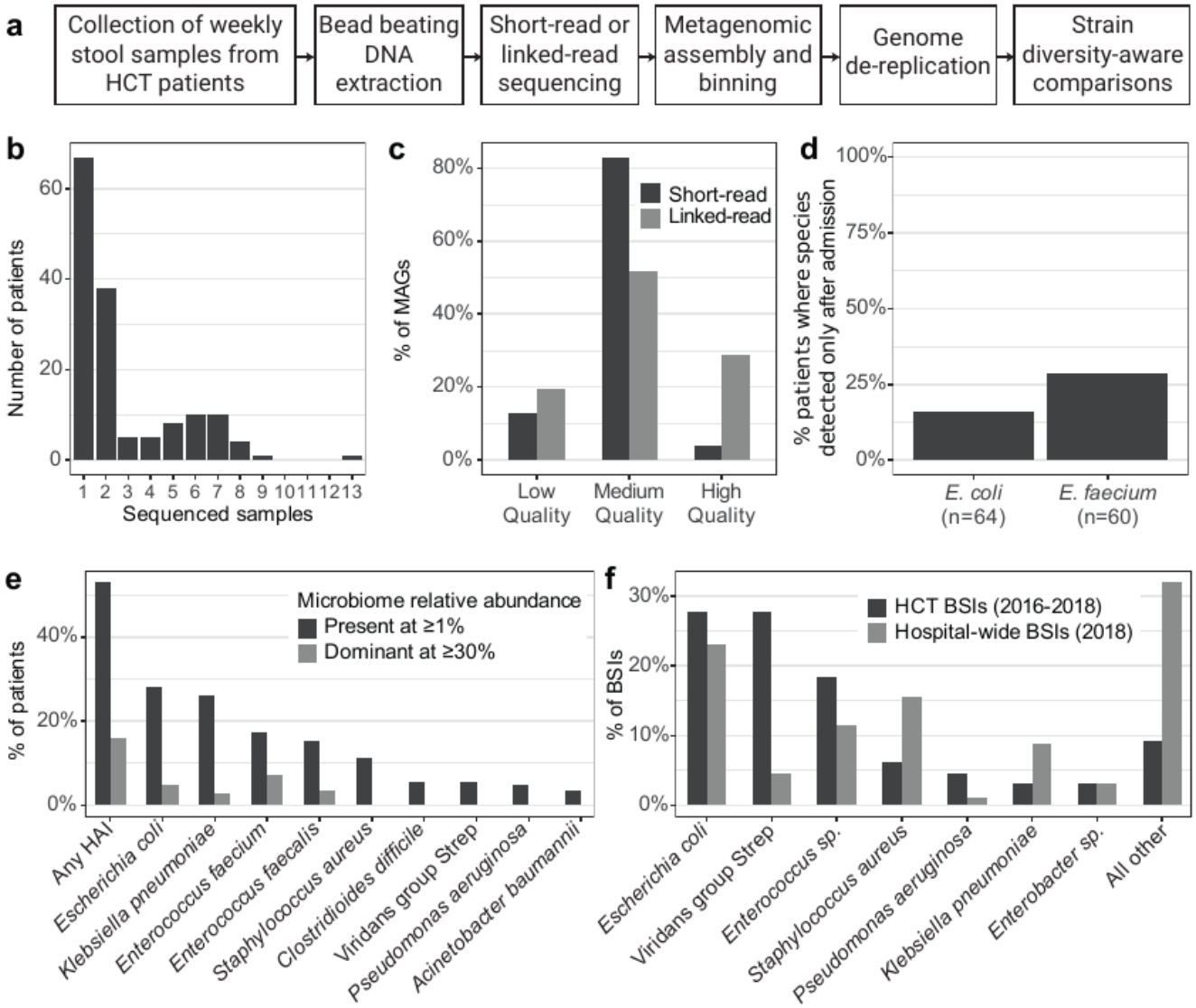


Figure 1: Overview of the methods, data generated, and clinical features of this sample set.

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 level[40], 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 (≥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 Kraken2[43] and abundance estimation with Bracken[44] using a custom database of bacterial, fungal, archaeal and viral genomes in NCBI Genbank (see methods) (Table S4, S5). A median of 33% ± 15% SD reads were classified to the species level with Kraken2 (72% ± 15% SD at the genus level), which was improved to 96% ± 7% SD using Bracken (97% ±8% SD at the genus level). Organisms that cause healthcare-associated Infections (HAI) were identified from the CDC list of pathogens[45]. Here, we report organisms as present if they achieve 1% relative abundance, but acknowledge that many microbes typically exist at lower concentrations, which may be undetectable with metagenomic sequencing.

Many HAI organisms were prevalent in the microbiomes of the studied HCT patients. 152 samples (38%) from 79 patients (53%) had at least one HAI organism identified at 1% relative abundance or above (Figure 1e). *Escherichia coli* was the most common HAI organism (present at ≥1% in 42/149 patients, 28.2%; 0.1%=80/149, 53.7%), followed by *Klebsiella pneumoniae* (39/149 patients, 26.2%; 0.1%=70/149, 47%) and *Enterococcus faecium* (26/149 patients, 17.4%; 0.1%=73/149, 49%). Rates of colonization with HAI organisms were much higher than in stool samples from healthy individuals in the Human Microbiome Project[46] where *E. coli* reaches 1% relative abundance in 2.1% of samples, and *K. pneumoniae* and *E. faecium* are never found at greater than 1% (present at 0.1% in 18.4%, 1.4% and 0.7% of samples, respectively).

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

207 these species are typically much more prevalent in the oral cavity[2,47] compared to the gut microbiome
208 (individual species in the group only reach 1% relative abundance in 8/149 patients, 5%).

209 Detection of *E. coli* and *E. faecium* becomes more common during a patient's 210 hospital stay

211 We investigated the detection of *E. coli* and *E. faecium* in patients with time course samples
212 (82/149 patients, 55%) as a proxy for understanding if these organisms were acquired or became more
213 abundant during the observed hospital stay. Of the 1615 de-replicated MAGs, nine were identified as *E.*
214 *coli* and five were identified as *E. faecium*. We mapped sequencing reads from all samples to these
215 MAGs and evaluated the maximum coverage breadth, the fraction of the reference genome covered with
216 at least one sequencing read. We used a breadth cutoff of 50% to determine "detection" or "absence" in
217 a sample. This threshold is likely specific (it is difficult to achieve 50% breadth by read mis-mapping or
218 homology with a different organism) but not extremely sensitive (it will likely miss very lowly abundant
219 organisms that are truly present).

220 64/82 (78%) patients with time course samples have *E. coli* present at 50% breadth in at least
221 one sample. Of these, 10/64 (16%) have *E. coli* below 50% breadth in the first sample. 60/82 (73%)
222 patients with time course samples have *E. faecium* present at 50% breadth in at least one sample. Of
223 these, 17/60 (28%) have *E. faecium* below 50% breadth in the first sample (Figure 1d). In these patients,
224 *E. coli* or *E. faecium* may have been newly acquired into the gut microbiome during the hospital stay.
225 Alternatively, the organism could have also been present at low abundance and below our limit of
226 detection in the first sample. It is possible that antibiotic use in the weeks after HCT kills off many of
227 natural microbiome colonizers, allowing antibiotic resistant HAI organisms to increase in relative
228 abundance past our limit of detection. Overall, the relative abundance of *E. coli* and *E. faecium* was not
229 significantly different in first samples compared to later samples (Wilcoxon rank-sum test). While more
230 patients have newly detectable E. faecium than E. coli, the difference was not statistically significant
231 ($p=0.09$, binomial likelihood ratio test).

232 Antibiotic use and its effect on HCT patient microbiomes

233 HCT patients are frequently prescribed antibiotic, antiviral and antifungal drugs, especially in the
234 days immediately after transplant. These drugs can have a significant impact on the microbial
235 populations in the gut and contribute to the loss of microbial diversity frequently observed following
236 HCT[6]. Antibiotic use likely impacts the dynamics of bacterial transmission in this patient population, as
237 both natural colonizers, which may provide resistance to newly invading species, and potentially
238 transmitted species can be killed by the drugs. We gathered electronic health record data to understand
239 the characteristics of antibiotic prescription in our patient cohort and its potential impact on the gut
240 microbiome composition.

241 Patients were prescribed a median of five different antibiotics (range 1 - 10) and had a median of
242 90 cumulative antibiotic-days (range 14 - 416). The most common antibiotics prescribed were
243 ciprofloxacin (98% of patients prescribed for at least one day), IV vancomycin (80%), cefepime (66%)
244 and piperacillin-tazobactam (57%). Prescription of most antibiotics peaked in the 14 days following HCT,
245 while administration of antifungal drugs like posaconazole and antiviral drugs like ganciclovir were higher
246 up to 50 days following HCT (Figure 2a). We found that antibiotic usage in the prior seven days before a
247 stool sample was collected was strongly negatively associated with bacterial diversity (linear regression
248 on cumulative antibiotic-days and Shannon diversity, R^2 =0.18, 0.14, p=8.9e-11, 2.6e-8 for species and
249 genus level, respectively). Samples with a single species dominant at 30% relative abundance or higher
250 also typically had higher antibiotic usage in the past seven days (p=0.001, Wilcoxon rank-sum Test)
251 (Figure 2b-d).

252 Patients who share time and space in the hospital do not converge in microbiome 253 composition or frequently share strains

254 To understand the overall impact of temporal and spatial overlap on patient microbiomes, we first
255 studied the taxonomic (Bray-Curtis) similarity between samples from different patients. For each sample
256 pair, we calculated the maximum time the two patients had overlapped in the hospital or as roommates,
257 determined by the earlier sample time. There was not a significant linear association between time of
258 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 inStrain[48] 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.

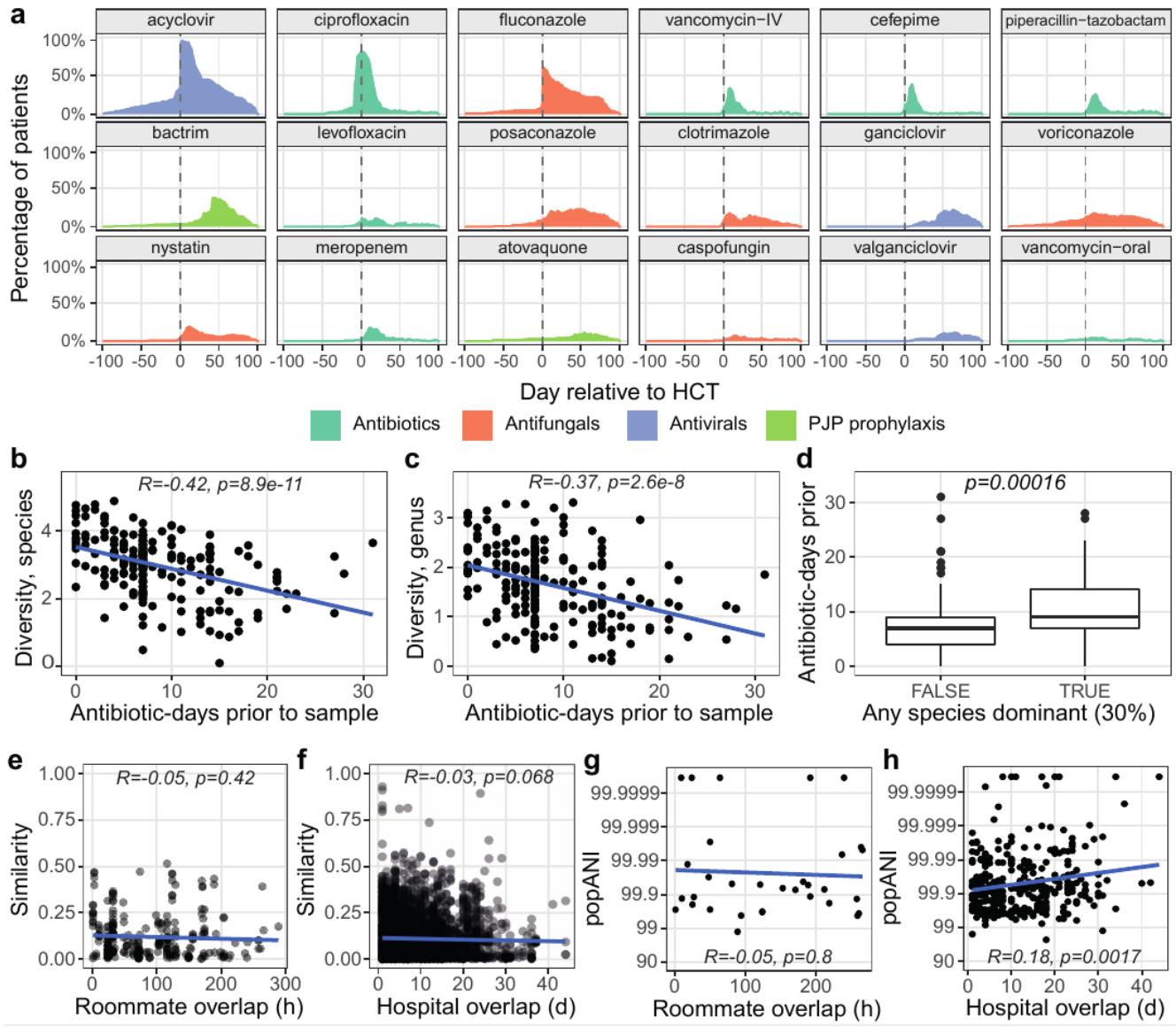


Figure 2: The impact of antibiotic prescription and geographic overlap on patient microbiomes.

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 (≥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 EC958[54] and JJ1886[55]. ST131 is an extraintestinal, pathogenic, multidrug resistant *E. coli* strain
313 which frequently causes urinary tract infections[56]. *E. coli* ST131 often carries extended-spectrum β-
314 lactamase (ESBL) genes which convey a wide range of antibiotic resistance. This sequence type is

315 believed to colonize the intestinal tract even in healthy individuals without antibiotic exposure[57], and there
316 are reports of this pathogen causing urinary tract infections in multiple individuals within a household[58].
317 The second clade contained 12 MAGs from 5 patients with >99.8% ANI to the pathogenic ST648
318 representative IMT16316[59]. ST648 is also an ESBL-producing *E. coli* strain, but it is not as widespread
319 as ST131. Both STs have been isolated from wastewater[60] and ST648 has been isolated from the gut of
320 humans[61] and other mammals[62]. Our finding that *E. coli* ST648 is also prevalent in HCT patient
321 microbiomes suggests that it may become a pathogen of interest in this patient population in the future.
322 To understand the antibiotic resistance capabilities of the *E. coli* strains colonizing these HCT
323 patients, we searched for β-lactamase genes in the *E. coli* MAGs (see methods). The most commonly
324 detected genes were *ampH* and *ampC*, which are part of the core *E. coli* genome and likely do not
325 contribute to antibiotic resistance[63] (Figure S2A). CMY-132 was detected exclusively in MAGs in the
326 ST131 clade, and mutations conveying resistance in *gyrA* were detected in MAGs in multiple clades.
327 CTX-M-type β-lactamases were detected in several samples, but often within the metagenome rather
328 than within the *E. coli* MAG, indicating they may be on plasmids or mobile genetic elements that did not
329 bin with the rest of the *E. coli* genomes.
330

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 ≥99% identical, suggesting a single ST is present in most patients[50]. These genomes
341 matched closest to *E. faecium* ST117, a well-described vancomycin-resistant strain that frequently
342 causes bloodstream infections[64]. Notably, five other MAGs had ~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 Com12[50]. To understand the vancomycin resistance
345 capabilities of these strains, we searched for the seven genes in the *vanA* operon[65]. 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 ~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

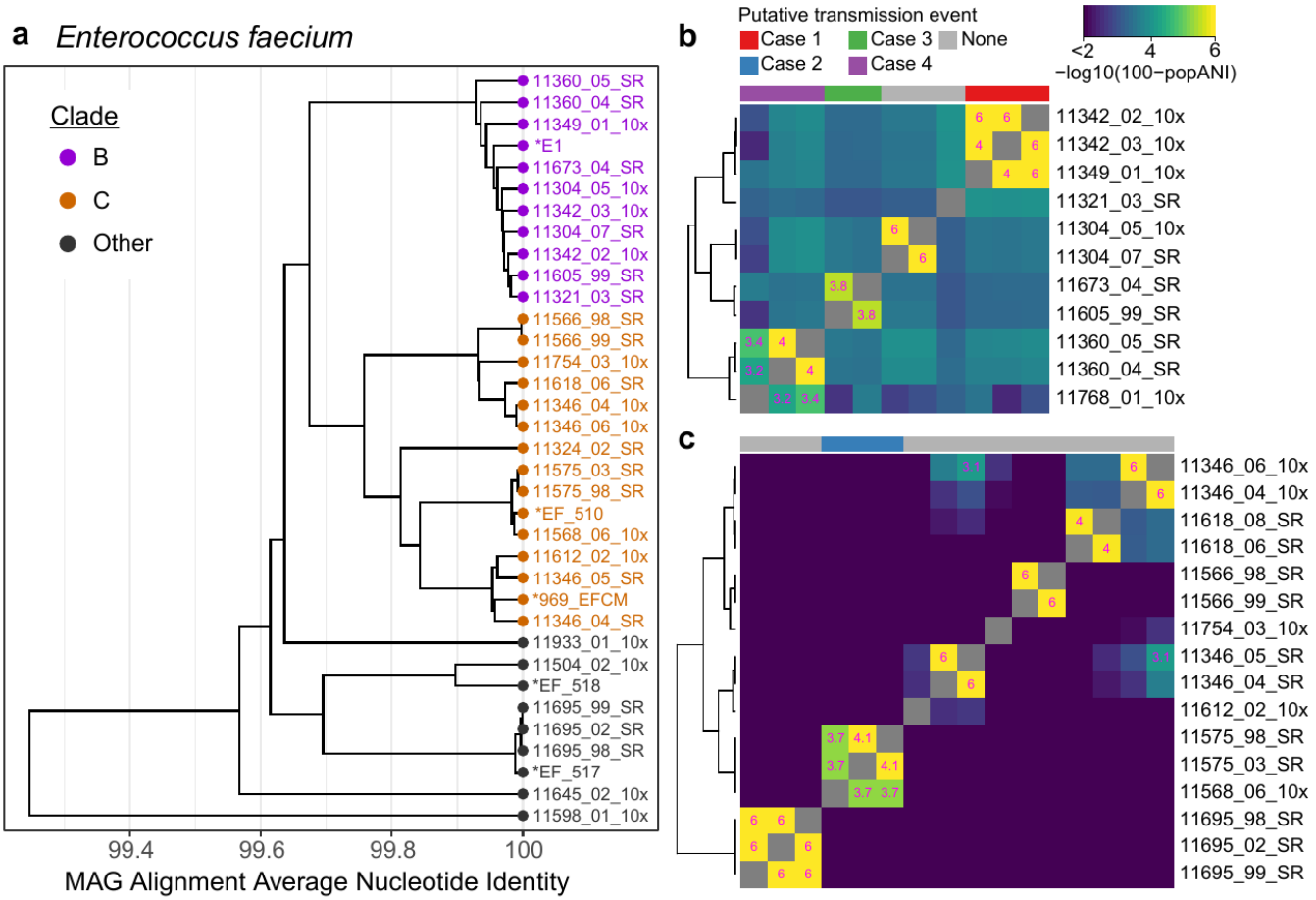


Figure 4: *Enterococcus faecium* strains compared between patients.

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.

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

We used the results from the inStrain comparison to search for nearly identical bacterial strains, which may be indicative of transmission between patients. To determine a threshold for putative transmission, we examined comparisons in several "positive control" datasets where we expect to find identical strains, either as the result of persistence or transmission: time course samples from the same HCT patient, stool samples from mother-infant pairs[18] and samples from fecal microbiota transplantation donors and recipients[23]. We often observed 100% popANI in these "positive control" comparisons, indicating that there were no SNPs that could differentiate the strain populations in the two samples (Figure S3, S4). Due to expected noise and errors in sequencing data, we set the lower bound for transmission in our HCT cohort at 99.999% popANI, equivalent to 30 population SNPs in a 3 megabase (Mb) genome. The same threshold was used to identify identical strains by the authors of inStrain[48].

Escherichia coli

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

Enterococcus faecium

We performed the same analysis in *E. faecium* and observed four examples where two patients shared a strain with ≥99.999% popANI (Figure 4b,c). In one case, the two patients were roommates and direct transmission appears to be the most likely route. In the other three cases, epidemiological links were less clear, suggesting the patients may have acquired a similar strain from the hospital environment

408 or through unsampled intermediates. In the following descriptions, samples are referred to by the day of
409 collection, relative to the first sample from patients in the comparison.

410 **Case 1:** Patients 11342 and 11349 overlapped in the ward for 21 days and were roommates for
411 11 days (Figure 5a). Patient 11342 had a gut microbiome that was dominated by *E. faecium*; the two
412 samples from this patient have 60% and 87% *E. faecium* relative abundance. A single sample from
413 patient 11349 was obtained 14 days after starting to share a room with patient 11342. This sample is
414 dominated by *Klebsiella pneumoniae,* and *E. faecium* is at 0.4% relative abundance. InStrain
415 comparisons between the *E. faecium* strains in 11342 (the presumed "donor") and 11349 (the presumed
416 "recipient") of the strain revealed 0-2 population SNPs (popANI 100% - 99.9999%) with 87% of the
417 reference MAG (2.24 Mb) covered ≥5x in both samples. MAGs from each patient were also structurally
418 concordant (representative dotplots in Figure S6a). These genomes were the most similar out of all *E.*
419 *faecium* genomes compared from different patients. Samples from these patients were extracted in
420 different batches and sequenced on different lanes, minimizing the chance that sample contamination or
421 "barcode swapping"[66] (see Supplemental Note) could be responsible for this result. No other strains were
422 shared between these two patients.

423 **Case 2:** Patients 11575 and 11568 overlapped on the ward for 36 days but were never
424 roommates (Figure 5b). Samples from patient 11575 span 97 days, during which this patient experienced
425 a BSI with *Klebsiella pneumoniae* and treatment with intravenous (IV) vancomycin, ciprofloxacin,
426 meropenem (Figure S7c). Antibiotic treatment likely resulted in a reduction in microbiome diversity and
427 domination by *E. faecium* in samples collected on days 16 and 28. Two samples were collected from
428 patient 11568 on days 28 and 119. The first sample from 11568 was also dominated by *E. faecium,* but
429 strains from the two patients were distinct (99.95% popANI). 91 days later, the second sample from
430 11568 has a lower relative abundance of *E. faecium* but a nearly identical strain to patient 11575. Five
431 population SNPs (99.9997% popANI) were detected with 88% of the reference MAG covered ≥5x in both
432 samples (representative dotplot in Figure S6b). This suggests that the *E. faecium* strain in 11568 was
433 replaced by a different strain with high identity to the strain in 11575. Patient 11568 was discharged from

434 the HCT ward during the period between the two samples. The shared strain may represent an
435 acquisition from a common environmental source or transmission from unobserved patients, rather than
436 a direct transmission event between these two patients. While the *E. faecium* strain was different at the
437 two time points from patient 11568, an *E. faecalis* strain remained identical.

438 **Case 3**: Patients 11605 and 11673 did not overlap in the ward (Figure 5c). Two samples were
439 collected from 11605 on days 0 and 14. This patient experienced a BSI with *E. faecium* and treatment
440 with meropenem (Figure S7e) prior to a sample dominated by the same species on day 14. Patient
441 11673 experienced a BSI with *E. coli* and treatment with vancomycin, meropenem and cefepime (Figure
442 S7f) prior to the single sample we collected from this patient. Comparing *E. faecium* strains between the
443 two patients revealed 2 population SNPs (99.9998% popANI) with 48% of the reference MAG covered
444 ≥5x in both samples (representative dotplot in Figure S6c). Although slightly below the 50% coverage
445 threshold, the high degree of similarity caused us to consider this result. While *E. faecium* strains in the
446 two patients were nearly identical, the samples were collected 161 days apart and the patients had no
447 overlap in the ward. This suggests both patients may have acquired the strain from the hospital
448 environment, through transmission from unsampled patients, or another source such as healthcare
449 workers.

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

455

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

457 The *E. faecium* and *E. coli* strains we observe in our patients may be unique to this patient
458 population and hospital environment. Alternatively, they may be hospital acquired strains that are present
459 in other settings around the globe. We searched through several published datasets to differentiate

460 between these possibilities. Our comparison dataset included metagenomic shotgun sequence data from
461 189 stool samples from adult HCT patients[67], 113 stool samples from pediatric HCT patients[3,68,69], 732
462 stool samples from hospitalized infants[70] and 58 vancomycin-resistant *E. faecium* isolates[71]. Sequence
463 data were downloaded from SRA and processed in the same manner as other short-read data. Each
464 sample was aligned against the *E. faecium* and *E. coli* MAGs used in the inStrain analysis above,
465 profiled for SNPs, and compared against samples collected from our HCT patients. Comparisons within
466 our data and comparisons within individual external datasets frequently achieved popANI values of
467 ≥99.999%, typically from comparisons of samples from the same patient over time. Meanwhile,
468 comparisons between our samples and external samples had lower popANI values (Figure S5).
469 Comparisons of *E. faecium* strains in samples from patient 11346 in our dataset and patient 688
470 in the HCT microbiome dataset collected at Memorial Sloan Kettering Cancer Center[67] demonstrated a
471 maximum of 99.9993% popANI (16 population SNPs detected in 2.3 Mb compared). While direct
472 transmission is likely not involved here, this observation does align with the nearly identical *E. faecium*
473 strains we observed in patients with no geographic or temporal overlap (case 3 and 4) and speaks to the
474 global dissemination of vancomycin-resistant *E. faecium* ST 117. Comparing *E. coli* to external datasets
475 revealed a maximum of 99.996% popANI (200 population SNPs detected in 5.0 Mb compared).
476

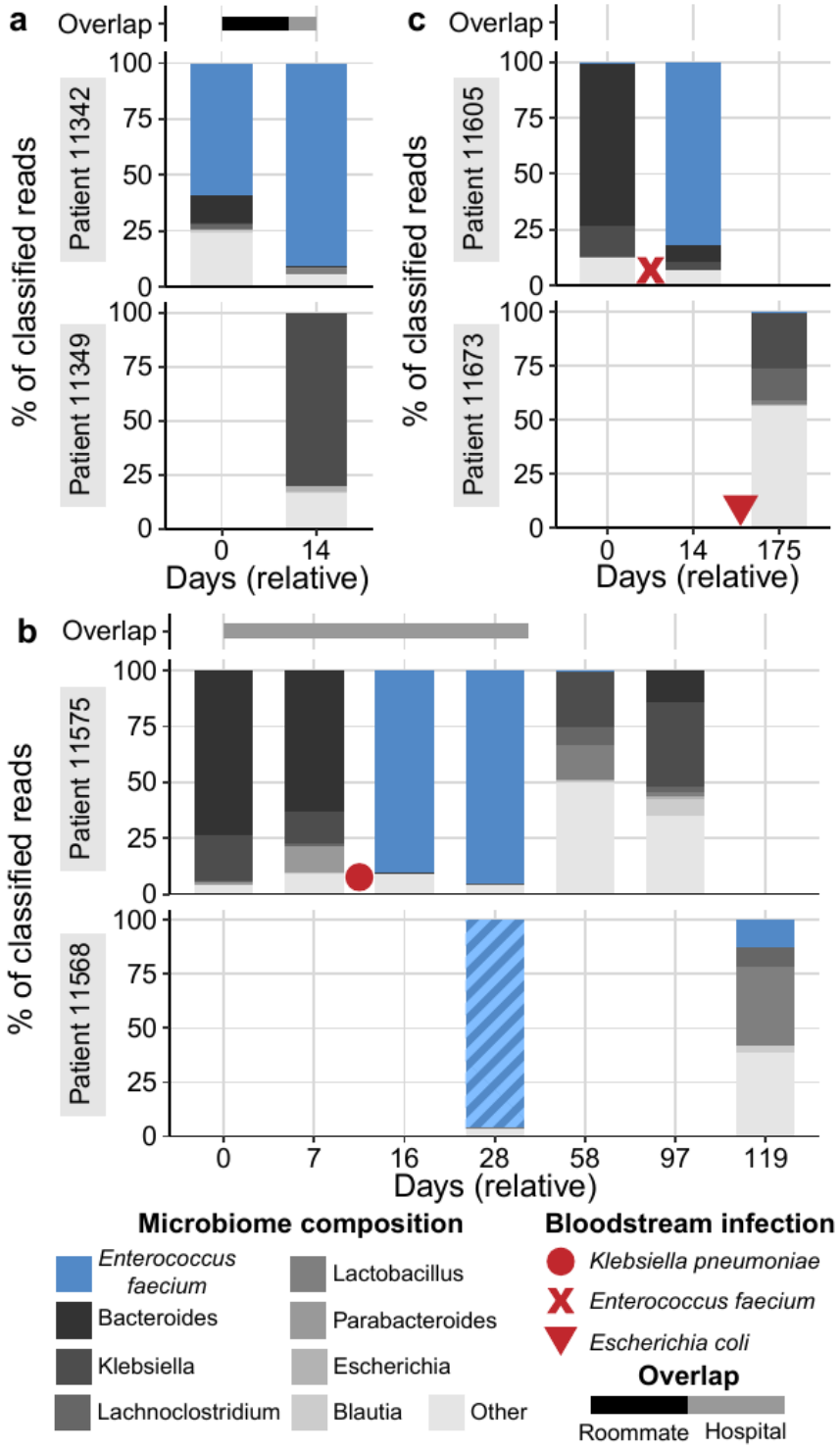


Figure 5: Microbiome composition of patients with putative *Enterococcus faecium* transmission events. Each panel shows the composition of two patients over time. The height of each bar represents the proportion of classified sequence data assigned to each taxon. Samples are labeled relative to the date 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

Putative transmission of commensal bacteria

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

Hungatella hathewayi

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

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

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.

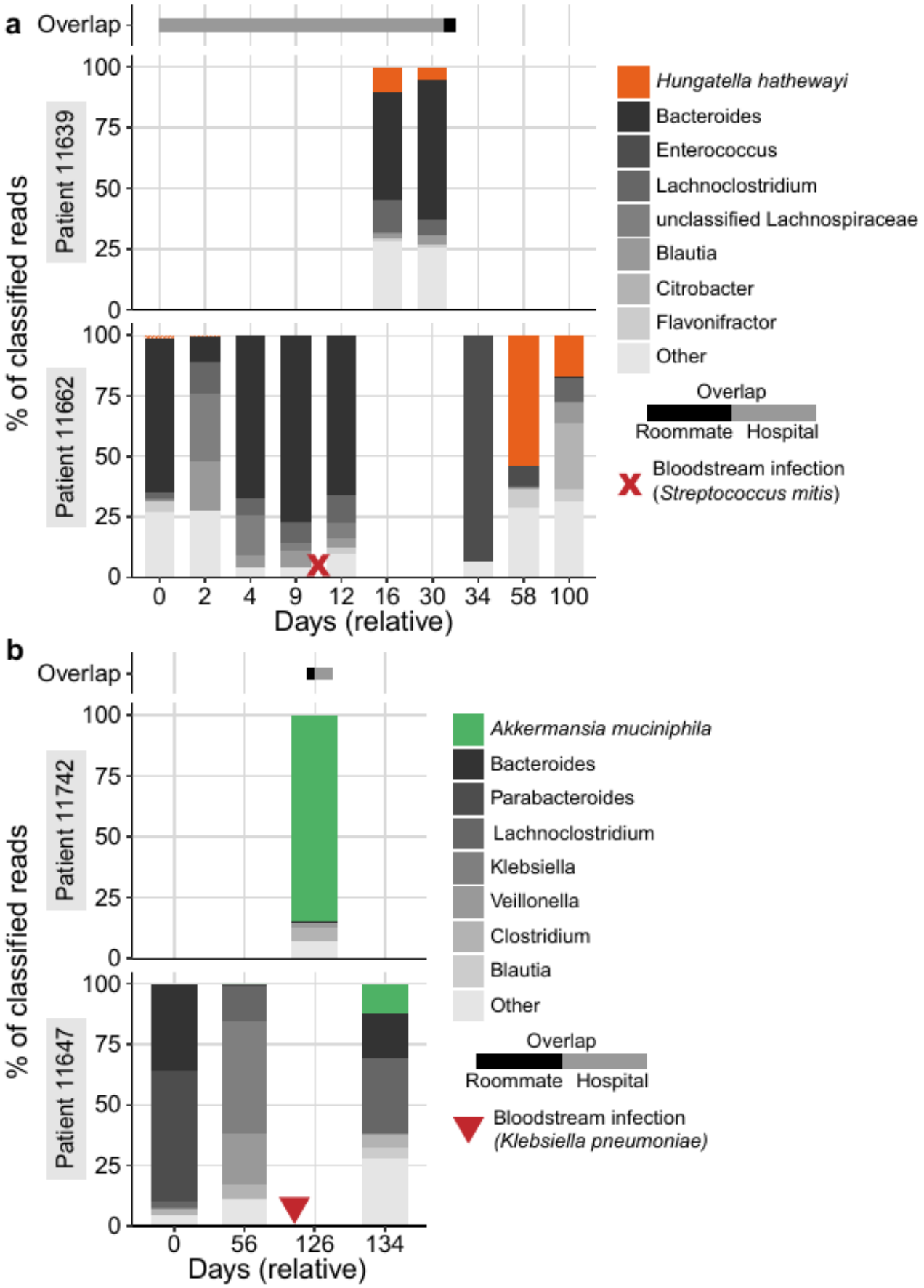
518 *H. hathewayi* is known to form spores and is able to persist outside a host for days[72]. Although
519 these patients were only roommates for a single day, 11662 remained in the same room for 4 days after
520 11639 was discharged, increasing the chance that a *H. hathewayi* spore could be transmitted 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 ≥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 inflammation[74]. 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 time[77]. 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


a
Overlap
Patient 11639
Patient 11662
% of classified reads
100
75
50
25
0
0
2
4
9
12
16
30
34
58
100
Days (relative)
Hungatella hathewayi
Bacteroides
Enterococcus
Lachnoclostridium
unclassified Lachnospiraceae
Blautia
Citrobacter
Flavonifractor
Other
Overlap
Roommate
Hospital
Bloodstream infection
(Streptococcus mitis)
b
Overlap
Patient 11742
Patient 11647
% of classified reads
100
75
50
25
0
0
56
126
134
Days (relative)
Akkermansia muciniphila
Bacteroides
Parabacteroides
Lachnoclostridium
Klebsiella
Veillonella
Clostridium
Blautia
Other
Overlap
Roommate
Hospital
Bloodstream infection
(Klebsiella pneumoniae)

547

548 **Figure 6:** Microbiome composition of patients with putative *Hungatella hathewayi* or *Akkermansia*
549 *muciniphila* transmission events. Each panel shows the composition of two patients over time. The height
550 of each bar represents the proportion of classified sequence data assigned to each taxon. Samples are
551 labeled relative to the date of the first sample in each set. Bars above each plot represent the
552 approximate time patients spent in the same room (black bars) or in the hospital (grey bars). Red
553 symbols indicate approximate dates of bloodstream infection with the specified organism. Hypothesized
554 direction of transmission progresses from the top to the bottom patient. Fractions of the bar with
555 >99.999% popANI strains in each panel are indicated with solid colors, and different strains are indicated
556 with hashed colors. All taxa except *H. hathewayi* or *A. muciniphila* are shown at the genus level for
557 clarity.
558 a) Putative case of *H. hathewayi* transmission from 11639 to 11662.
559 b) Putative case of *A. muciniphila* transmission from 11742 and 11647.
560

561 ## Widespread strain sharing of commercially available probiotic organisms

562 Several organisms were found with identical or nearly identical genomes across multiple patient
563 microbiomes without clear epidemiological links. The largest clade was found for *Lactobacillus*
564 *rhamnosus*, which included 11 samples collected from eight patients over a span of 2.5 years (Figure
565 7a,d). All 11 samples in this clade had pairwise popANI of ≥99.999%, and in a subset of eight samples
566 from seven patients, all pairs were identical from a popANI perspective (100%). Of the eight patients in
567 this clade, only two pairs were roommates or overlapped in the hospital (patients 11537/11547 and
568 11647/11662, roommates for three days and one day, hospital overlap for 20 and 44 days, respectively).
569 All 11 samples in this clade were collected after HCT (median time from HCT to first sample 43 days,
570 range 12-93 days), and *L. rhamnosus* always had <0.1% relative abundance in pre-HCT samples, when
571 present. Five of eight patients were discharged from the hospital after HCT and prior to acquiring *L.*
572 *rhamnosus,* which we observed in a sample collected during a subsequent admission. We also observe
573 *L. rhamnosus* falling below 0.1% relative abundance in a subsequent sample in five patients, suggesting
574 that this strain may be a transient colonizer of the microbiome (Figure 7E). We evaluated antibiotic
575 prescriptions in these patients and found that acquisition and loss of *L. rhamnosus* typically occurs
576 independently of antibiotic use. For example, in patient 11537, *L. rhamnosus* is first detected and
577 expands to 23.4% relative abundance while the patient is prescribed ciprofloxacin and cefepime. *L.*
578 *rhamnosus* declines in relative abundance after the antibiotic prescription ends (Figure S9). Similar
579 clades of high-identity genomes from different patients were found for *Lactobacillus gasseri* (Figure 7b)
580 and *Streptococcus thermophilus* (Figure 7c).

581 Given that we did not observe hospital or roommate overlap between most patients in the *L.*
582 *rhamnosus* cluster, the most likely explanation is that patients acquired this strain from a common
583 source. *L. rhamnosus* is a component of several commercially available probiotic supplements, is
584 present in certain live active-culture foods such as yogurt, and is among the most commonly prescribed
585 probiotic species in US hospitals[78]. However, HCT recipients were not allowed to take probiotics or
586 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 8530[79]. 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 collection[41] 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 Collection[80] (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 ≥99.99% ANI; inStrain-based SNP comparisons had ≥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

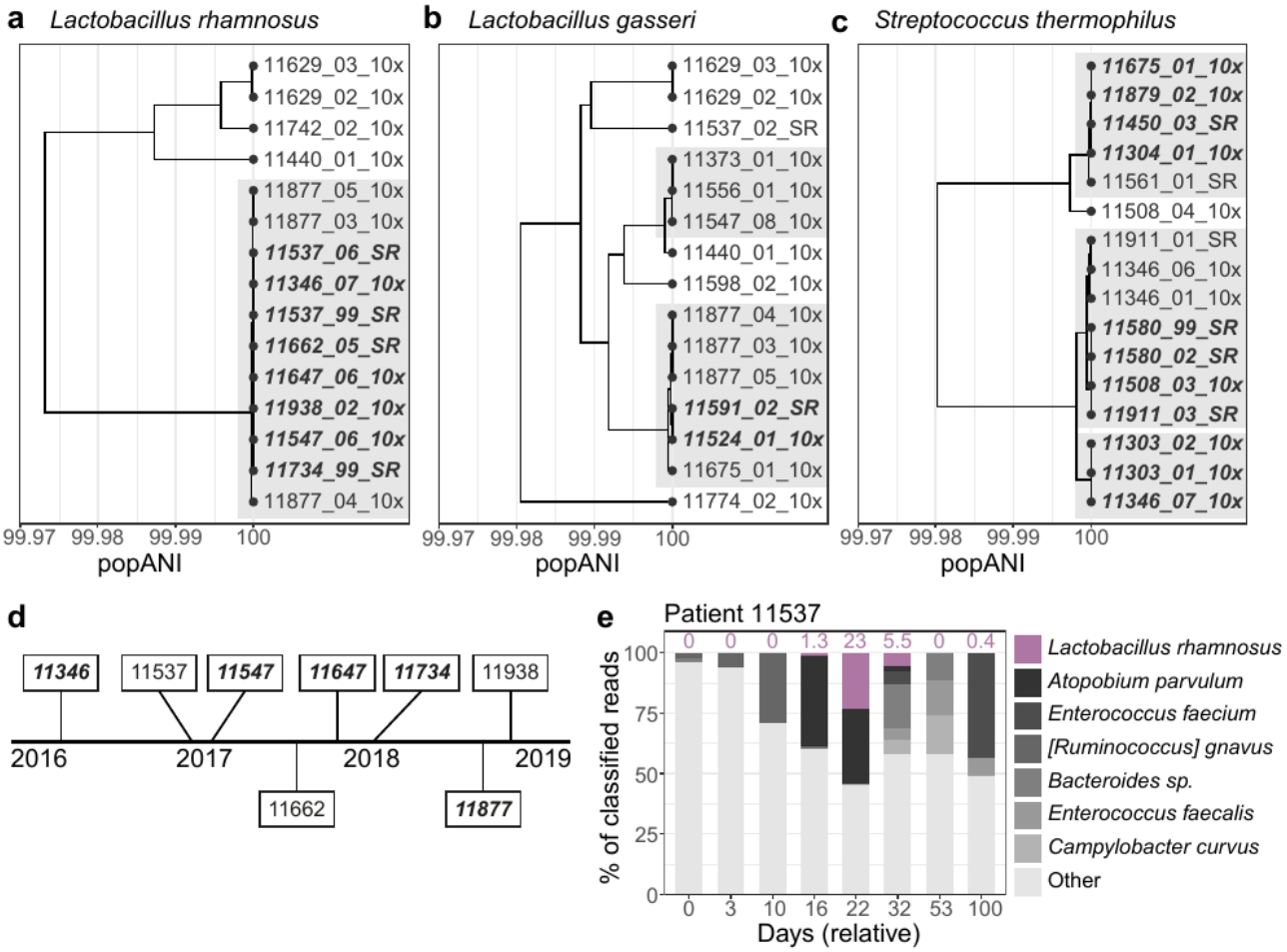


Figure 7: *Lactobacillus* and *Streptococcus* strains are acquired after HCT and identical between many patients.

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 ≥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 microbiomes[67] that had above 99.999% popANI to a strain in our sample collection. While this

644 finding is likely not the result of patient-patient transmission, it does indicate that very similar strains may
645 exist within patients in different geographic locations.

646 We then expanded the transmission analysis to examine all species that were present in the
647 microbiome of multiple patients. Identical *Hungatella hathewayi* strains were found in two patients who
648 were in the hospital together for 34 days and roommates for a single day. Earlier samples from patient
649 11662 had a significantly different *H. hathewayi* strain than the strain present in later samples. It is
650 possible that the earlier colonization with the same species exhibited a priority effect[81] and primed this
651 individual to be re-colonized. In another set of patients who overlapped in the hospital for 11 days and
652 were roommates for nine days, we identified identical *Akkermansia muciniphila* strains. In both
653 examples, the likely "recipient" patient experienced BSI prior to the putative transmission event. The
654 subsequent antibiotic treatment initiated for the treatment of BSI resulted in vast microbiome modification
655 and simplification, which may have opened a niche for the new organism to engraft into. Both *H.*
656 *hathewayi* and *A. muciniphila* can survive outside the host for periods of time, but *H. hathewayi* can form
657 spores that enable it to live in aerobic conditions for days[72]. In the case of *H. hathewayi* transmission,
658 patient 11662 remained in the same room with a shared bathroom after the single day of overlap with
659 patient 11639. Spores surviving on surfaces may be responsible for transmission, given the relatively
660 short period of overlap. In these cases, patient-patient transmission may help in the recovery of
661 microbiome diversity following BSI and may play a role in ameliorating post-HCT inflammatory
662 processes, such as acute graft-vs-host disease.

663 Finally, we observed identical probiotic species in multiple patients without clear geographic or
664 temporal links, including *Lactobacillus rhamnosus*, *Lactobacillus gasseri*, and *Streptococcus*
665 *thermophilus*. Acquisition from a commercially available probiotic or live-active culture food appears to be
666 the most likely explanation. While patients hospitalized for HCT were not allowed to consume probiotics
667 or high-bacterial dairy foods, a majority of patients were discharged after HCT and prior to a subsequent
668 admission, upon which *L. rhamnosus* was detected. Many patients lost the strain in later samples,
669 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 ends[20].
672 The healthy adult gut microbiome is relatively resistant to perturbation and colonization with new
673 strains or species[19]. 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 gut[82] 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 centers[86]. Our findings suggest that microbiome transmission does not play a large role in

696 spreading infections among HCT patients, and that established contact precautions and procedures for
697 patient isolation were working as intended. Recently, the HCT ward at our hospital moved to a new
698 location with exclusively single rooms, which may further reduce the opportunity for transmission.

699 Our analysis of transmission of microbes between HCT patients does have several limitations.
700 We analyzed hundreds of samples collected over many years, but most sampling was done on a weekly
701 basis. We did not explicitly collect samples on the day of admission or discharge. Our sample collection
702 also ignores previous hospital stays, either in a different ward in our hospital, or other hospitals entirely,
703 that may be responsible for the acquisition of HAI organisms. Many patients in our study had one or two
704 sequenced samples, limiting our inference about microbiome changes over time. By contrast, the second
705 largest study using shotgun metagenomics to study the microbiome of HCT patients had much more
706 dense sampling, analyzing 395 samples from 49 patients[67]. We also did not perform any sampling of the
707 hospital environment, healthcare workers or visitors, which would allow us to track transmission patterns
708 in detail and more conclusively state where newly acquired microbes originated[87]. Our work is also
709 entirely based on metagenomic sequencing data, which has its own challenges and sources of bias,
710 including "barcode swapping"[66], which could contribute to false positive transmission findings. To
711 address this, we measured the impact of barcode swapping in linked-read data, and eliminated linked-
712 read and short-read comparisons where a finding of identical strains could be the result of barcode
713 swapping (see supplemental note). Additionally, metagenomic sequencing may fail to detect lowly-
714 abundant colonizers, especially when samples are contaminated with host DNA[88], which is often the
715 case in stool samples from HCT patients. These challenges undoubtedly affect our sensitivity and
716 specificity in measuring acquisition and transmission of both pathogenic and commensal microbes. While
717 our comparison methods were sensitive to strain populations in the gut microbiome, we did not attempt
718 to phase strain haplotypes. Haplotype phasing with long-read sequencing technology like Nanopore or
719 PacBio[89–91] could help us determine whether sets of SNPs occurred in the same or different strains.

720 Our study leaves several questions unanswered that we hope future work on microbiome
721 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.

760 12. Medications prescribed to be taken by eyes, ears, topical application, or “swish and spit” were
761 excluded.
762 13. Medication prescriptions occurring outside the window of HCT date +/- 100 days were excluded
763 for the aggregated analysis.
764 We do acknowledge the challenge of working with electronic health record data, and recognize that there
765 is a disconnect between medications prescribed and medications consumed by a patient.
766

767 DNA Extraction, library preparation and sequencing

768 DNA was extracted from stool samples using a mechanical bead-beating approach with the Mini-
769 Beadbeater-16 (BioSpec Products) and 1-mm diameter zirconia/silica beads (BioSpec Products)
770 followed by the QIAamp Fast DNA Stool Mini Kit (Qiagen) according to manufacturer's instructions.
771 Bead-beating consisted of 7 rounds of alternating 30 s bead-beating bursts followed by 30 s of cooling
772 on ice. For samples subjected to linked-read sequencing, DNA fragments less than approximately 2 kb
773 were eliminated with a SPRI bead purification approach[92] using a custom buffer with minor modifications:
774 beads were added at 0.9×, and eluted DNA was resuspended in 50 µl of water. DNA concentration was
775 quantified using a Qubit fluorometer (Thermo Fisher Scientific). DNA fragment length distributions were
776 quantified using a TapeStation 4200 (Agilent Technologies).
777 Short-read sequencing libraries were prepared with either the Nextera Flex or Nextera XT kit
778 (Illumina) according to manufacturer’s instructions. Linked-read sequencing libraries were prepared on
779 the 10X Genomics Chromium platform (10X Genomics). Linked-read libraries have a single sample
780 index, and were pooled to minimize the possibility of barcode swapping between samples from patients
781 who were roommates (see supplemental note). Libraries were sequenced on an Illumina HiSeq 4000
782 (Illumina).

783 Sequence data processing

784 TrimGalore version 0.5.0[93] was used to perform quality and adapter trimming with the flags “–
785 clip_R1 15–clip_R2 15–length 60”. SeqKit version 0.9.1[94] was used to remove duplicates in short-read

786 data with the command “seqkit rmdup–by-seq”. Due to excessive processing time, this step was skipped
787 for linked-read data. Reads were mapped against the GRCh38 assembly of the human genome using
788 BWA version 0.7.17-r1188[95] and only unmapped reads were retained. Quality metrics were verified with
789 FastQC version 0.11.8[96]. Bioinformatics workflows were implemented with Snakemake[93].

790 Short-read classification with Kraken2

791 We classified all short-read data with a Kraken2[42] database containing all bacteria, viral and
792 fungal genomes in NCBI GenBank assembled to complete genome, chromosome or scaffold quality as
793 of January 2020. Human and mouse reference genomes were also included in the database. A
794 Bracken[40] database was also built with a read length of 150 and k-mer length of 35. Classification results
795 were processed into matrices and taxonomic barplots with the workflow available at
796 https://github.com/bhattlab/kraken2_classification. Bray-Curtis distances were calculated with the R
797 package vegan[97] version 2.5-7.

798 Assembly and binning

799 Short-read sequencing samples were assembled using SPAdes version 3.14.0[33] using the ‘--
800 meta’ flag. Linked-read sequencing samples were assembled with Megahit version 1.2.9[34] to generate
801 seed contigs, which were then assembled with the barcode-aware assembler Athena[29]. Metagenome-
802 assembled genomes (MAGs) were binned with Metabat2 version 2.15[35], Maxbin version 2.2.7[36] and
803 CONCOCT version 1.1.0[37] and aggregated using DASTool version 1.1.1[38]. MAG completeness and
804 contamination was evaluated using CheckM version 1.0.13[39] and MAG quality was evaluated by the
805 standards set in[40]. All assembled contigs were classified with Kraken2 as described above. To generate
806 bin identifications, contig classifications were pooled such that contigs making up at least two thirds the
807 length of the bin were classified as a particular species. If a classification could not be assigned at the
808 species level, the process was repeated at the genus level, and so on.

809 Genome de-replication and SNP profiling

810 MAGs were filtered to have minimum completeness 50% and maximum contamination 15% as
811 measured by CheckM, then were de-replicated with dRep version 2.6.2[53] with default parameters except

812 the primary clustering threshold set to 0.95. In further steps, a single de-replicated genome will be
813 referred to as a cluster. Reads from all samples were mapped against the de-replicated set of genomes
814 with BWA. Clusters that had greater than 1x average coverage in at least two samples were retained for
815 further analysis. Individual bam files were extracted for each sample-cluster pair with at least 1x
816 coverage. Bam files were randomly subsetted to a maximum of 2 million reads for computational
817 efficiency. Alignments were profiled and then compared across samples with inStrain version 1.3.11[48]
818 using default parameters. A Snakemake workflow for dRep and inStrain analysis is available at
819 https://github.com/bhattlab/bhattlab_workflows.

820 Building phylogenetic trees

821 MAG Average Nucleotide Identity (ANI) trees (Figure 2 and 3) were created using the pairwise
822 alignment values from dRep, which uses the MUMmer program[99]. MAGs were filtered to be at least 75%
823 the length of the mean length of reference genomes used in the tree. Reference genomes were selected
824 by searching literature for collections of well-described isolates with genomes available. References that
825 were not relevant and clustered in isolated sections of the tree were removed. Pairwise ANI values were
826 transformed into a distance matrix and clustered using the 'hclust' function with the 'average' method in
827 R version 4.0.3[100]. Heatmaps were created using pairwise popANI values from inStrain, transformed into
828 a distance matrix, and hierarchically clustered using the 'ward.D2' method.

829 Determining transmission thresholds

830 To determine the ANI threshold to call a comparison a "putative transmission event" we evaluated
831 the distributions of ANI values for within- and between-patient comparisons for different species (Figure
832 S3). We often detected zero population SNPs in time course samples from the same patient, including *E.*
833 *faecium* in a pair of samples collected from the same patient 323 days apart. Meanwhile, between-
834 patient comparisons typically had lower ANI values. To verify that transmission events would also result
835 in population ANI values near 100%, we examined external datasets where transmission of bacteria in
836 the microbiome is known to occur as a "positive control". We gathered sequencing data from stool
837 samples of matched mother-infant paris[18] and fecal microbiota transplantation donors and recipients[23]

838 and processed them with the same methods. In these datasets, we regularly observed genomes with
839 100% popANI between matched individuals, and did not find cases of 100% popANI between unmatched
840 individuals (Figure S4). In the ideal cases, we expect transmission of bacteria between the microbiomes
841 of HCT patients to result in genomes with 100% popANI. However, the measured genomes may not
842 reach this level of identity, due to mutations or genetic drift since the transmission event, sequencing
843 errors, or other factors. Therefore, we set the transmission threshold at 99.999% popANI, equivalent to
844 30 population SNPs in a 3 megabase (Mb) genome. Although this threshold is stringent, we recognize
845 that it may allow for false positives where two closely related strains exist in different patients solely by
846 chance.

847 Despite our efforts to minimize the impact of barcode swapping on detecting transmission (see
848 Supplemental Note), we still identified many comparisons with >99.999% popANI that we believe to be
849 false positives. These were filtered out based on the following criteria: short read samples from different
850 patients that were sequenced on the same lane and shared one index sequence or linked read samples
851 that were sequenced on the same lane and reads mapping to the organism shared >40% of barcodes.
852 We also removed pairs that could be affected by "secondary" swapping, where the two samples were not
853 directly affected, but an interaction between other samples from the two patients could cause false
854 positives. In total, we removed 31 comparisons from the final table with >99.999% popANI.

855 Pairwise MAG comparison

856 MAGs were aligned with the mummer program using default settings[99] and filtered for 1-1
857 alignments. Dotplots were visualized with the "Dot" program[101] filtering for non-repetitive alignments ≥1
858 kb.

859 Antibiotic resistance gene detection

860 Antibiotic resistance genes (ARGs) were profiled in contigs from all samples using Resistance
861 Gene Identifier (RGI) and the Comprehensive Antibiotic Resistance Database (CARD)[102] with default
862 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

Acknowledgements

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

Author Contributions

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

Competing Interests statement

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/gbsc-gcp-lab-bhatt_public/transmission/MAGs.tar.gz. Workflows for data
907 processing, Krarken2 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.

925 7. Pairwise alignment-based ANI for Escherichia coli and Enterococcus faecium genomes analyzed
926 in Figures 2 and 3. MAGs are identified by a concatenation of the seq_id and Bin columns in
927 Table S3.
928 8. InStrain results for all comparisons made between samples in this manuscript. Filtered to remove
929 comparisons <99.99% popANI, <0.5 percent_compared and potential barcode swapping results.
930 9. Matrix of number of days patients overlapped in the hospital.
931 10. Matrix of number of days patients overlapped as roommates.
932 11. Antibiotic prescription data for each patient, limited to ∓ 100 days relative to HCT.
933
934
935
936

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 genome[103]. 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. EC958[53] We set a threshold of 40% overlap of barcode sets to call a comparison “swapped” and remove it from analysis. By counting reads believed

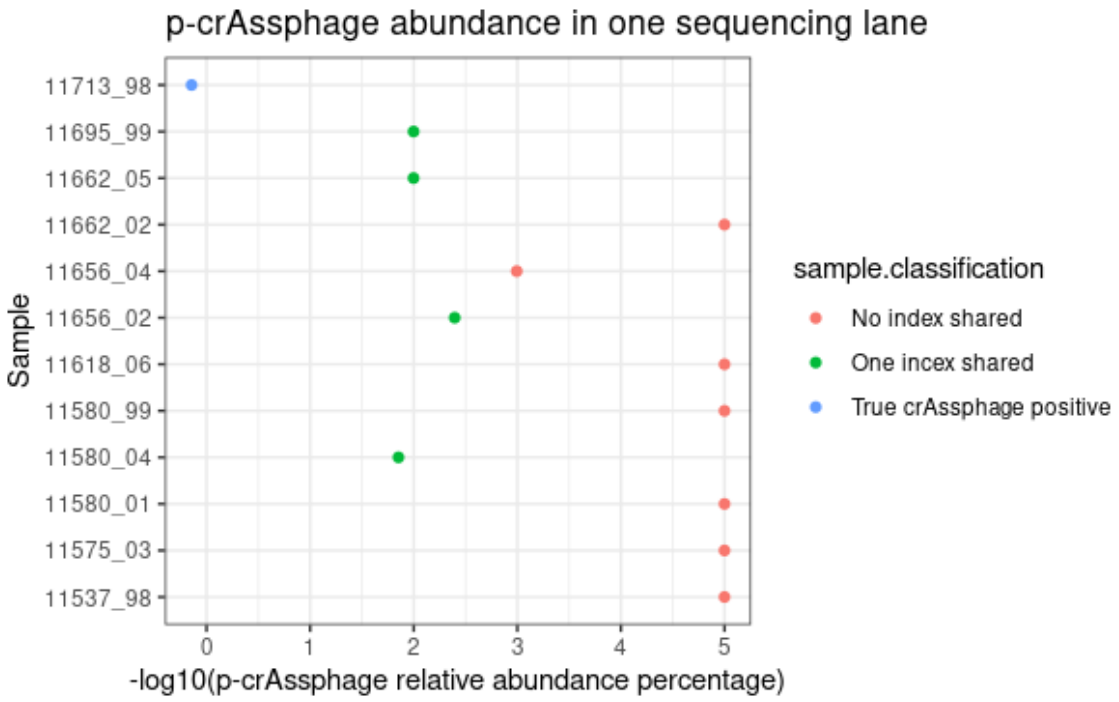
962 to be assigned to improper samples because of barcode swapping, we estimate the rate in our linked-
963 read data to be 0.1-0.2%.

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

988 between other samples from the two patients could cause false positives. While this filtering may discard
989 legitimate transmission events, we believe it is necessary to lower the number of false positives.
990 Previous DNA extraction and short-read sequencing efforts did not follow the randomization
991 constraints above and we cannot guarantee that laboratory contamination did not happen at some point
992 in the process. However, we note that cases of laboratory contamination or barcode swapping would
993 result in the entire microbiome composition of one sample being transferred to another. After our
994 stringent filters, we only discovered one case where patients shared two separate species. As these
995 were both *Lactobacillus* species, our hypothesis about probiotic consumption is a possible explanation.



996

997

998 # References

999 1. Young, J.-A. H. et al. Infections after Transplantation of Bone Marrow or Peripheral Blood Stem Cells
1000 from Unrelated Donors. Biol. Blood Marrow Transplant. 22, 359–370 (2016).

1001 2. See, I. et al. Mucosal Barrier Injury Laboratory-Confirmed Bloodstream Infection: Results from a
1002 Field Test of a New National Healthcare Safety Network Definition. Infect. Control Hosp. Epidemiol. 34,
1003 769–776 (2013).

1004 3. Kelly, M. S. et al. Gut Colonization Preceding Mucosal Barrier Injury Bloodstream Infection in
1005 Pediatric Hematopoietic Stem Cell Transplantation Recipients. Biol. Blood Marrow Transplant. 25, 2274–
1006 2280 (2019).

1007 4. Tamburini, F. B. et al. Precision identification of diverse bloodstream pathogens in the gut
1008 microbiome. Nat. Med. 1 (2018) doi:10.1038/s41591-018-0202-8.

1009 5. Zhai, B. et al. High-resolution mycobiota analysis reveals dynamic intestinal translocation preceding
1010 invasive candidiasis. Nat. Med. 1–6 (2020) doi:10.1038/s41591-019-0709-7.

1011 6. Taur, Y. et al. Intestinal Domination and the Risk of Bacteremia in Patients Undergoing Allogeneic
1012 Hematopoietic Stem Cell Transplantation. Clin. Infect. Dis. 55, 905–914 (2012).

1013 7. Ubeda, C. et al. Vancomycin-resistant Enterococcus domination of intestinal microbiota is enabled
1014 by antibiotic treatment in mice and precedes bloodstream invasion in humans. J. Clin. Invest. 120, 4332–
1015 4341 (2010).

1016 8. Jenq, R. R. et al. Intestinal Blautia Is Associated with Reduced Death from Graft-versus-Host
1017 Disease. Biol. Blood Marrow Transplant. 21, 1373–1383 (2015).

1018 9. Mathewson, N. D. et al. Gut microbiome–derived metabolites modulate intestinal epithelial cell
1019 damage and mitigate graft-versus-host disease. Nat. Immunol. 17, 505–513 (2016).

1020 10. Peled, J. U. et al. Microbiota as Predictor of Mortality in Allogeneic Hematopoietic-Cell
1021 Transplantation. N. Engl. J. Med. 382, 822–834 (2020).

1022 11. Shono, Y. et al. Increased GVHD-related mortality with broad-spectrum antibiotic use after
1023 allogeneic hematopoietic stem cell transplantation in human patients and mice. Sci. Transl. Med. 8,
1024 339ra71 (2016).

1025 12. Taur, Y. et al. The effects of intestinal tract bacterial diversity on mortality following allogeneic
1026 hematopoietic stem cell transplantation. Blood 124, 1174–1182 (2014).

1027 13. Weber, D. et al. Microbiota Disruption Induced by Early Use of Broad-Spectrum Antibiotics Is an
1028 Independent Risk Factor of Outcome after Allogeneic Stem Cell Transplantation. Biol. Blood Marrow
1029 Transplant. 23, 845–852 (2017).

1030 14. L. Livornese Jr, L. et al. Hospital-acquired Infection with Vancomycin-resistant Enterococcus
1031 faecium Transmitted by Electronic Thermometers. Ann. Intern. Med. (1992).

1032 15. Raven, K. E. et al. Complex Routes of Nosocomial Vancomycin-Resistant Enterococcus faecium
1033 Transmission Revealed by Genome Sequencing. Clin. Infect. Dis. 64, 886–893 (2017).

1034 16. Bäckhed, F. et al. Dynamics and Stabilization of the Human Gut Microbiome during the First Year
1035 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-
1148 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-Waldrip, 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 Env. 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		
Any BSI	54	36%
Bacillus	1	1%
Enterobacter	2	1%
Enterococcus	7	5%
Escherichia	8	5%
Gemella	1	1%
Klebsiella	6	4%
Pseudomonas	2	1%
Rothia	4	3%
Staphylococcus	14	9%
Streptococcus	9	6%

Table 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%	

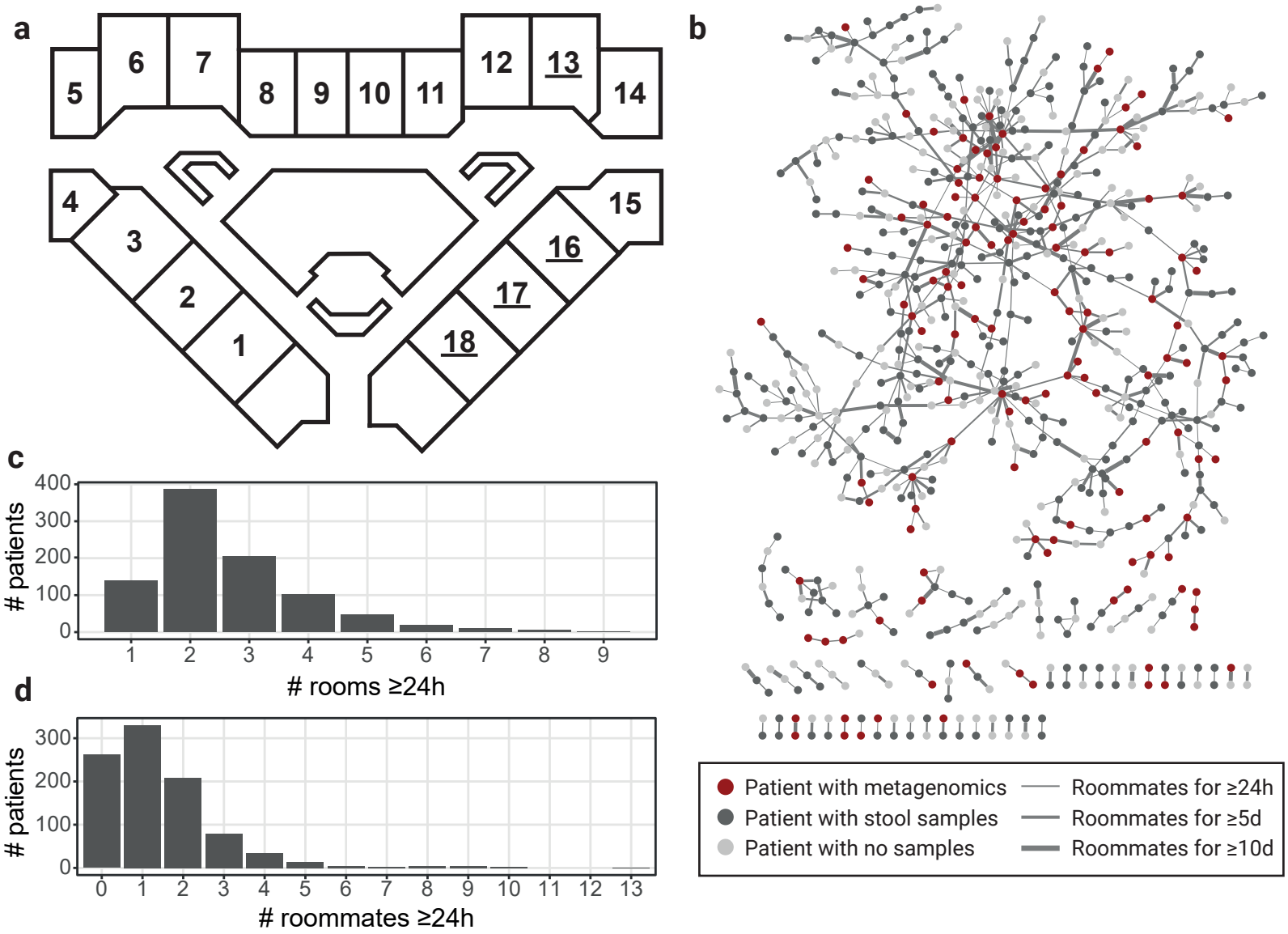


Figure S1: Analysis of hospital geography.
a) Layout of rooms in the HCT ward. Room numbers are indicated and double occupancy rooms are underlined.
b) Network view of patients who were roommates for at least 24 hours. Each node represents a single patient, colored according to if they have a banked 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.

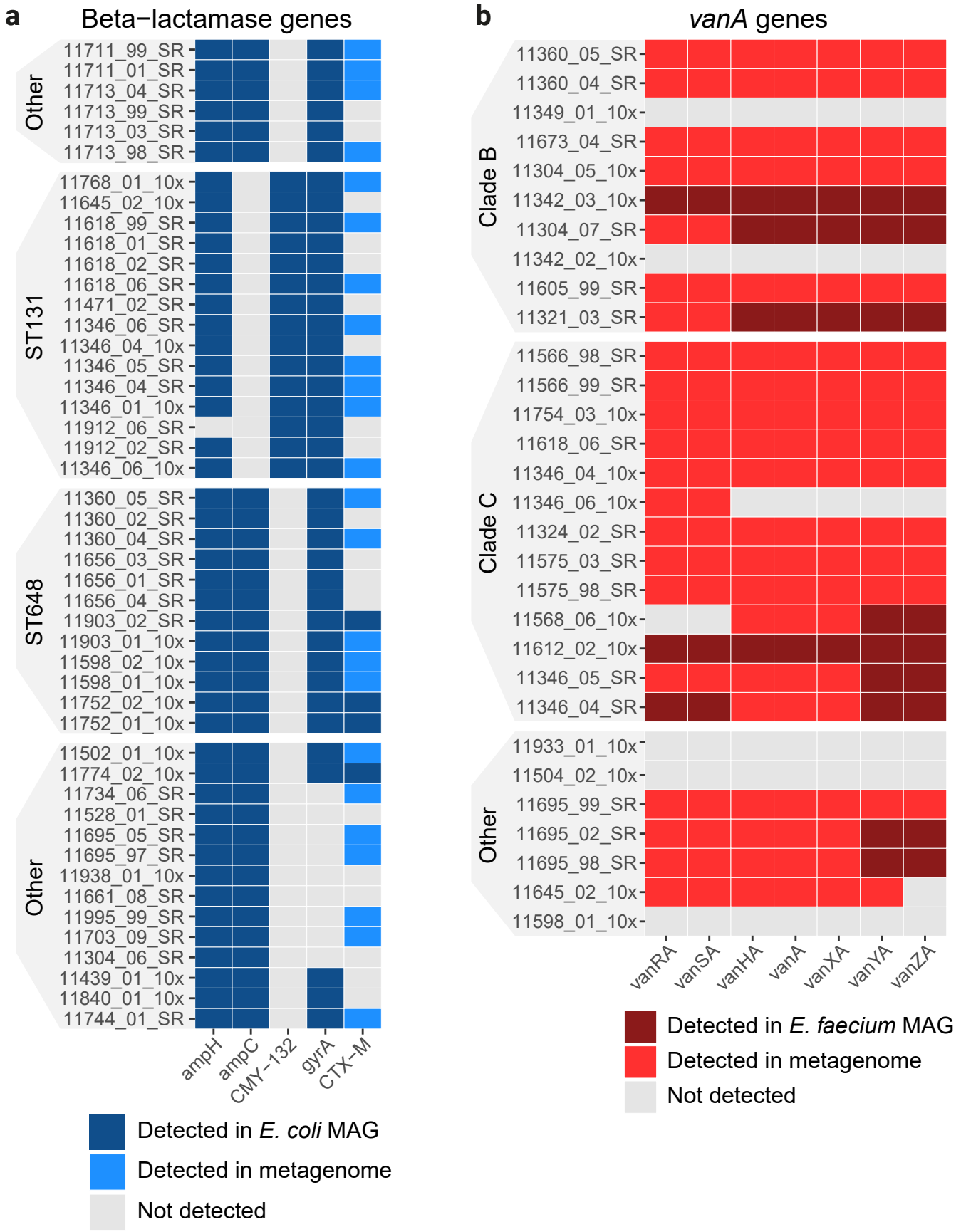


Figure S2: Antibiotic resistance genes detected. In each panel, samples are rows and resistance genes are columns. Samples are ordered and clades are highlighted corresponding to the respective figure in the main text. Cells are colored whether the gene was detected in the respective MAG from the sample, 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.

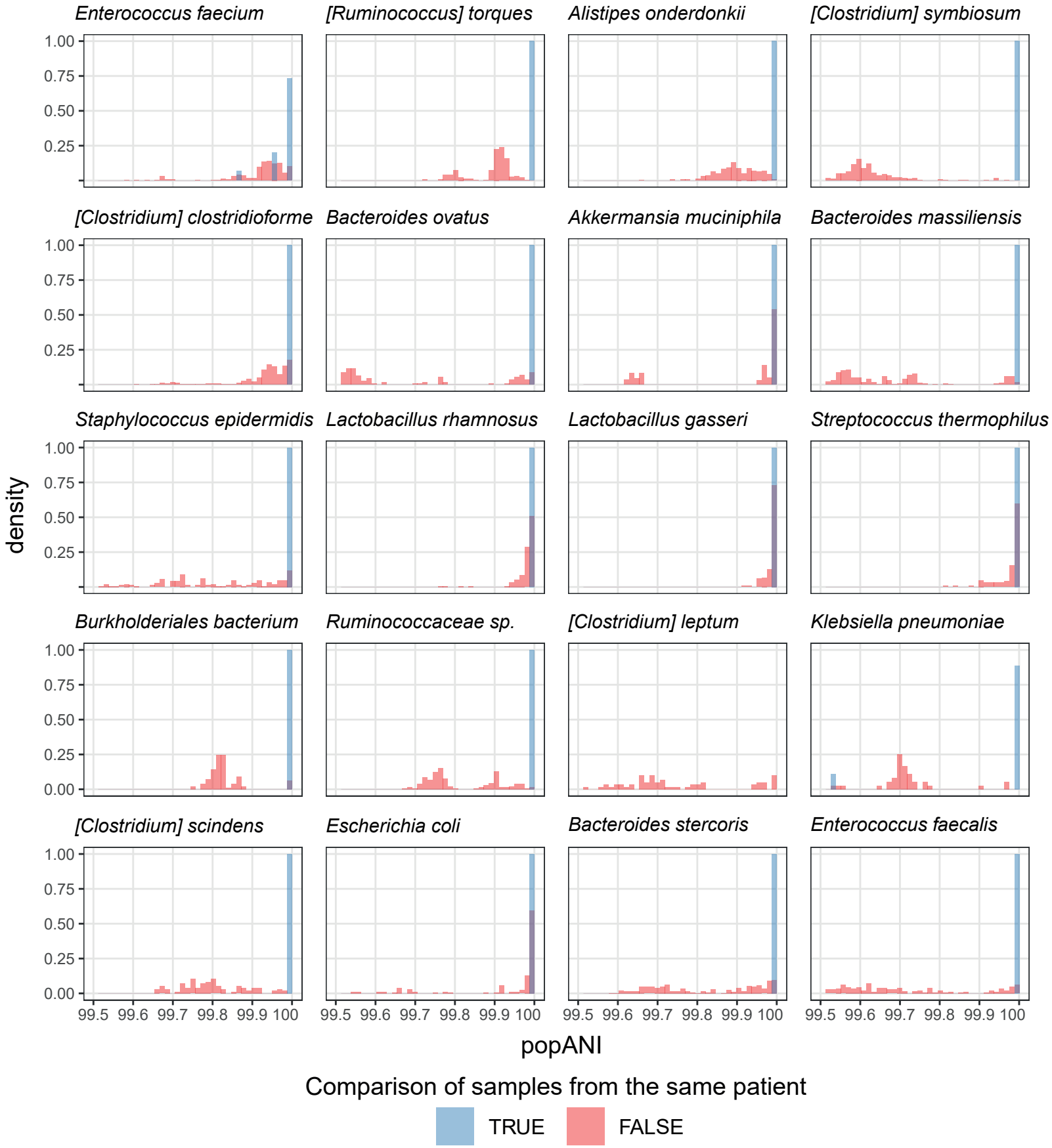


Figure S3: Distribution of popANI values comparing samples from the same or different patients. Distributions are split by species and the most common 25 species are shown. While in many cases the two distributions overlap, very rarely did popANI values comparing samples from different patients exceed the 99.999% transmission threshold. Comparisons with <99.5% popANI are omitted from the figure for clarity.

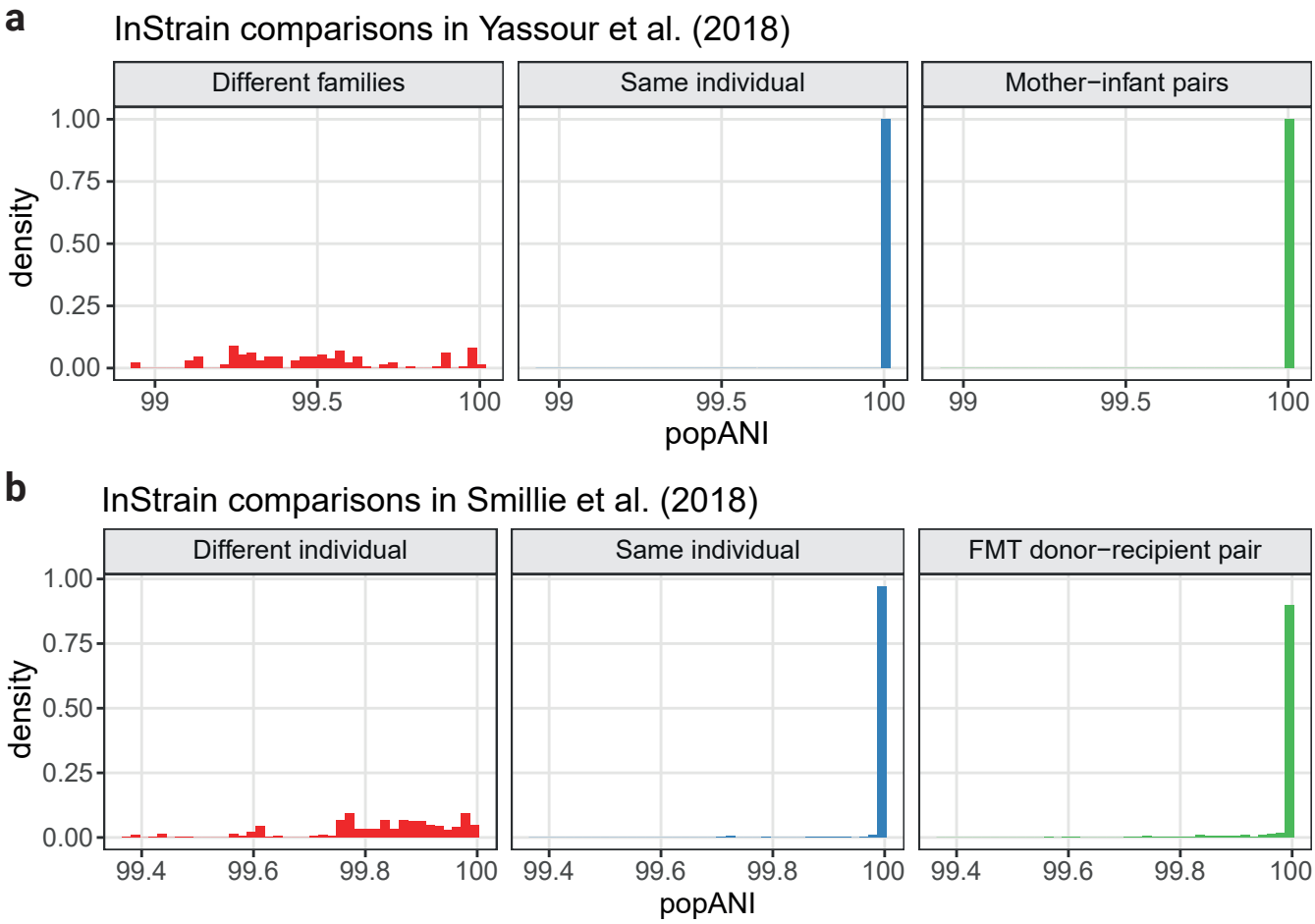


Figure S4: InStrain analysis of the five most common species in external datasets where transmission is expected to occur. Distributions of popANI values are separated based on the individuals the samples came from, with putative transmission events contained in the far right panel.
a) Metagenomic sequencing datasets from mother-infant pairs[18]. The maximum popANI value obtained when comparing samples from different families was 99.995%.
b) Metagenomic sequencing datasets from fecal microbiota transplantation donors and recipients[67]. The maximum popANI value obtained comparing samples from individuals not related by FMT was 99.998%.

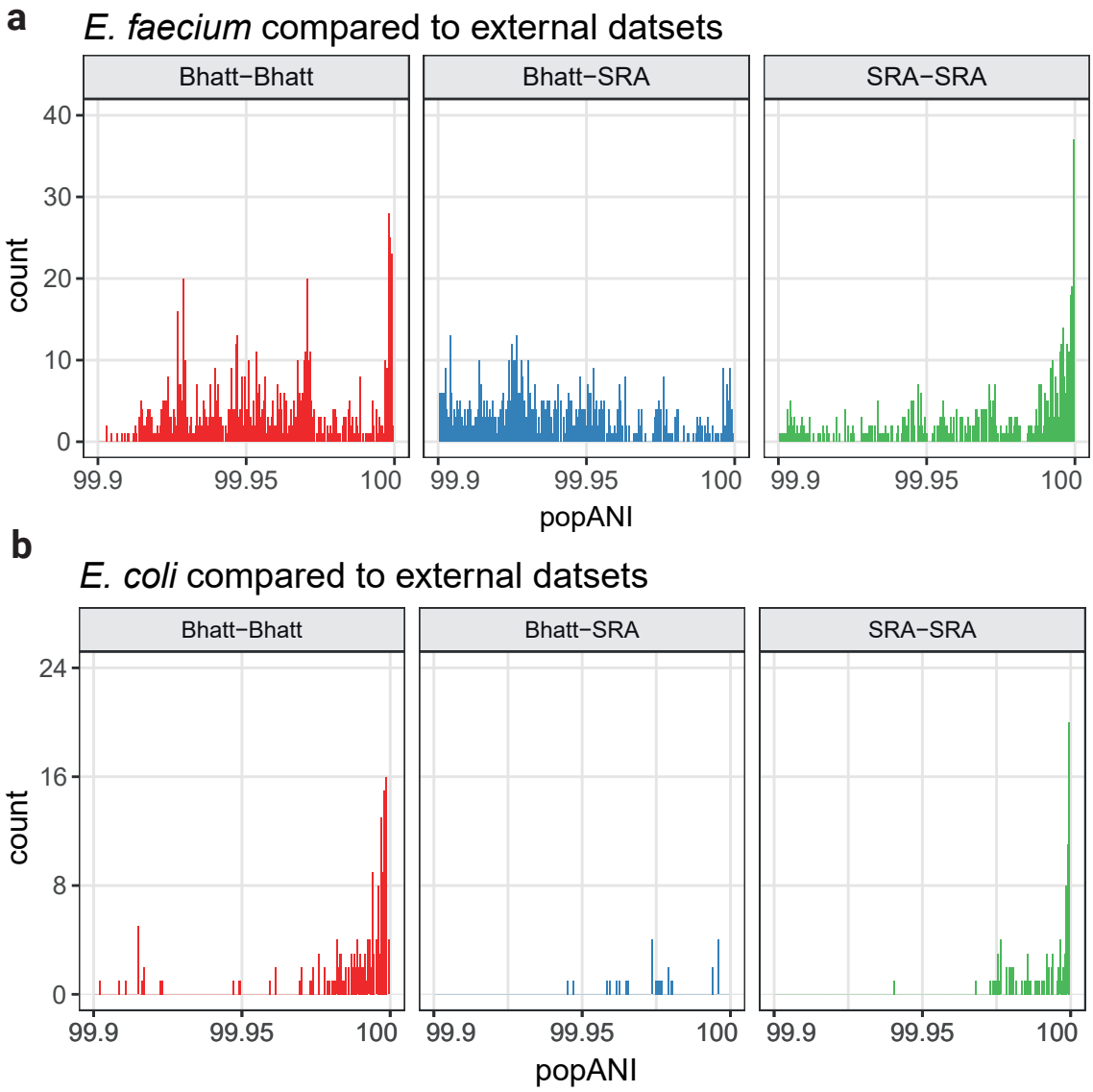


Figure S5: *Enterococcus faecium* (a) and *Escherichia coli* (b) strains compared to external datasets, including hospitalized adult and pediatric HCT patients, hospitalized infants and vancomycin-resistant *E. faecium* isolates[3,69-73]. Panels are separated according to whether comparisons were made within the data in this manuscript (Bhatt-Bhatt), between our data and external data (Bhatt-SRA) or within external data (SRA-SRA).

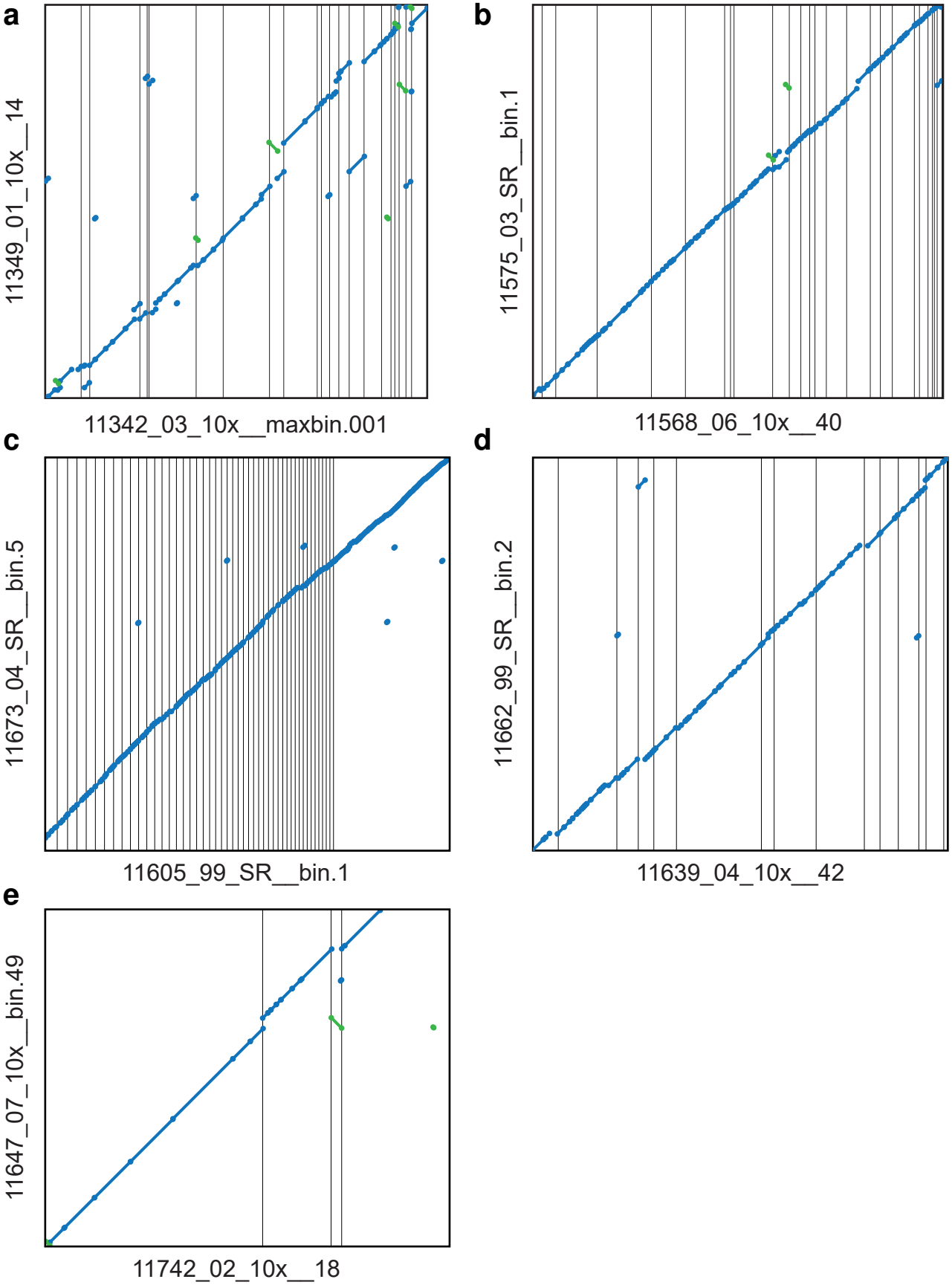


Figure S6: Dotplots showing pairwise alignment of MAGs in cases of putative transmission of the given species. Blue lines along the diagonal indicate 1-1 homology between the two sequences. Green lines indicate inversions that are likely the result of assembly or binning errors.
a) *E. faecium* MAGs 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.

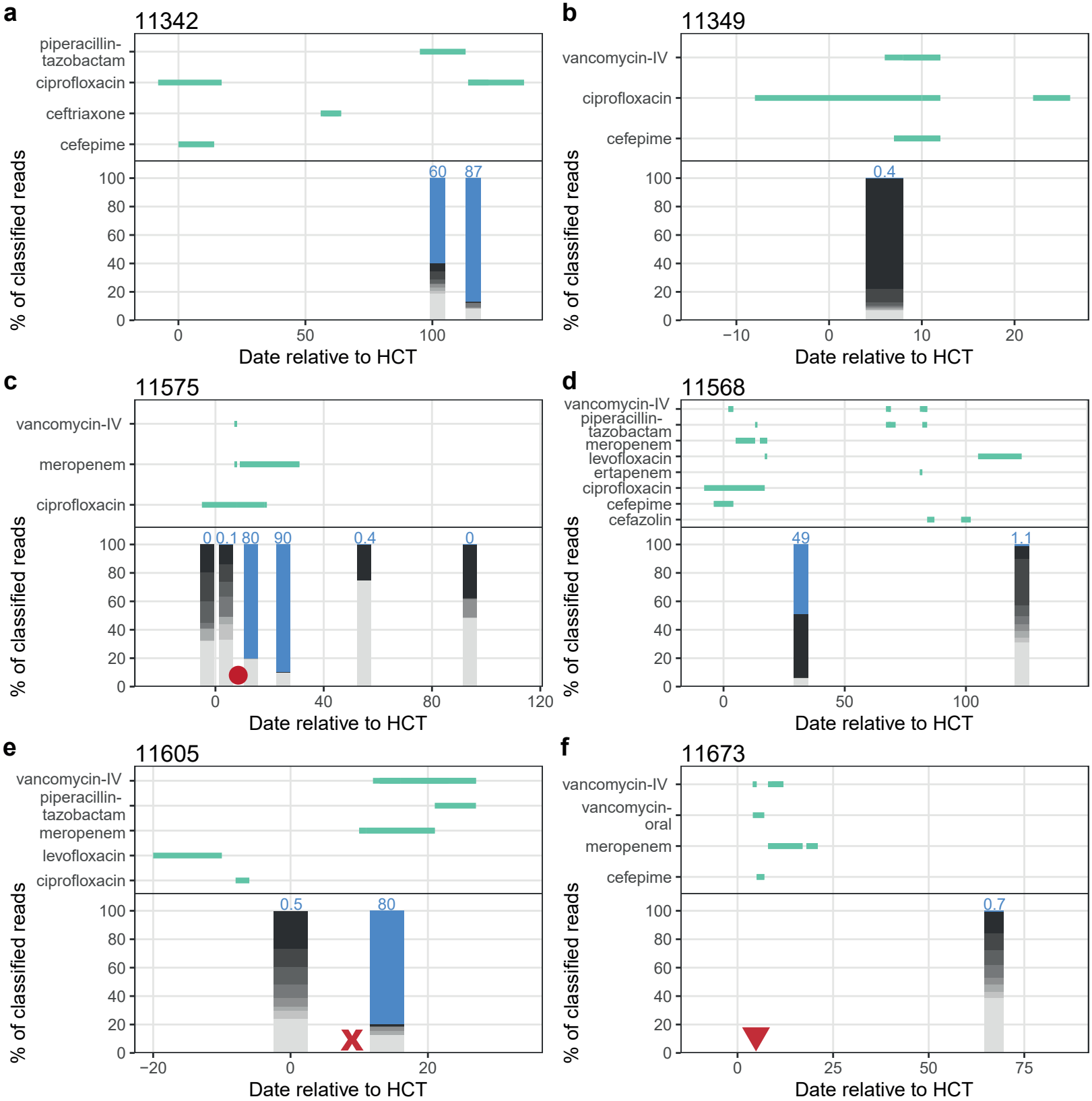


Figure S7: Antibiotic prescription and taxonomic composition of patients with nearly identical *Enterococcus faecium* strains. *E. faecium* abundance is shown in blue and indicated with text. Other taxa are shown in grey. All dates are relative to HCT for the particular patient. Approximate dates of BSI are shown with red symbols. Circle: *Klebsiella pneumoniae*, X: *Enterococcus faecium*, triangle: *Escherichia coli.*

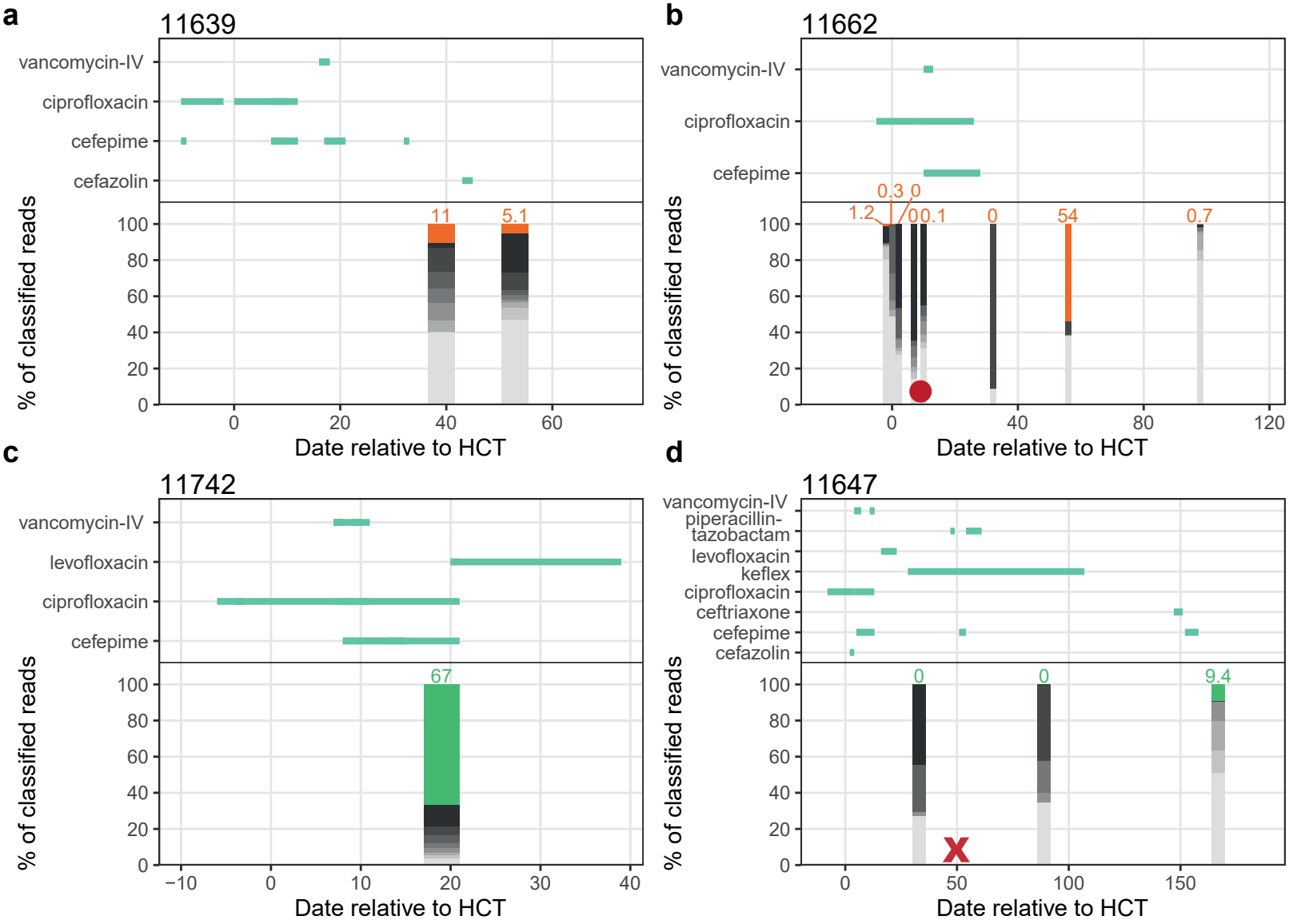


Figure S8: Antibiotic prescription and taxonomic composition of patients with nearly identical *Hungatella hathewayi* (shown in orange) or *Akkermansia muciniphila* (shown in green) strains. Other taxa are shown in grey. All dates are relative to HCT for the particular patient. Approximate dates of BSI are shown with red symbols. Circle: *Streptococcus mitis*, X: *Klebsiella pneumoniae.*

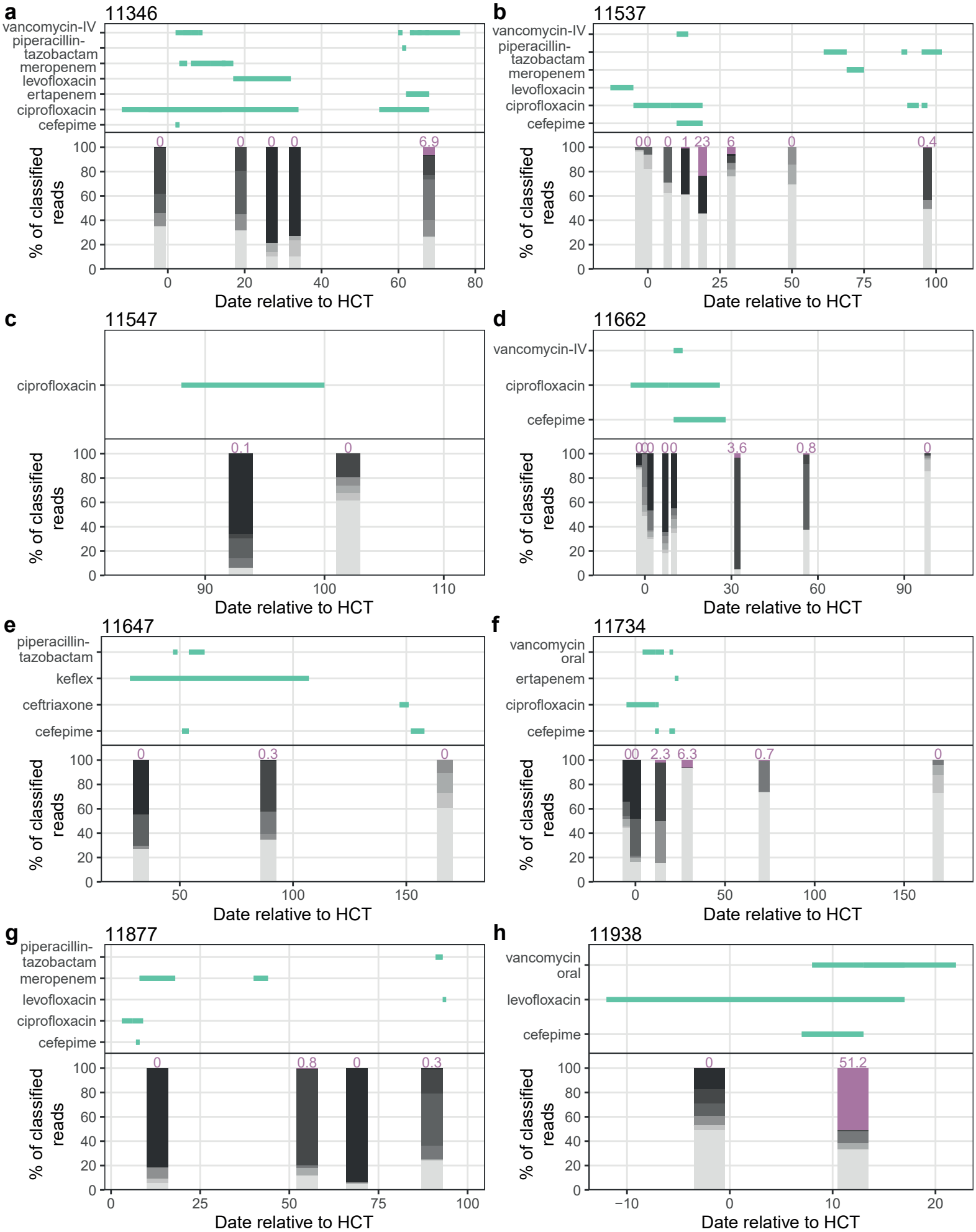


Figure S7: Antibiotic prescription and taxonomic composition of patients with nearly identical *Lactobacillus rhamnosus* strains. *L. rhamnosus* abundance is shown in purple and indicated with text. Other taxa are shown in grey. All dates are relative to HCT for the particular patient.