Science Translational Medicine
AAAS

Supplementary Materials for

Gut pathogen colonization precedes bloodstream infection in the neonatal intensive care unit

Drew J. Schwartz *et al.*

Corresponding author: Gautam Dantas, dantas@wustl.edu

Sci. Transl. Med. **15**, eadg5562 (2023)
DOI: 10.1126/scitranslmed.adg5562

The PDF file includes:

Materials and Methods
Figs. S1 to S11
Legends for data files S1 to S5
References (*57*–*73*)

Other Supplementary Material for this manuscript includes the following:

Data files S1 to S5
MDAR Reproducibility Checklist

Materials and Methods

Metadata associations with the microbiome

To determine associations between antibiotic receipt, BSI, and changes in the gut microbiome, MaAsLin2 was used with default parameters (*41*). Metagenomic sequencing data for this analysis was downloaded from SRA from previously published shotgun metagenomic sequencing data from the overall cohort (*20, 37*). MetaPhlAn3 (v3.0.7) (*56*) was used to assign taxonomy from sequenced reads. DOL and antibiotics were used as fixed effects and individual was used as a random effect. For associations between BSI organism and the gut microbiome in Figure 5, a more stringent q value <0.05 was used. A minimum prevalence filter of 0.05 was set instead of the default 0.1 for BSI with GBS and *Serratia marcescens* because of the rarity of finding these organisms in the gut microbiome. For this analysis we included all samples sequenced for this manuscript as well as previously sequenced samples (*20, 37*). Antibiotic start and stop dates and dates of collection of the ultimately positive blood culture were used to determine number of antibiotic days before bacteremia. Antifungals and antivirals were excluded. When an infant was administered combinations of antibiotics (for example, ampicillin and gentamicin together), each day of each antibiotic was summed. If an antibiotic was started and stopped on the same or following day, a score of 1 was assigned.

Case-control study design

From the overall cohort (*35, 36*), 139/977 infants (14%) had at least one positive blood culture after DOL 3 (Fig. S1) with an average onset of 21 +/- 16 DOL. Eleven infants had a positive blood culture with a different organism simultaneously or later in life for 151 unique

BSI organisms. Consistent with other NICUs in the United States (*3*), staphylococci (usually coagulase negative) were the most frequently isolated with 104/151 (69%) positive cultures followed by 26/151 (17%) with *Enterobacteriaceae*. Streptococci were isolated in 10/151 (7%), enterococci in 4/151 (3%) and 5% with other organisms. We included case participants in our nested case-control study if they suffered BSI after the 3rd day of life consistent with LOS, if the bloodstream isolate was available and stored at -80 °C and not thawed, and if the neonates had at least 1 stool on the day of or in the 7 days prior to the date of infection that was not volume-limited (Fig. S1). We excluded neonates for whom the BSI isolate could not be definitively linked with bacteremia, including positive blood cultures with *Clostridium spp.*, *Bacillus spp.*, and coagulase-negative *Staphylococcus*, which could represent cutaneous contamination (*3*). We included only 6 infants with BSI caused by *Staphylococcus aureus* to evaluate its presence in the gut. Two cases with concomitant bacteremia and necrotizing enterocolitis (NEC) were included. With these criteria, we included 19 infants with LOS BSI in our analysis (Table 1, Data file S1). Relevant events (BSI, NEC) from previously interrogated infants from prior studies (*20, 22, 36, 37*) are also detailed in Data File S1.

We aimed to match two control participants to each case participant. Control participants were selected from the same cohort who did not have a culture-positive BSI during their stay in the NICU (Fig. S1). Sixteen cases were from the SLCH NICU and 3 were from the Norton (formerly Kosair) Children's Hospital NICU. We matched cases to controls by minimizing Euclidean distance on the following clinical factors that were transformed to a range between 0 and 1: GA, antibiotic score, birthweight, sex, and delivery mode (Table 1, Data file S1). Number of antibiotic days prior to bacteremia or equivalent control DOL was determined by summing the number of days of each antibiotic prior to that DOL. Non-systemic antimicrobials such as mupirocin, antifungals, and antivirals were excluded. Cumulative antibiotic exposure was calculated for each participant to quantitatively describe the participant's antibiotic exposure during their stay in the NICU. For each antibiotic to which a participant was exposed, we calculated an antibiotic score by multiplying the antibiotic spectrum score (ranging from 1-4), based on the Duke University antibiotic stewardship program (Data file S3) (*43*), by the number of days the treatment lasted. The participant's antibiotic score was calculated by summing each treatment antibiotic score. For sample-based comparisons a per-sample antibiotic score was calculated to the date that the sample was obtained. Given the known impact of GA on the gut microbiome (*20, 36*), we initially matched on GA +/- 2 weeks, then birthweight +/- 200 grams, followed by antibiotic score, then sex and delivery mode. We did not specifically control for year of birth. Last, control participants were further selected based on sufficient stool availability during matched DOL with case participants. Previously sequenced samples from individuals without BSI were also included (*20, 37*). This process resulted in 1 control being used for 2 cases. Further, 1 control differed by more than 2 weeks in GA from its case, and 2 controls exceeded 200g difference in birthweight. Four controls did not match the delivery mode and/or sex of the paired cases (Data file S1). Thirty-one controls from St. Louis Children's Hospital (SLCH), 5 from Norton (formerly Kosair) Children's Hospital NICU, and 1 from University of Oklahoma NICU were ultimately included. We defined the date of BSI as the day the first positive blood culture was drawn. The DOL of BSI for control participants was assigned the same DOL of BSI for the corresponding cases (Table 1, Data file S1). The proportion of each daily feed (in mLs) from each category was summed over time for each neonate and divided by the total number of feeds or mLs if feeds were split (Table 1).

Library preparation and sample handling.

Approximately 25-100 mg of frozen stool was chipped from the original aliquot for subsequent use. Bacterial isolates were stored in 20% glycerol at -80 °C prior to resuscitation on agar plates or in broth. Stool samples were processed with the DNeasy PowerSoil Pro Kit (Qiagen) per manufacturer instructions with two rounds of bead beating for 2 min at 2500 oscillations/min on a Mini-Beadbeater-24 (Biospec Products) with 5 min on ice in between. Bacterial whole genomic DNA was obtained from 5mL of a pure overnight culture using the BiOstic Bacteremia DNA kit (Qiagen). DNA and sequencing library quantification was performed with the Quant-iT PicoGreen Fluorescence Assay (Invitrogen) and Qubit fluorometer dsDNA HS assay (Invitrogen). Nextera libraries were created with 0.5ng input DNA with the Nextera flex reagents (Illumina) and purified using the Agencourt AMPure XP system (Beckman Coulter). Pooled libraries were submitted for 2x150 paired-end sequencing on the Illumina NextSeq 550 and NovaSeq 6000 platforms. Stool samples were sequenced to a target read depth of 5 million paired reads per sample and bacterial isolates were sequenced to a depth to achieve 100x genome coverage based on the genome size of the pathogenic organism. Read counts for each sample are provided in data file S1. Samples with less than 100,000 post-processed reads were excluded from downstream analyses leaving 449 samples for metagenomic analysis.

Bioinformatic read processing and assembly

All isolate genomic and fecal metagenomic reads were processed using FastQC (v0.11.7) (*57*), Trimmomatic (v0.36) (*58*) with parameters "ILLUMINACLIP: NexteraPE-PE.fa:2:30:10:1:TRUE LEADING:10 TRAILING:10 SLIDINGWINDOW:4:15 MINLEN:60", Deconseq (v0.4.3) (*59*) with reference to the human chromosome to remove human DNA contamination, and then repaired with the repair.sh script in BBMap (v38.90) (*60*) to order read sides with disordered mates. Processed genomic DNA from bloodstream isolates was assembled using SPAdes (v3.14.0) (*61*) with the following parameters "-k 21,33,55,77,87,97,107,117,127 –isolate". Contigs smaller than 1kb were removed. The assembly with the largest kmer (k=127), termed *isolate reference*, was used for downstream analysis. Mash v2.2 (*62*) was used to confirm the species identity of the bacteremia isolate from the determination from the clinical microbiology laboratory (*62*). CheckM v1.0.13-python-2.7.15 (*63*) and Quast v3.2 (*64*) were used to perform quality control on the isolate genome assemblies (*64*). Genome assemblies were retained if they were >90% complete, <5% contamination, <500 contigs and >1000bp long.

Metagenomic taxonomic assignment and assembly

Processed metagenomic reads were used as input to MetaPhlAn3 (v3.0.7) (*56*). Species calls from MetaPhlAn3 were used to calculate the Shannon diversity of metagenomic samples and Bray-Curtis dissimilarity. Bray-Curtis dissimilarities were used for PcoA analysis and visualization of sample pair-wise dissimilarity distributions. Repeat measures PERMANOVA was utilized to analyze Bray-Curtis distances accounting for longitudinal sampling. Briefly, we adapted prior implementations (*48, 65*) and assigned participant-specific metadata such as individual, birth mode, case vs. control assignment, sex, gestational age, birthweight, BSI organism family, or sample-specific metadata such as DOL, antibiotic days, antibiotic score, days receiving human milk (HM), days of formula, days without oral nutrition, non per os (NPO) days, and DPI. The contribution to microbiome variance and p value was calculated individually with Participant ID blocked with the script repeated_measures_PERMANOVA.R (*65*) then the p-values were adjusted with Benjamini-Hochberg correction. To determine relative abundance quartiles in DPI bins, the abundance rank for the causative species was computed for each sample. When more than one sample existed within a DPI bin, the median value of the rank abundance was used.

Gene annotations and antibiotic resistance analysis

Processed metagenomic reads were supplied to ShortBRED (v0.9.4) (*66*). The antibiotic resistance databases supplied were referenced to CARD2020 (*67*) and results from prior functional metagenomic screens (*66*). ARGs conferring resistance to beta-lactams, aminoglycosides, and vancomycin were determined based on CARD2020 annotation. Sample antibiotic resistance richness was defined as the number of unique antibiotic resistance factors present with reads per kilobase of reference sequence per million sample reads (RPKM) greater than 0. Antimicrobial susceptibility testing was performed on Mueller-Hinton agar (Hardy Diagnostics) with Kirby-Bauer disk diffusion. Antibiotic-impregnated discs were purchased from Hardy Diagnostics and Becton, Dickinson. Susceptible, intermediate, and resistant were determined by reference to Clinical and Laboratory Standards Institute criteria (*68*). Resfinder v4.0 (*69*) was used to assign resistance gene annotation to BSI isolates.

Strain similarity analyses

Isolate reference sequences were indexed using BWA (v0.7.17) with the command "bwa index" (*70*). Processed reads from stool samples and isolate samples were individually mapped onto the corresponding participant isolate reference using the command "bwa mem." The resulting SAM files were converted into BAM files using Samtools (v1.5) through the command "samtools view -S -b" (*71*). The BAM files and isolate references were used as input for inStrain (v1.3.7) with the command "inStrain profile" (*26*). To determine the most stringent definition of strain similarity, we mapped isolate reads to its own assembly. We identified 0-11 SNS, depending on sequencing depth from 100,000 to 1 million reads (Data file S4). We used population ANI >0.99999 to indicate the same strain, which based on 2-6 megabase genomes, would be a maximum of 20-50 SNS between isolate and metagenome. We used the output from inStrain profile to determine differences in population ANI between BSI families. We first performed negative Log (1-"popANI_reference"). We excluded samples with popANI of 1 with breadth less than 0.5 as these indicate low coverage between the metagenome and the isolate reference (*26*). For strain sharing analyses in Figure 5, metagenomic sample reads were preprocessed and mapped to isolate references from case participants according to the same procedures above. We first masked self-comparisons and restricted to breadth >0.5 to be certain we had effective read mapping (*26*). Then, we plotted -Log(Breadth)*(1-SNSrate) which incorporates organism presence (breadth) as well as penalizing for substitutions (SNSrate). We analyzed differences in strain sharing for *Enterobacteriaceae*/*Enterococcaceae* versus *Staphylococcaceae*/*Streptococcaceae* using Fisher's Exact test with all possible unique donor-recipient combinations in the denominator. We used a cutoff of Log (breadth/SNSrate) of 5 to limit false positives and ensure that the BSI isolate genome was present in sufficiently high breadth. Above this threshold included all instances of metagenome-isolate pairs with population ANI > 0.99999.

Coverage analysis

The coverage of a metagenomic-isolate reference pair was defined as the median coverage value among all isolate reference nucleotide positions. The sample normalized coverage was defined as $C_{sample} = \frac{M}{R}$, where M is the coverage and R is the total number of processed reads in the metagenomic sample. The participant normalized coverage was defined as $C_{participant} = \frac{M - \min(s)}{\max(s)}$, where $\min(s)$ is the minimum coverage among samples produced by participant s and $\max(s)$ is the maximum coverage sample produced by participant s. For determination of coverage over time relative to BSI (Fig. 4B-C), only participants that produced at least one stool with breadth >99% and fewer than 20 substitutions were included. To determine Log breadth for Fig. 4D, we calculated Log(1-breadth). For samples with a breadth of 1, Log(1-breadth) was set to 7. For pairs that produced 0 substitutions, negative Log SNS/genome was set to 7.

Enterotype modeling

Enterotyping of stool metagenomes was performed using Dirichlet Multinomial Mixtures (DMM) modeling which clusters samples based on microbial community structure (72). Taxonomic abundance matrices at the species and genus levels were first transformed by a factor of 5000 to increase the spread of values. The R package DirichletMultinomial v1.26.0 was then run with 1000 iterations of clustering at different starting seeds and a maximum of 20 clusters (73). AIC and Laplace approximations were generated for each cluster model and the best model was chosen based on lowest Laplace score. The taxonomic contribution was then extracted and plotted for each cluster within this model. The sample distribution within each enterotype was calculated and analyzed with Chi-squared test and the pairwise Nominalindependence test from rcompanion V2.4.15 with “fdr” correction.

Supplementary Figures

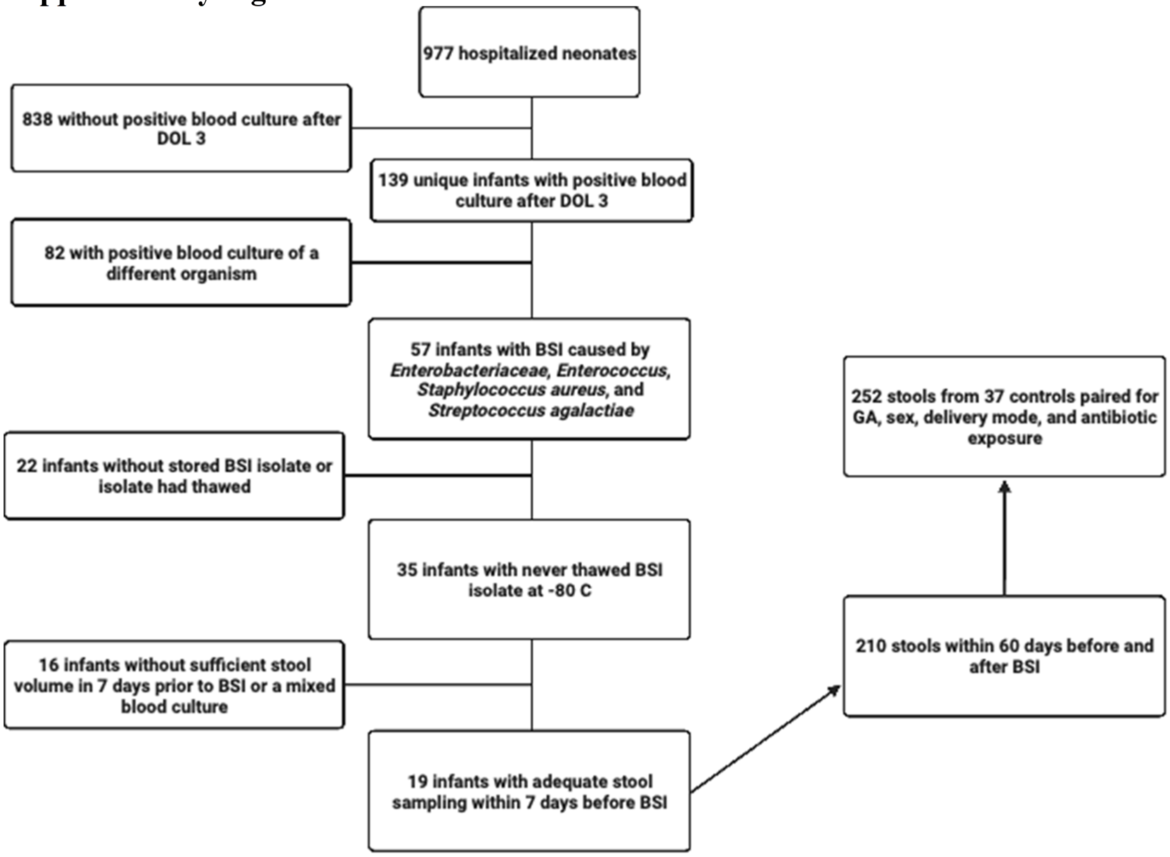
977 hospitalized neonates
838 without positive blood culture after DOL 3
139 unique infants with positive blood culture after DOL 3
82 with positive blood culture of a different organism
57 infants with BSI caused by *Enterobacteriaceae*, *Enterococcus*, *Staphylococcus aureus*, and *Streptococcus agalactiae*
252 stools from 37 controls paired for GA, sex, delivery mode, and antibiotic exposure
22 infants without stored BSI isolate or isolate had thawed
35 infants with never thawed BSI isolate at -80 C
210 stools within 60 days before and after BSI
16 infants without sufficient stool volume in 7 days prior to BSI or a mixed blood culture
19 infants with adequate stool sampling within 7 days before BSI

Figure S1. Inclusion and exclusion of infants and sample availability. Created with BioRender.com.

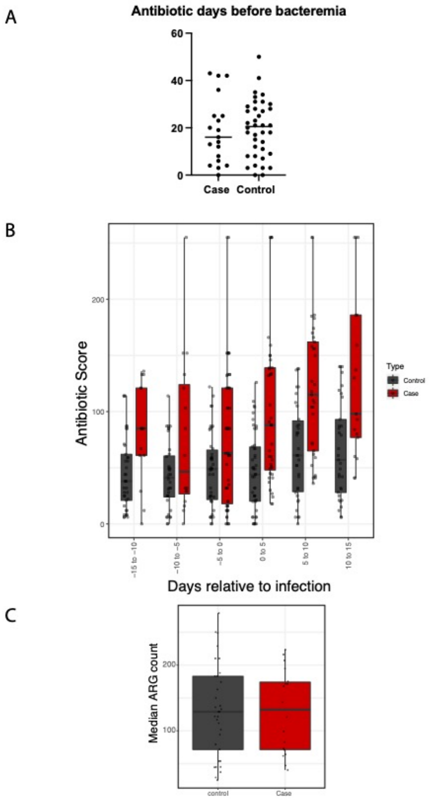

Figure S2. Antibiotic days, score, and resistance gene content do not differ between cases and controls. Sum of antibiotic days before BSI (A), cumulative antibiotic score (B), and median ARG count (C) do not differ between cases and controls. A) ns, Mann-Whitney U test. B-C) linear mixed effect model, data file S3

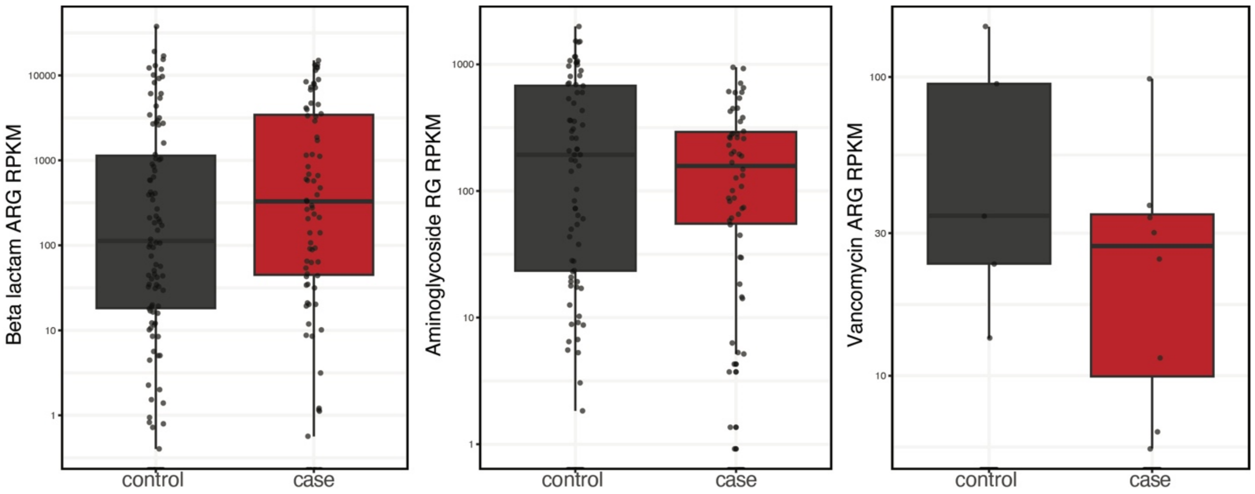

Figure S3. **Beta-lactam, aminoglycoside, and vancomycin resistance genes identified in metagenomes before BSI and in controls.** ShortBRED was used to profile ARGs in metagenomes in cases before BSI and in control infants over matched timeframes. ARGs annotated from the CARD database and displayed in reads per kilobase million (RPKM). Ns, Mann-Whitney U test

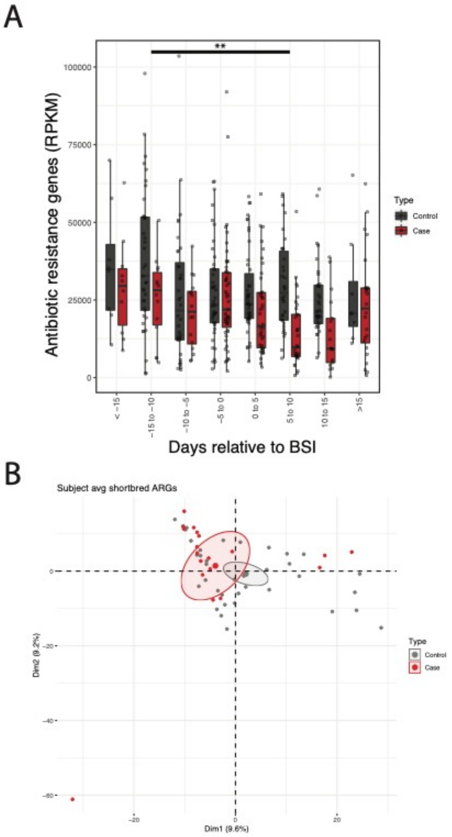

Figure S4. Control infants 2 weeks before BSI have higher ARG content. ShortBRED assigned ARG RPKM is higher in the 2 weeks prior to BSI for control infants versus case immediately after BSI and control infants 5 to 15 days after with linear mixed effect models (Data file S2). B) PCoA of Bray-Curtis dissimilarity of participant-averaged ARG content did not differ between cases and controls using PERMANOVA.

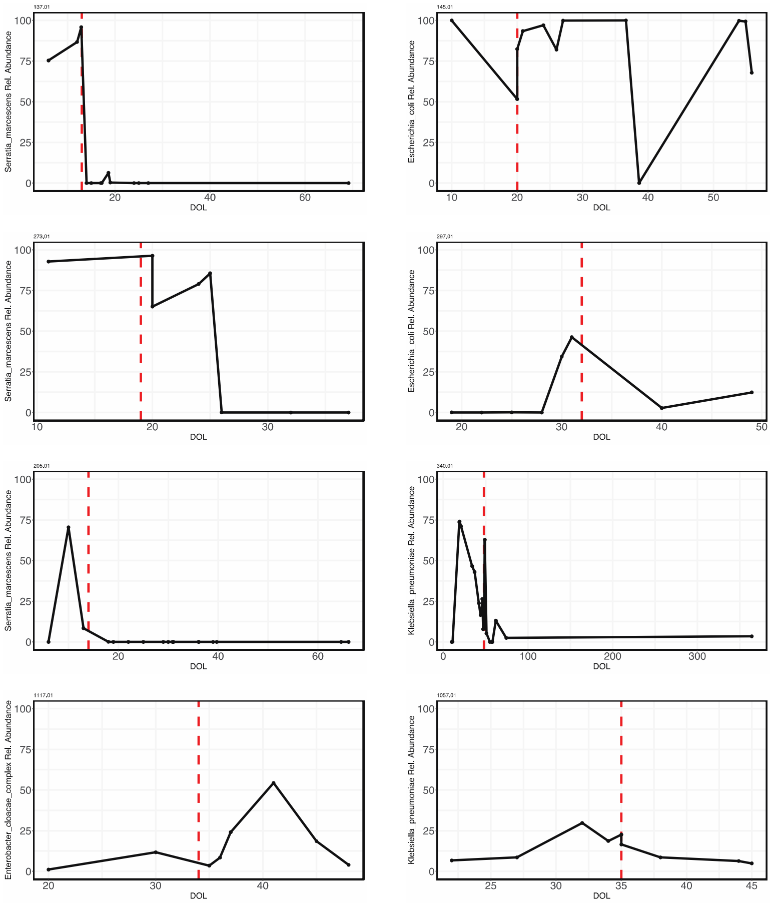

Figure S5. ***Enterobacteriaceae* relative abundance in stool of cases with BSI caused by those species using MetaPhlAn3.** Dashed line indicates BSI DOL.

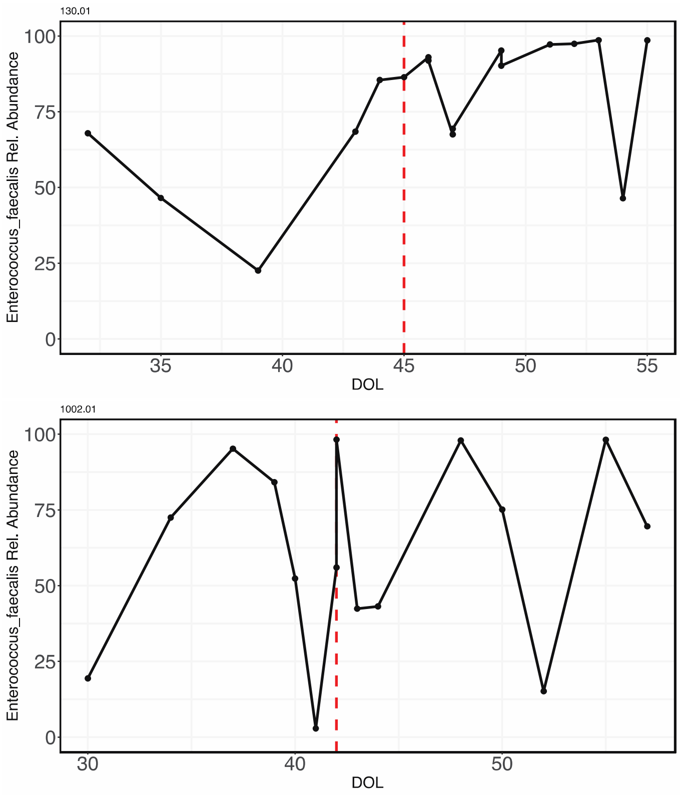

Figure S6. ***Enterococcus faecalis*** **relative abundance in stool of cases with BSI caused by those species using MetaPhlAn3.** Dashed line indicates BSI DOL.

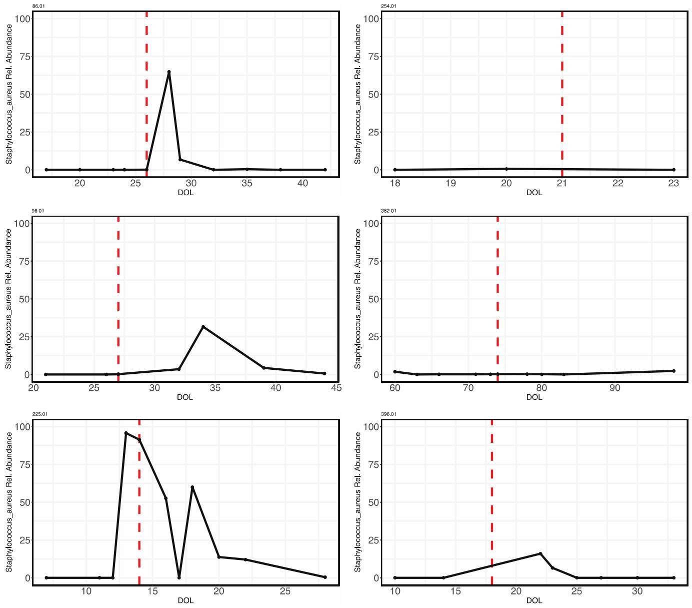

Figure S7. ***Staphylococcus aureus*** **relative abundance in stool of cases with BSI caused by those species using MetaPhlAn3.** Dashed line indicates BSI DOL.

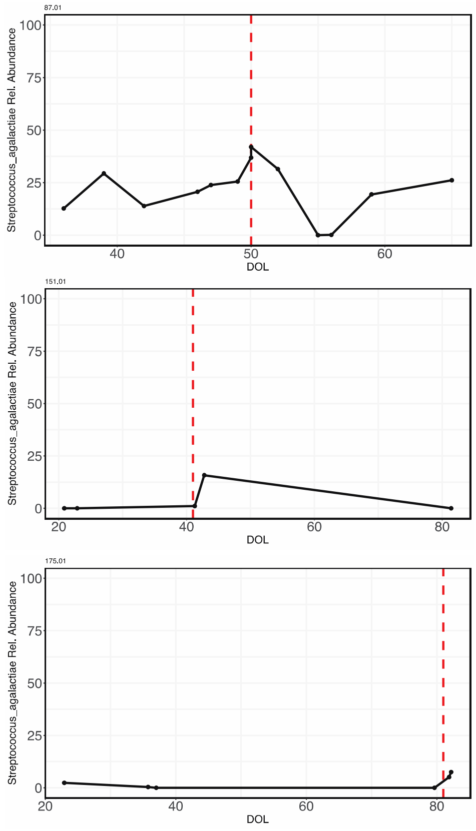

Figure S8. *Streptococcus agalactiae* relative abundance in stool of cases with BSI caused by those species using MetaPhlAn3. Dashed line indicates BSI DOL.

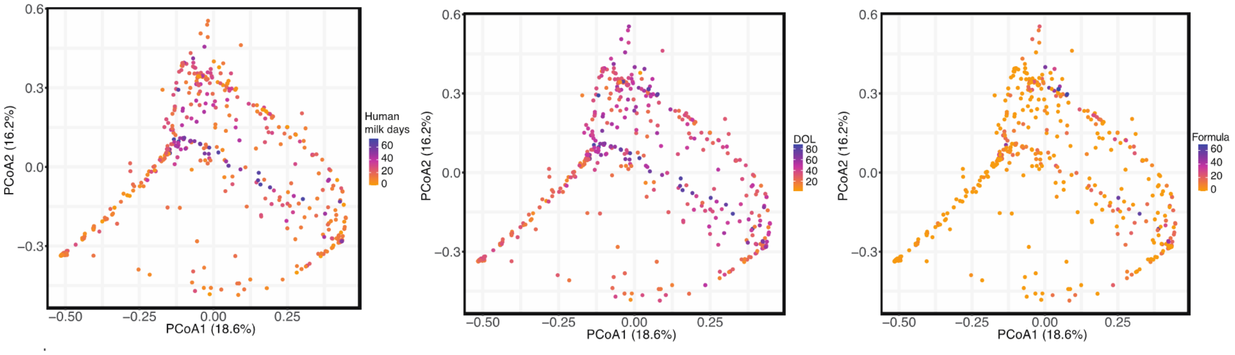

Figure S9. Gut metagenomes cluster by clinical metadata. Principle coordinate analysis of Bray-Curtis difference distinguish metagenomes by number of days of human milk, DOL, and number of days of formula. Repeated measures PERMANOVA variance and BH-corrected p-value from Fig. 2E.

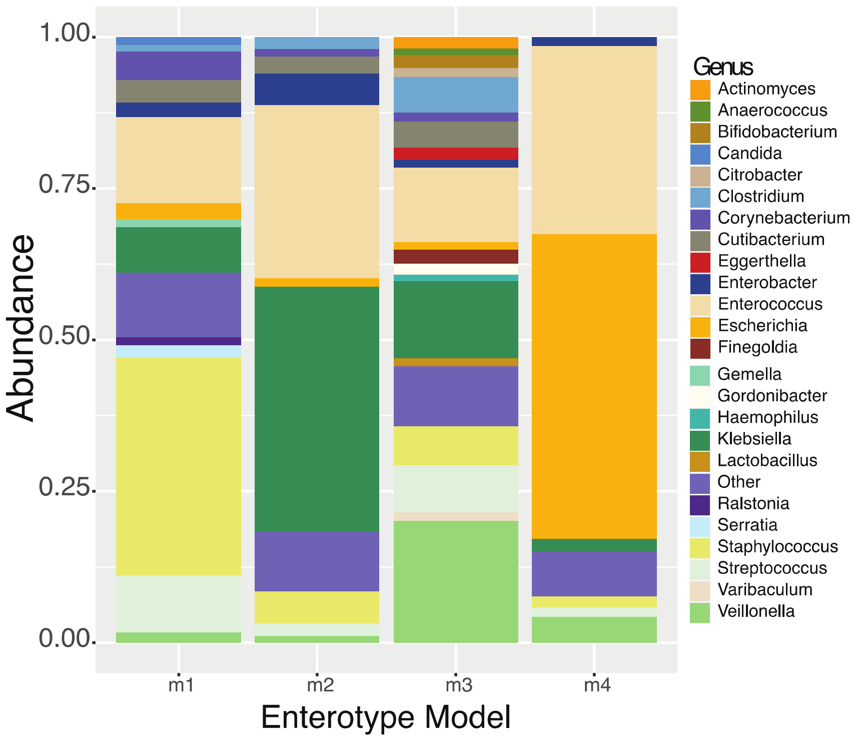

Figure S10. Enterotype modeling with Dirichlet Multinomial Mixtures identified 4 enterotypes among our case-control cohort. The representation within these enterotypes differed depending on BSI-causing organism (Data file S2). Specifically, 48% of control metagenomes before the day of BSI of their paired cases clustered into m1 with 19%, 18%, and 15% in m2-4, respectively. This distribution was statistically different from enterotype groupings from all other BSI-causing families (ChiSq, q<0.05, data file S2). Metagenomes from cases with BSI caused by Enterobacteriaceae were over-represented in m1 (55%) compared to 34% in m2 and 11% in m3 and m4 combined. Conversely, metagenomes from individuals with E. faecalis BSI were 50% in m4 and 43% in m1. Metagenomes from BSI cases with S. aureus were only observed in m1 (75%) and m3 (25%). Similarly, S. agalactiae metagenomes were dominant in m3 (53%) and m1 (40%). Representation within enterotypes differed based on BSI-causing family separately from all other BSI-causing families and controls. chisq, all comparisons q<0.01, data file S2.

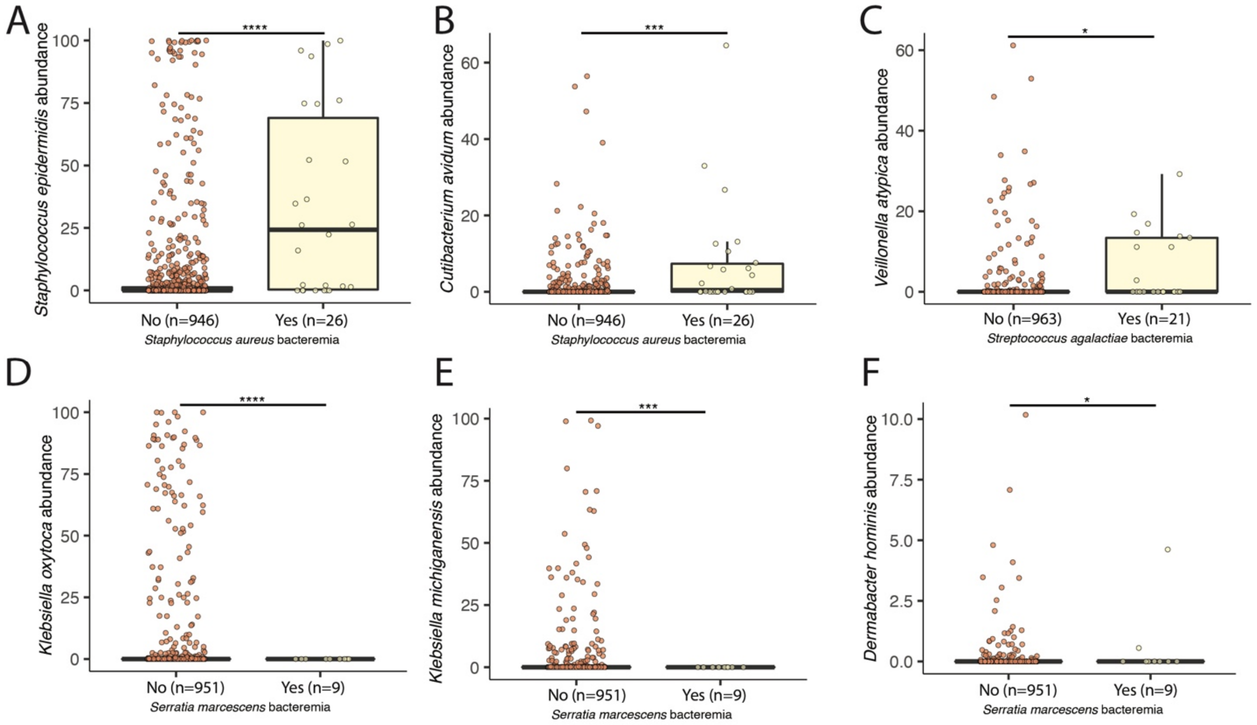

Figure S11. Gut microbiome species differentiate BSI cases from controls. Non-BSI causing species also differ between BSI caused by *S. aureus* (A-B), *S. agalactiae* (C), and *S. marcescens* (D-F). Generalized linear mixed effect models using MaAsLin2 with DOL as an additional fixed effect and participant as random effect. "Yes" indicates samples obtained prior to BSI from that organism and no indicates for participants without bacteremia from the causative organism. Stools after BSI were not included in this analysis.

The following data files are available online:

Data file S1. Sample and clinical metadata.

Data file S2. Statistical analyses results.

Data file S3. Isolate assembly statistics and antibiotic susceptibility and ARGs, clinical information and antibiotic histories.

Data file S4. inStrain mapping statistics for Figure 4 and isolate read-assembly self mapping.

Data file S5. inStrain mapping statistics for Figure 5.