

UC Berkeley

UC Berkeley Previously Published Works

Title

Evidence for persistent and shared bacterial strains against a background of largely unique gut colonization in hospitalized premature infants

Permalink

<https://escholarship.org/uc/item/1wn5d9zb>

Journal

The ISME Journal: Multidisciplinary Journal of Microbial Ecology, 10(12)

ISSN

1751-7362

Authors

Raveh-Sadka, Tali
Firek, Brian
Sharon, Itai
[et al.](#)

Publication Date

2016-12-01

DOI

10.1038/ismej.2016.83

Peer reviewed

Evidence for persistent and shared bacterial strains against a background of largely unique gut colonization in hospitalized premature infants

Tali Raveh-Sadka¹, Brian Firek², Itai Sharon¹, Robyn Baker³, Christopher T Brown¹, Brian C Thomas¹, Michael J Morowitz² and Jillian F Banfield¹

¹ Department of Earth and Planetary Science, UC Berkeley, Berkeley, CA, USA; ² Department of Surgery, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA and ³ Department of Pediatrics, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

Correspondence: JF Banfield, Department of Earth and Planetary Sciences, and Department of Environmental Science, Policy, and Man, 369 McCone Hall, UC Berkeley, Berkeley CA 94720, USA. E-mail: jbanfield@berkeley.edu

Abstract

The potentially critical stage of initial gut colonization in premature infants occurs in the hospital environment, where infants are exposed to a variety of hospital-associated bacteria. Because few studies of microbial communities are strain-resolved, we know little about the extent to which specific strains persist in the hospital environment and disperse among infants. To study this, we compared 304 near-complete genomes reconstructed from fecal samples of 21 infants hospitalized in the same intensive care unit in two cohorts, over 3 years apart. The genomes represent 159 distinct bacterial strains, only 14 of which occurred in multiple infants. *Enterococcus faecalis* and *Staphylococcus epidermidis*, common infant gut colonists, exhibit diversity comparable to that of reference strains, inline with introduction of strains from infant-specific sources rather than a hospital strain pool. Unlike other infants, a pair of sibling infants shared multiple strains, even after extensive antibiotic administration, suggesting overlapping strain-sources and/or genetic selection drive microbiota similarities. Interestingly, however, five strains were detected in infants hospitalized three years apart. Three of these were also detected in multiple infants in the same year. This finding of a few widely dispersed and persistent bacterial colonizers despite overall low potential for strain dispersal among infants has implications for understanding and directing healthy colonization.

Introduction

Gut microbes have important roles in health and disease. Colonization of the gut may begin *in utero* (Funkhouser and Bordenstein, 2013; Moles et al., 2013; Aagaard et al., 2014), but progresses rapidly after birth. According to 2012 data from the Center for Disease Control (Macdorman et al., 2013), 99% of US infants are born in hospitals, suggesting that very early stages of gut colonization occur in the hospital environment. We know very little about the extent to which hospital-associated microbial communities can influence

gut colonization. However, given that many pathogens are found in hospital settings, it is likely that acquisition of hospital-derived organisms can lead to aberrant colonization and long-term detrimental health effects (Arrieta et al., 2014). This is of particular importance for premature infants that are hospitalized for long periods, especially given frequent antibiotic administration at birth, their immature gastrointestinal tracts and the fact that their immune system is more vulnerable than that of term infants (Groer et al., 2014). Hospitalized children and adults, especially those with compromised immune systems, are also vulnerable to acquisition of hospital-associated pathogens, particularly those that persist in the room environment.

Some studies of hospital outbreaks have focused on a small number of specific strains of interest, typically bacterial pathogens (Chin et al., 2010; Köser et al., 2012; Snitkin et al., 2012; He et al., 2013; Loman et al., 2013), and tracked them among hospitalized individuals. However, an outbreak is a very specific phenomenon that involves both spread and infection by a single, potentially highly virulent organism. Consequently, lessons from outbreak studies provide only limited insight into the behavior of the larger consortia of hospital-associated bacteria that may be relevant for initial gut colonization. Numerous recent studies followed microbial communities colonizing hospitalized neonates over time (Morowitz et al., 2011; Brown et al., 2013; Costello et al., 2013; Sharon et al., 2013; Brooks et al., 2014; La Rosa et al., 2014; Raveh-Sadka et al., 2015; Sim et al., 2015; Zhou et al., 2015), providing a broader view of hospital-associated gut-colonizing bacteria. These studies often reveal high inter-patient variability, raising questions about reservoirs and sources of bacteria that establish in the gut environment. However, characterization of members of the microbial community is typically done with low taxonomic resolution (Arboleña et al., 2012; Barrett et al., 2013; Aujoulat et al., 2014; La Rosa et al., 2014), thus precluding the possibility of tracking and comparing individual microbial strains across hospitalized patients. Subsequently, we are still lacking a broad understanding of the extent to which gut colonizers are shared between neonates housed in the same environment.

De novo assembly of shotgun-sequencing data from fecal samples is a cultivation-independent method that provides community-wide strain-resolved view of early gut colonizers, including potentially important phage and mobile elements (Sharon et al., 2013), but until recently has only been applied to one (Morowitz et al., 2011; Brown et al., 2013; Sharon et al., 2013) or two (Brooks et al., 2014) hospitalized infants. Following continued reduction in sequencing costs and improvements in assembly and binning algorithms, we applied this technique to study the gut microbiome in a group of 10 premature infants hospitalized over 2 months in 2014 (Raveh-Sadka et al., 2015). This short-term study was conducted during a reported outbreak of necrotizing enterocolitis (NEC), a common and life-threatening gastrointestinal disease that primarily afflicts preterm infants in their first

weeks of life (Schnabl et al., 2008; Neu and Walker, 2011; Carlisle and Morowitz, 2013) and ruled out a single shared virulent strain as the source of disease as well as suggested that identical strains are rarely detected among co-hospitalized infants.

Here, we focus on strain sharing not only between co-hospitalized infants, but rather study questions of longer-term strain persistence and dispersal in the hospital environment by combining our strain-resolved analyses with long sampling periods. We analyze microbial communities in a total of 202 fecal samples from 21 infants, nine of which developed NEC. The data set includes 147 samples collected during the first month of life of a cohort of 11 preterm infants hospitalized in 2011–2012, and 55 samples from the second cohort of 10 infants hospitalized three years later, in 2014 (Raveh-Sadka et al., 2015). All infants were hospitalized in the neonatal intensive care unit (NICU) in Magee-Womens Hospital of the University of Pittsburgh. Our results reveal few persistent and dispersing bacterial colonizers against a background of largely unique gut colonization. Our approach could be used for routine monitoring to detect rapidly dispersing and potentially clinically relevant bacteria in the hospital environment.

Methods

Patients and sample collection

2011–2012 cohort—fecal samples from 11 preterm infants hospitalized in the NICU in Magee-Womens Hospital of UPMC (Pittsburgh, PA, USA) over a period of ~5 months in 2011–2012 were collected. Four of these infants developed NEC (defined as Bell’s Stage II or III disease). One infant (#17A) developed sepsis. Two deaths (Infants #17A and #21A) occurred among this cohort of infants.

2014 cohort—fecal samples from 10 infants hospitalized in the same unit were collected over 2 months in 2014. Five of these infants developed NEC.

Excepting one-term infant who developed NEC (Infant #9B), all infants were of low gestational age (<33 weeks) and low birth weight (<2500 g).

The study was performed with approval from the University of Pittsburgh Institutional Review Board as protocol number PRO10090089, and written parental consent was obtained on behalf of the neonates.

See Raveh-Sadka et al. (2015) and Supplementary Table S1 for clinical information.

Fecal samples were collected from spontaneously expelled diaper samples or using an established perineal stimulation procedure (Morowitz et al., 2011). Collected samples were placed promptly in a –20 °C freezer, and were transferred to –80 °C for long-term storage (few minutes to few days during weekends). Samples were collected daily, as available, during days of life 5–21, and then weekly while hospitalized in the NICU. For the 2011–2012 cohort, a subset of 147 samples yielding adequately high quantity and

quality of genomic DNA was subjected to shotgun sequencing. For the 2014 cohort 55 samples were sequenced (see detailed description in Raveh-Sadka et al. (2015)).

DNA extraction

Genomic DNA was extracted from fecal samples using the MO BIO PowerSoil DNA Isolation kit. Samples were added directly into bead tubes and incubated at 65 °C for 10 min followed by 95 °C for 10 min. After addition of 60 µl of Solution C1 the bead tubes were then shaken horizontally on a lab mixer for 10–16 min at maximum speed using a MO BIO vortex adaptor. All remaining steps followed the manufacturer's protocol.

Sequencing and assembly

Sample preparation and sequencing were performed at the sequencing unit in the University of Illinois at Urbana-Champaign. Reads were trimmed using the Sickle software (Joshi and Fass, 2011) and Bowtie2 (Langmead and Salzberg, 2012) was used to map and remove reads mapping to the human genome.

For the 2014 cohort, DNA from 55 samples was prepared for paired-end sequencing using Illumina's Nextera kit with average insert sizes of 250–500 bp. Samples were sequenced on four lanes of a HiSeq2500 platform. Trimmed reads (2.2–7.3 Gbp per sample) were assembled using `idba_ud` (Peng et al., 2012). For more details see Raveh-Sadka et al. (2015).

For the 2011–2012 cohort, DNA from 147 samples was prepared for paired-end sequencing using Illumina's Nextera kit with average insert sizes of 500/900 bp. Samples were sequenced on 13 lanes of a HiSeq2000 platform. In total, ~5.14 billion high-quality reads remained after trimming and discarding trimmed reads of length <60 bp (93.9% of 5.5 billion total raw reads).

Sequence reads were assembled using a coverage-based iterative approach that assembles differentially abundant genomes separately. With this approach, the parameters for the assembly of different genomes can be optimized for each coverage bin, similar to the approach described in Sharon et al. (2013). At each iteration a specific sample was selected and reads were assembled using Velvet (Zerbino and Birney, 2008), where assembly parameters were chosen to optimize the assembly for the highest coverage bin still remaining in this sample. Resulting scaffolds longer than 500 bp that fit the range of the targeted coverage bin were added to the assembly, and reads mapping to these scaffolds (using Bowtie2) were removed from further analysis. When there were no evident coverage peaks left in any single sample, remaining reads from all samples of the same infant were combined and the process was reiterated.

Assembly gaps, represented by Ns added in the scaffolding process, were resolved when possible, by reassembling reads that map close to the gap edges, along with their paired reads (Sharon et al., 2013).

Potential mis-assemblies (assembly and scaffolding errors) were detected using several approaches: (1) identification of significantly long regions with zero-insert coverage and regions in which coverage changed markedly (Sharon et al., 2013) and (2) detection of regions with high frequency of mate-pair problems (small/large inserts, mis-orientation of paired reads, mapping to different scaffolds). Scaffolds were split at positions in which mis-assembly was suspected.

When possible, scaffolds were extended and merged if paired-end information consistently suggested that they were linked (the paired reads for reads mapping near edges of one scaffold mapped consistently to a second scaffold). The read pairs mapping near the edges of such scaffold pairs were re-assembled and the resulting assemblies were merged with the original scaffolds using minimus2 (Sommer et al., 2007).

Scaffold annotations

Scaffolds >1000 bp were annotated. We used prodigal (Hyatt et al., 2010) (v2.60; available at <https://github.com/hyatt/Prodigal/releases/tag/v2.60>) to predict genes using default settings for metagenomics gene prediction. Protein sequences were searched against KEGG (Kanehisa et al., 2014) (KEGG FTP Release 2014-07-07; available at <http://www.kegg.jp/kegg/download/>), UniRef100 (release 2014_07; available at ftp://ftp.uniprot.org/pub/databases/uniprot/previous_releases/release-2014_07/), and UniProt (Leinonen et al., 2004) (same as UniRef) using USEARCH (Edgar, 2010) (v7.0.1001; available at <http://www.drive5.com/>). In addition, reciprocal best-blast hits were determined. All matches with bit scores >60 were saved, and reciprocal best hits with a bit score >300 were also cataloged. In total, 16S ribosomal RNA (rRNA) gene sequences were identified based on hidden Markov model searches using the cmsearch program (cmsearch -hmmonly -acc -noali -T -1) from the Infernal (Nawrocki and Eddy, 2013) package (v1.1; available at <http://infernal.janelia.org/>). To identify 16S rRNA gene sequences, all assembled contigs were searched against the manually curated structural alignment of the 16S rRNA provided with SSU-Align. Transfer RNA (tRNAs) were predicted using tRNAscan_SE (Lowe and Eddy, 1997) (v1.23-r2; available at <http://lowelab.ucsc.edu/tRNAscan-SE/>). Scaffolds, gene predictions and all associated annotations were uploaded to ggKbase.berkeley.edu (http://ggkbase.berkeley.edu/project_groups/necevent_samples and http://ggkbase.berkeley.edu/project_groups/nih-year-1).

Binning

2011–2012 cohort—Scaffolds were clustered into bins representing genomes using the Databionic implementation of ESOM (Emergent Self Organizing Maps) (Ultsch and Moerchen, 2005) based on their normalized read coverage pattern over time (Sharon et al., 2013) as well as their GC content. Each scaffold was broken into 3kbp segments and the coverage for each segment was calculated separately. Coverage values were normalized by the total number of reads in each sample as well as by the sum of coverage values for each segment across all time points, in order to eliminate large values that might skew the output.

In cases where there were not enough time points to allow clustering based on coverage profiles (Infants #8A and #17A), clustering was done based on nucleotide tetramer frequencies in each scaffold (Dick et al., 2009).

Parameters for clustering were chosen as previously described (Dick et al., 2009), and clusters were extracted manually from the resulting maps (similar to the approach described in Sharon et al. (2013)).

Binning was further refined in ggKbase manually based on a specific phylogenetic profile, and/or scaffold coverage and/or GC information (Raveh-Sadka et al., 2015). ggKbase was also used to assess the completeness of the resulting genome bins based on the number of predicted single-copy genes (Sharon et al., 2013). The abundance of genomes over time relative to the total number of reads mapped to all binned scaffolds was calculated based on the median coverage of scaffolds in the corresponding bin at each time point.

Inspection of the abundance profiles over time revealed four cases (Infant #8A day of life (DOL) 28, Infant #19A DOL 13, Infant #21A DOL 15 and Infant #23A DOL 13) in which a sample swap event was suspected. Unlike other data points, these four samples showed species composition that was not supported by neighboring time points or by 16S data for the same sample (data not shown), and in fact seemed to fit in well with data from one of the other three infants (for example, the sample labeled as Infant #23A DOL 13 is most likely from Infant #21A DOL 15). All of these four samples were sequenced on the same sequencing lane. We excluded these time points from downstream analyses.

For the 2014 cohort, online binning tools in ggKbase were used to bin samples based on coverage, GC content and taxonomy. Bins were validated using ribosomal gene and single-copy gene counts as well as using time series and tetra nucleotide frequency ESOMs. For more details see Raveh-Sadka et al. (2015).

Phylogenetic profile

Taxonomic assignments for each scaffold were determined using ggkbase online binning tools (ggkbase.berkeley.edu) based on the similarity of each of the predicted proteins in the scaffold to proteins in the UniProt database. A taxonomic label was assigned if >50% of predicted proteins had best

matches to the same taxon. For more detailed description, see Raveh-Sadka et al. (2015).

Comparison of assemblies from all time points with those obtained from single samples

To verify that the two approaches yield similar assemblies, we selected two organisms that were present in a single infant for a relatively long period of time and assembled their genomes from reads in single time points. Specifically, this was done for *E. coli* in infant #8A, DOL 7 and 31 and *Enterobacter* in infant #11A DOL 13, 51. In both cases, the assemblies recovered from single time points were over 98% identical across >99% of their length, way above the threshold for defining them as the same 'strain types'.

Strain comparison

All bacterial genomes from the same genus (or in the same Family/Order in the rare cases of relatively novel genomes whose genus could not be determined) that were fairly complete (bin length >0.5 Mbp, >50% (25/51) single-copy genes detected), and were not very fragmented (N50>10 Kbp) were compared with each other using nucmer (Delcher et al., 2002). All scaffolds from one genome bin were aligned against all others (using parameters -maxmatch -c 100 -b 100). For each scaffold, the number of basepairs contained in alignments with >98% identity were counted. Basepairs that were aligned to more than one scaffold were only counted once, and the fraction of such aligned basepairs from all basepairs in the genome bin was computed. If for one of the genomes in the compared pair, >95% of basepairs were >98% identical, the genomes were considered as belonging to the same 'strain type'.

Mutation rate analysis for persistent strains

Breseq (Deatherage and Barrick, 2014) was used to map reads from one infant in one cohort to the closely related genome assembled from an infant in the second cohort. Single nucleotide polymorphisms (SNPs) in scaffolds >3000 bp were extracted. SNPs in problematic regions of the original assemblies (containing Ns and so on) were discarded. Total number of remaining SNP was used to evaluate whether strains could have diverged in the period between 2011 and 2014 using bounds on expected number of mutations based on a possible range of doubling times and mutation rates (Hendricks, 1972; Guttman and Dykhuizen, 1994; Fossum et al., 2007; Moran et al. 2009; Lee et al., 2012).

Analysis of pathogenicity and metabolic potential

Analysis of metabolic and pathogenic potential was done in ggKbase. ggKbase annotates each predicted protein using its best sequence matches in well-known databases (KEGG (Kanehisa et al., 2014), UniRef and UniProt (Leinonen et al., 2004)). Functional profiling of genomes is then performed

based on searches for specific annotation terms (for example, terms like 'macrolide' are used to when functionally profiling antibiotic resistance genes). The number of genes annotated for a list of selected annotation terms is displayed. The list function in ggKbase can be used to display the annotation terms used to define each functional group. In addition to functional analysis based on gene annotation, for analysis of antibiotic resistance the CARD (McArthur et al., 2013) database was used, and provides breakdown of antibiotic resistance genes based on their type.

Phylogenetic trees

For each species of interest, four single-copy universal genes (*gyrase A*, *recombinase A*, *Histidyl tRNA synthetase* and *Phenylalanyl tRNA synthetase subunit alpha*) were used to construct phylogenetic trees. Draft or complete genomes in which all of these genes were annotated on the IMG database (Markowitz et al., 2014) or in assemblies were used. For *E. coli*, owing to the large number of sequenced strains, only complete genomes were used. Genomes in which these genes were truncated or missing were discarded. Gene sequences were aligned using MAFFT V7.050b with default parameters and the aligned sequences were concatenated, In case of duplicated reference sequences, a single genome was randomly chosen for analysis. Trees were constructed using RAxML (substitution model GTRGAMMA) on concatenated and aligned sequences.

Branches with bootstrap values of over 70 (out of 100) are highlighted.

Results

Samples, assemblies and binning

To augment the published collection of 55 metagenomic samples from the 2014 cohort of 10 hospitalized premature infants (Infants #1B-10B, average of six samples per infant) (Raveh-Sadka et al., 2015), sequence data sets were generated for fecal samples collected from a second cohort of eleven preterm infants (<2100 g at birth, <33 wk gestation) hospitalized over the course of ~5 months during 2011-2012 (see Supplementary Table 1 for clinical details, Supplementary Figure 1 for timeline). Four of the eleven infants developed NEC in addition to five infants from the 2014 cohort (Supplementary Table 1). All 21 infants were housed in the same NICU in the period 2011-2014.

Samples from the 2011-2012 cohort were collected daily, as available, during days of life 5-21, and then weekly while hospitalized in the NICU (consistent with the approach used in the prior study). A subset of 147 samples (average of 13 samples per infant) with sufficient quantities of extracted DNA was selected for shotgun-sequencing on 13 lanes of Illumina HiSeq2000 (Supplementary Figure 1). We generated a total of 5.5 billion 100-bp raw paired-end reads (93% of which were of high quality) with an average of ~50 Gbp per infant and ~3.71 Gbp per sequenced sample (data sets comparable to those used in the prior study where 2.2-7.3 Gbp per sample

were generated). Given that the prior study (Raveh-Sadka et al., 2015) indicated that strains tend to persist in each infant over time, a single assembly was generated for each infant using a coverage-based iterative approach (Sharon et al., 2013) (see Methods). We verified that this approach generates genome assemblies highly similar to those obtained from single sample assemblies on test cases (see Methods). An average of 94% of all high-quality reads from each sample mapped back to the corresponding assembly (Supplementary Table 2). For the entire set of samples from 2011–2012, 490 Mbp of genome sequence was reconstructed on scaffolds >2 kb. Scaffolds were clustered into bins representing genomes based on their read coverage pattern over time (Sharon et al., 2013), as well as their GC content. Online binning tools (Raveh-Sadka et al., 2015) were used to further curate binning and to evaluate the completeness and purity of each bin, based on taxonomic profiles, bin length and predictions for the presence of universal single-copy genes (See Methods). A total of 103 near-complete (>94% of single-copy genes detected) high-quality draft genomes and 51 partial genomes (including a partial genome for the Eukaryote *Candida albicans*), as well as 29 complete or partial mobile elements (that is, plasmid, phage and transposons) were recovered (see Supplementary Tables S3 and S4 for detailed statistics on the genomes assembled for each infant; Supplementary Figure 2 for high-level bioinformatic pipeline).

Species composition over time supports lack of microbial agent responsible for NEC

While species composition varied considerably among infants (Figure 1, Supplementary Figures S3 and Supplementary Table S5), species diversity within individual samples was remarkably low (Supplementary Table S3), in accordance with previous reports (Magne et al., 2006; Wang et al., 2009; Morowitz et al., 2011; Raveh-Sadka et al., 2015). On average in the 2011–2012 cohort, only six genomes with relative abundance >0.1% were detected in each sample, and in some samples, one or two genomes accounted for as much as 99% of the relative abundance (for example, in Infant #11A, 46th DOL, Figure 1). Species diversity was not significantly different in infants that developed NEC either when comparing mean Shannon index in all collected samples (t -test p -value 0.31) or when limiting comparison to samples collected prior to NEC diagnosis (t -test p -value 0.32) (Supplementary Table S3).

In terms of taxonomic classification, the preterm infant gut communities from the 2011–2012 cohort were dominated by a small group of species from the Proteobacteria (mostly Enterobacteriaceae genera *Klebsiella*, *Escherichia* and *Enterobacter*) and Firmicutes (mostly *Clostridium*, *Veillonella*, *Staphylococcus* and *Enterococcus*) phyla. Actinobacteria (mostly *Propionibacterium*) and Bacteroidetes (*Bacteroides*) were also present in some communities but at lower abundances. No single common species was detected (or was notably absent) in all nine NEC cases in comparison with the 12 healthy infants (including *Clostridium* (De la

Cochetiere et al., 2004) and *Propionibacterium* (Morrow et al., 2013) species that were previously implicated with NEC). Abundant species detected in NEC cases were different from the ones detected in the 2014 cohort (Raveh-Sadka et al., 2015). Moreover, the NEC diagnoses were not immediately preceded by marked changes in bacterial community composition.

Interestingly, in all cases for which medical records indicated positive microbiology cultures from either blood or tracheal aspirates (two cases in NEC infants #17A and #21A, and an additional two in Infant #19A; Figure 1, Supplementary Figure S3, Supplementary Table S1) almost all of the isolated bacterial pathogens, with the exception of a *Streptococcus viridans* in Infant #19A, were observed in stool before the culture was taken.

Antibiotic treatments correlate with shifts in microbial community composition

Almost all premature infants received 2 days of empiric ampicillin/gentamicin antibiotic treatment immediately after birth. In addition, depending on their medical condition, infants received additional courses of antibiotics, most often ampicillin/gentamicin or vancomycin/cefotaxime combined treatments. On average infants received 9 day of antibiotic treatment during the sampling period (Supplementary Figure S1, Supplementary Table S1).

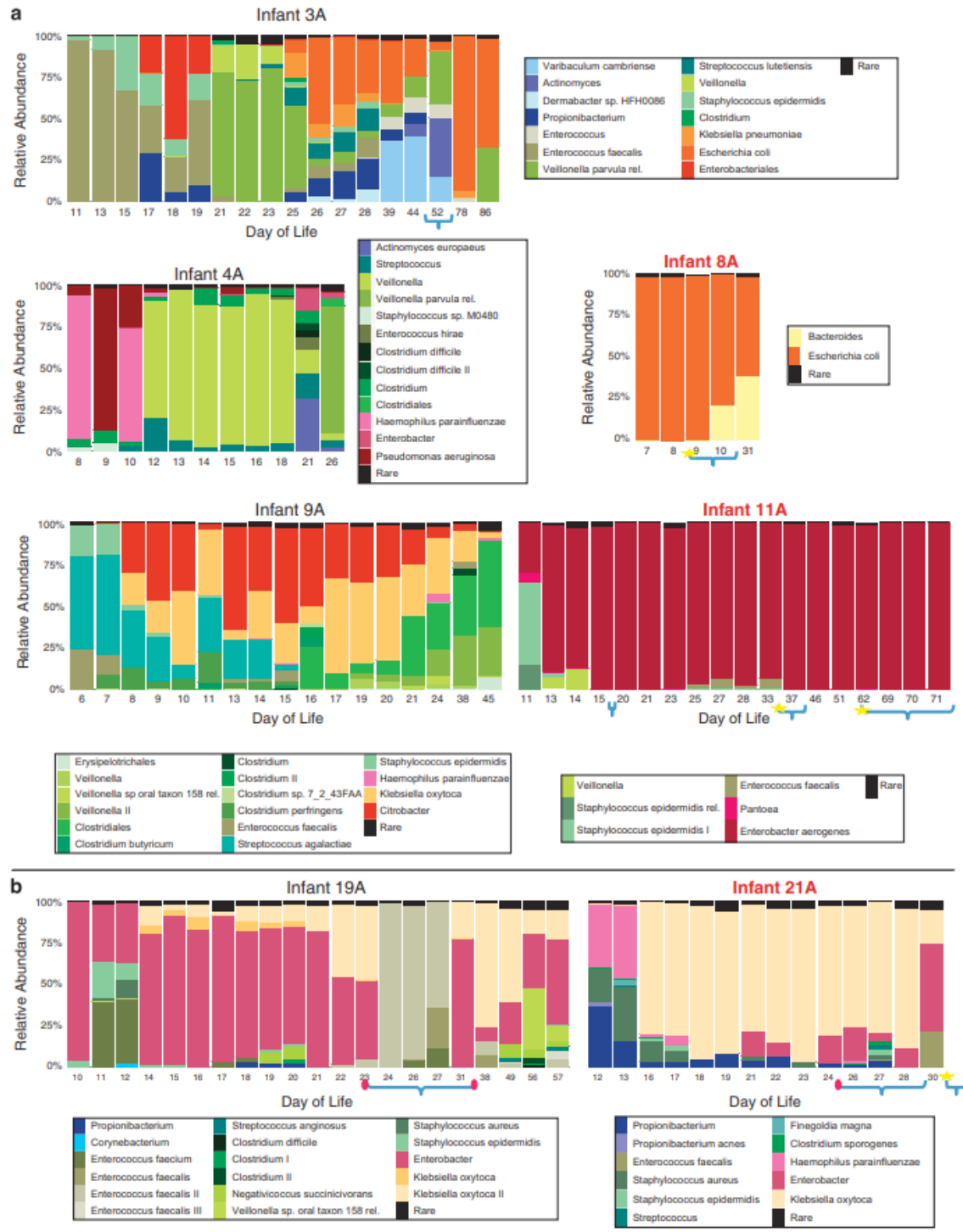


Figure 1 Bacterial community composition for (a) Infants #3 A, #4 A, #8 A, #9 A and #11 A (b) sibling Infants #19 A and #21 A (relative to each infant's birth date) from the 2011–2012 cohort. Red font indicates infants that were diagnosed with NEC. Curly brackets indicate periods in which antibiotics were administered (incomplete brackets indicate antibiotic treatment that extended outside of sampling period), stars indicate days in which NEC was diagnosed, red ovals indicate dates in which additional positive cultures (blood/tracheal aspirate) were available. In each time point, the abundance of all organisms with relative abundance <2% is summed and labeled as 'Rare' (see Supplementary Table S5 for the data underlying relative abundance plots).

A few distinct configurations of bacterial community composition could be defined in the time series for each baby (Figure 1, Supplementary Figure

S3,Supplementary Table S5). Transitions between these configurations were often, but not always, gradual. In a few cases, the most abundant species changed within a couple of days. Although in some cases the reason for large composition changes was not apparent (for example, in the 2011–2012 cohort—Infant #3A DOL 21, Infant #4A DOL 12, Infant #14A DOL 14), in other cases shifts in the bacterial community coincided with known external factors, mostly related to antibiotic administration (for example, Infant #19A DOL 24, Infant #23A DOL 8). Antibiotic administration typically decreased species richness (Infant #3A DOL 78, Infant #19A DOL 24), and in one case (following the administration of cefotaxime, vancomycin and cefepime antibiotics) resulted in a massive enterococcal bloom (Infant #19A). Patterns of recovery from antibiotics were not consistent between babies, and ranged from quick and almost complete recovery (Infant #19A) occurring over a few days to cases in which recovery was not observed in the studied time frame (Infant #3A).

Microbial diversity across infants and over time reveals little strain sharing, with notable exceptions

To gain better understanding of the background processes of strain dispersal and diversification in the NICU, we combined data from the 2014 cohort with that of the 2011–2012 cohort. In total, 304 genomes with >500 Kbp of recovered sequence per genome, that were at least 50% complete (see Methods) and not extremely fragmented ($N_{50} > 10$ Kbp) were analyzed. To measure genome similarity for related organisms, we calculated the percent of genome length that was >98% identical for pairs of genomes in the same genus (See Methods; Figure 2, Supplementary Figure S6). Genomes for which >95% of genome length was >98% identical were designated as 'strain types'. This measure is consistent with that used previously for the 2014 cohort (Raveh-Sadka et al., 2015), showing that genomes recovered from different samples of the same infant exhibit an extremely high level of similarity, indicating persistence of strains in the gut of individual infants over time. The 304 genomes from the combined cohort represent 159 strain types. In total, 145 of these were represented by genomes from a single infant. Only 14 strain types included genomes recovered from different infants. Of the strain types with genomes from multiple infants, three were found exclusively in the 2014 cohort, five were detected in both the 2011–2012 and the 2014 cohort, and six were found exclusively in the 2011–2012 cohort.

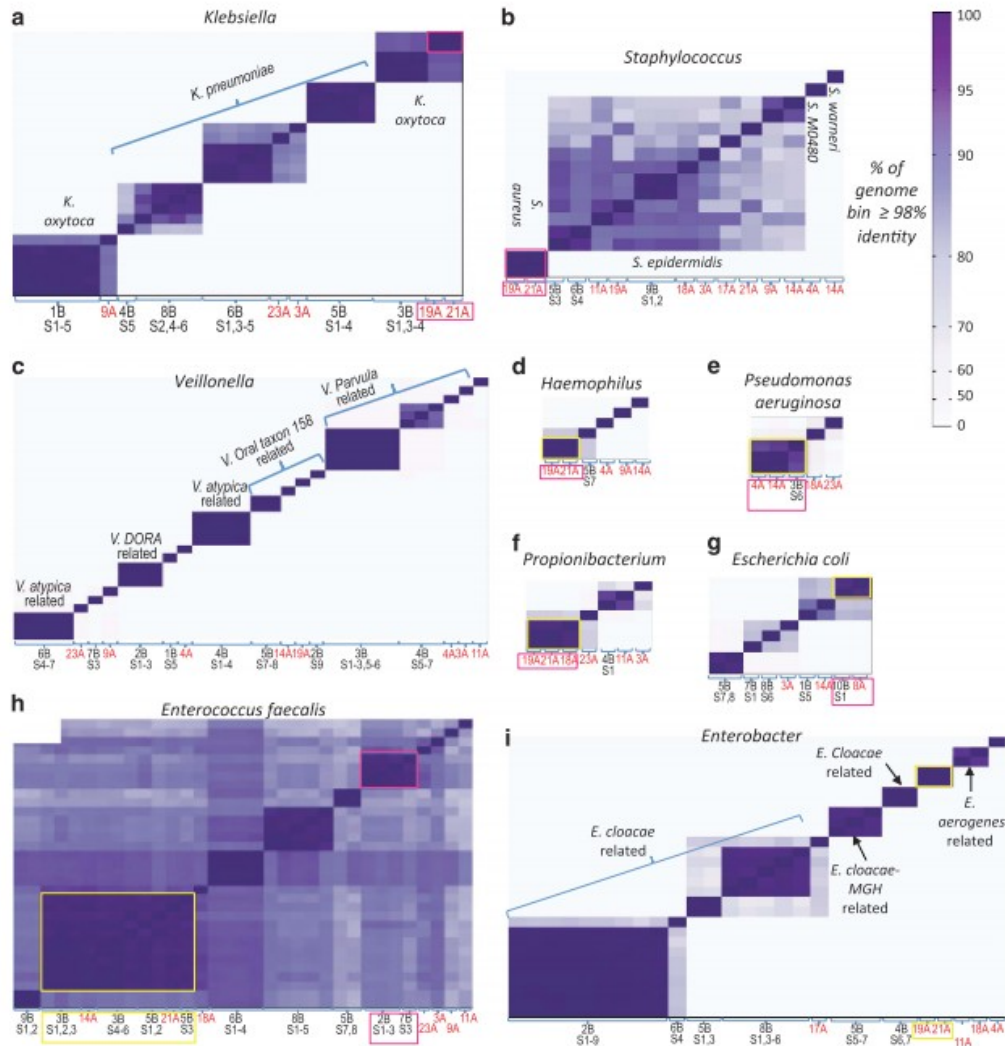


Figure 2 Pairwise similarity of high-quality genomes. Each column on the x axis (and each row on the y axis) represents a reconstructed genome from the taxonomic group showed in each panel. The intensity of color in cell (i, j) of the matrix corresponds to the similarity of genome i and genome j, where similarity was defined to be the fraction of genome length, that is, >98% identity. For each scaffold, the number of basepairs contained in alignments with >98% identity were counted. Strains where >95% of total genome basepairs were >98% identical, were considered as belonging to the same 'strain type'. The ordering of genomes is the same on the x and y axis and was determined by clustering the genomes according to their similarity to all other genomes. Strain types that were recovered from more than one infant are highlighted with boxes. X axis shows Infant number (and samples number for the 2014 cohort) from which the corresponding genome was recovered. Genomes recovered from the 2011–2012 cohort are marked with red font. (a–h) comparisons of strains of (a) *Klebsiella* (b) *Staphylococcus* (c) *Veillonella* (d) *Haemophilus* (e) *Pseudomonas aeruginosa* (f) *Propionibacterium* (g) *Escherichia coli* (h) *Enterococcus* (i) *Enterobacter*.

Extraordinary number of shared strains detected in a pair of sibling infants

Remarkably, four of the six strain types found in more than one infant but exclusively in the 2011–2012 cohort consisted of strains (*Enterobacter*, *Klebsiella oxytoca*, *Haemophilus influenzae*, *Staphylococcus aureus*) shared solely between the only pair of siblings in this study (two siblings, Infants #19A and #21A, from a fraternal triplet. For the third infant in this triplet, Infant #20A, no good samples were collected and it was thus excluded from our study). The fifth type included *Propionibacterium* strains shared between

the number of identified SNPs between the recovered genomes from each strain type. For four of the five strains detected in both cohorts, the number of identified SNPs was within the range of what could be expected for a 3-year period of *in situ* diversification according to published mutation rates (Hendricks, 1972; Guttman and Dykhuizen, 1994; Fossum et al., 2007; Moran et al., 2009; Lee et al., 2012) in bacteria (Supplementary Table S7; see Methods). For the 5th strain—the *E. coli* strain, mutation analysis indicated earlier divergence. However, bounds on mutation rates and bacterial doubling times provide extremely loose ranges of possible divergence times that are far from being conclusive.

To gain additional insight on the possible sources of persistent strains in the hospital environment, we constructed phylogenetic trees comparing persistent strains to reference genomes from the IMG database, based on concatenated alignment of *gyrase A*, *recombinase A*, *Histidyl tRNA synthetase* and *Phenylalanyl tRNA synthetase subunit alpha* single-copy genes (see Methods). For *E. coli*, *P. aeruginosa* and *E. faecalis*, the trees identified reference strains whose nucleotide sequence for the four studied genes was identical to that of the assembled genomes (Supplementary Figures S7–S10). In these cases whole-genome alignments were used to better assess the overall genome similarity. For the persistent *E. coli* strain, whole-genome comparisons showed that 96.3% of the reconstructed genome was >98% identical to JJ1886 clinical reference strain. However, despite high similarity, the persistent *E. coli* strain did not contain the pathogenic island reported for JJ1886. Similarly, 95.5% of the genome of persistent *P. aeruginosa* was >98% identical to CI27 clinical strain, and 95% of persistent *E. faecalis* was >98% identical to strain TX4248.

Interestingly, in the case of *E. coli*, genome similarity to reference genome was on par with similarity for persistent genomes recovered 3 years apart (96.9% of genome >98% identical). Combined with the mutation rate analysis, this result suggests that the *E. coli* strain may not be diversifying *in situ*, but rather that a similar *E. coli* strain was reintroduced to the hospital 3 years later.

Unlike other persistent strains, persistent *E. faecium* and *Clostridium parautrificum* were not extremely similar to sequenced reference genomes (with 86% of persistent strain genome >98% identical to closest *E. faecium* reference genome strain e4452 and, and 88% of persistent *C. parautrificum* >98% identical to the only available reference genome AGR2156), thus supporting *in situ* diversification for these strains.

Notably, in three out of the five cases, strains that were found in both cohorts were also found in multiple infants in one of the cohorts (Figure 3), suggesting their relative prevalence and persistence in the hospital environment. These persistent strain types varied in their pathogenic capacity (Figure 4a). Moreover, functional analysis of mobile elements and antibiotic resistance failed to identify features that distinguished the

persistent strains from genomes of other strains in the same species in a consistent manner (Methods; Figure 4b). Thus, it is still unclear what in these strains make them better adapted to the hospital environment, enabling them to persist and disperse. Accumulation of additional samples and genomes of persistent and non-persistent bacteria could potentially allow for the detection of subtle distinguishing features.



Figure 4 Pathogenicity potential of persistent strains. (a) pathogenicity potential for all persisting strains (note variability) based on gene annotations. For each genome, shown is the total number of genes in each category (see Methods) (b, c) pathogenicity potential of persistent (purple highlighting) *Escherichia coli* (b) and *Pseudomonas aeruginosa* (c) as well as other fairly complete genomes of the same species that were recovered based on gene annotations (top) and based on antibiotic resistance genes as detected by CARD (McArthur *et al.*, 2013). For CARD annotations, shown in each circle are the number of reported genes, the genome size, and color coded circle shows the breakdown of reported genes based on antibiotic resistance gene types (see color legend on the right).

Extensive strain diversity of *E. faecalis* and *Staphylococcus epidermidis* is likely due mostly to colonization by bacteria from infant-specific sources

Unlike for most species, many closely related yet genomically distinct strain types were recovered for *E. faecalis* and *S. epidermidis* (Figure 2). A total of 11 *E. faecalis* strain types in 14 out of 21 infants were identified from the set of compared genomes, and all of them were >98% identical across at least 78% of their genome lengths (mean $87\pm 4\%$). For *S. epidermidis* 10 strain types were identified in 11 infants, and all of them were >98% identical across at least 72% of their genome lengths (mean $84\pm 6\%$). In comparison, for example, for *E. coli*, seven strain types were identified in eight infants, but 11–90% of the length of compared genomes were >98% identical (mean $40\pm 28\%$).

To assess whether this diversity is largely the result of colonization by bacteria from infant-specific sources or a limited and diversifying pool of hospital bacteria, we inferred phylogenetic trees for both *E. faecalis* and *S. epidermidis* (see Methods) using all of the recovered genomes, as well as for reference genomes for which these genes were annotated in the IMG database (Markowitz et al., 2014) (Figure 5, Supplementary Figures 7 and 11). The sequences from strains recovered from the hospitalized infants are phylogenetically distributed across the tree and exhibit overall diversity that is comparable to that of the reference genomes. This suggests that although groups of persistent and relatively prevalent hospital strains do exist (like the persistent *E. faecalis* strain type discussed above), the major contributor to the observed diversity in these species is likely colonization by bacteria from infant-specific sources rather than *in situ* diversification of a small set of hospital-associated bacterial strains.

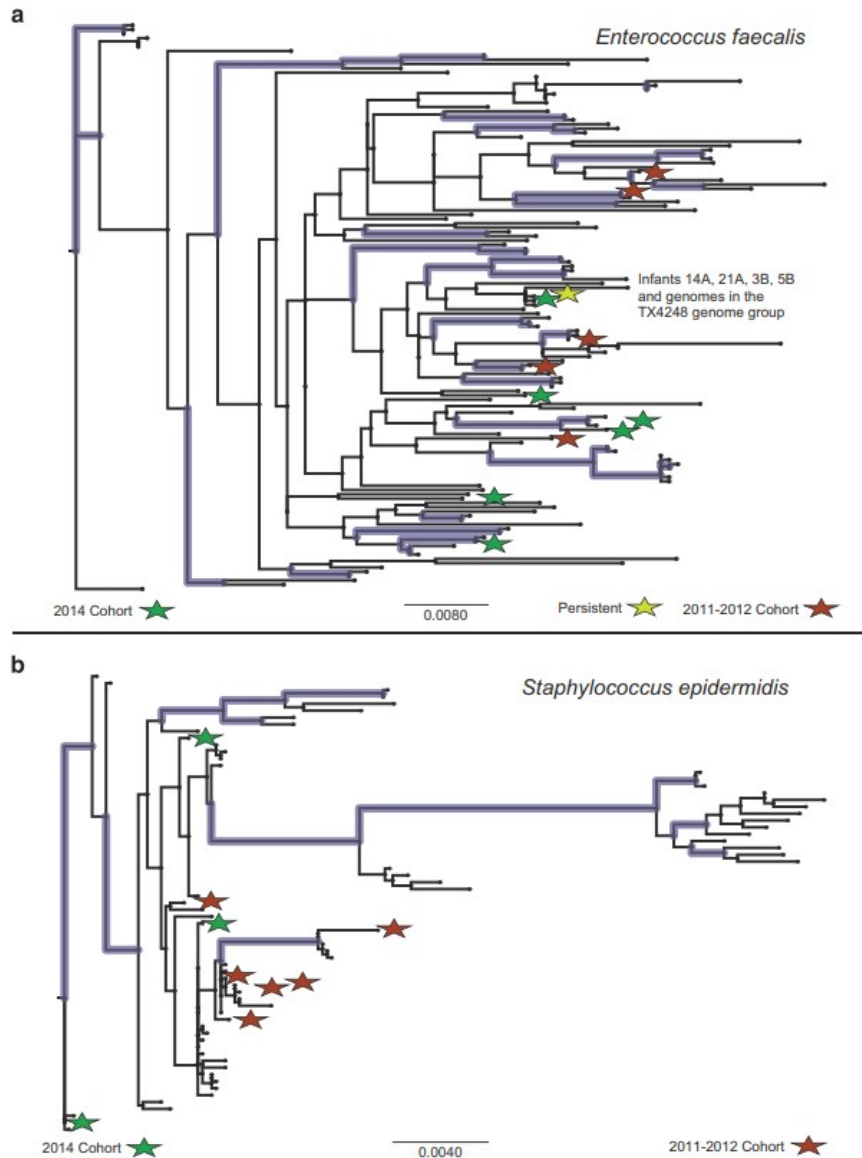


Figure 5 Schematic phylogenetic tree for concatenated genes of gyrase A, recombinase A, Histidyl tRNA synthetase and Phenylalanyl tRNA synthetase subunit alpha for (a) *Enterococcus faecalis* and (b) *Staphylococcus epidermidis*. Genomes recovered in this study are marked with red (2011–2012) and green (2014) stars. Branches with > 70% bootstrap values are colored (blue).

Discussion

In this study we tracked microbial communities colonizing the gastrointestinal tracts of premature infants hospitalized in the same NICU over a 3-year period. Generally, bacterial strains showed low potential for dispersal, with the vast majority of colonizing bacteria being unique to each infant, in accordance with prior analysis on a subset of these data. In particular, infants from the 2011–2012 cohort that were diagnosed with NEC did not share bacterial strains, supporting the observation made for the 2014 cohort regarding the lack of a single infecting pathogen that is responsible for NEC.

Genome similarity analysis of the relatively common infant gut colonizers, *S. epidermidis* and *E. faecalis*, point to gut colonization by bacteria from infant-specific sources rather than by bacteria diversifying from a limited pool of hospital strains.

We were able to detect 14 strain types that were shared among individuals, but notably, five of those were shared only by a pair of siblings, suggesting the importance of genetic selection and/or shared sources of inoculation. Extending this type of analysis to additional identical and non-identical twin/triplet sets and comparing those with sets of unrelated infants that are hospitalized in close proximity to one another could potentially decouple the possible contributors to the observed strain sharing.

Five additional strain types persisted over 3 years. Three out of these five types were identified in multiple infants from the same cohort, indicating that although most bacterial strains did not tend to disperse, some strains that dispersed widely persist in the NICU. These strains were not found to differ significantly in their virulence, antibiotic resistance or metabolism from non-persisting strains. The study of larger cohorts over longer time periods and improvement in the resolution and accuracy of functional gene annotations could potentially allow the detection of more subtle functional features that may be under/over represented in persistent taxa. In addition, isolating persistent strains and studying them in culture could potentially reveal additional traits that confer a fitness advantage within the hospital environment.

In this work we do not explicitly discuss or track the abundance of mobile elements, despite the fact that multiple partial and complete mobile elements were recovered from the samples and were even highly abundant in some. This is partly because many of these elements can be found both inside and outside their host genomes, making it more challenging, especially in the case of partial recovery, to quantify their abundance and track it between samples and infants. However, spatio-temporal dynamics of phage and plasmids can be highly relevant in the context of nosocomial infections, given that often these elements harbor antibiotic resistance genes and other genes of interest.

Overall, the approach presented here provides a way to monitor, with basepair resolution, microorganisms in the hospital environment over time and to distinguish persistent from newly introduced strains. With the accumulation of such data we can begin to think of planned manipulations to this environment in a way that will promote the dispersal of health promoting bacteria and hinder the proliferation of pathogens. In this sense, our approach may represent an important step towards understanding, and ultimately controlling, early gut colonization in hospitalized premature infants.

Data access

Read data sets have been deposited in NCBI sequence read archive and are available under bioproject PRJNA294605.

References

- Aagaard K, Ma J, Antony KM, Ganu R, Petrosino J, Versalovic J . (2014). The placenta harbors a unique microbiome. *Sci Transl Med* 6: 237ra65–237ra65.
- Arboleya S, Binetti A, Salazar N, Fernández N, Solís G, Hernández-Barranco A *et al.* (2012). Establishment and development of intestinal microbiota in preterm neonates. *FEMS Microbiol Ecol* 79: 763–772.
- Arrieta M-C, Stiemsma LT, Amenyogbe N, Brown EM, Finlay B . (2014). The intestinal microbiome in early life: health and disease. *Front Immunol* 5: e-pub ahead of print, doi: 10.3389/fimmu.2014.00427.
- Aujoulat F, Roudière L, Picaud J-C, Jacquot A, Filleron A, Neveu D *et al.* (2014). Temporal dynamics of the very premature infant gut dominant microbiota. *BMC Microbiol* 14: 2320.
- Barrett E, Kerr C, Murphy K, O’Sullivan O, Ryan CA, Dempsey EM *et al.* (2013). The individual-specific and diverse nature of the preterm infant microbiota. *Arch Dis Child Fetal Neonatal Ed* 98: F334–F340.
- Brooks B, Firek BA, Miller CS, Sharon I, Thomas BC, Baker R *et al.* (2014). Microbes in the neonatal intensive care unit resemble those found in the gut of premature infants. *Microbiome* 2: 1.
- Brown CT, Sharon I, Thomas BC, Castelle CJ, Morowitz MJ, Banfield JF . (2013). Genome resolved analysis of a premature infant gut microbial community reveals a *Varibaculum cambriense* genome and a shift towards fermentation-based metabolism during the third week of life. *Microbiome* 1: 30.
- Carlisle EM, Morowitz MJ . (2013). The intestinal microbiome and necrotizing enterocolitis. *Curr Opin Pediatr* 25: 382–387.
- Chin C-S, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR *et al.* (2010). The origin of the haitian cholera outbreak strain. *N Engl J Med* 364: 33–42.
- Costello EK, Carlisle EM, Bik EM, Morowitz MJ, Relman DA . (2013). Microbiome assembly across multiple body sites in low-birthweight infants. *MBio* 4: e00782–13.
- De la Cochetiere M-F, Piloquet H, des Robert C, Darmaun D, Galmiche J-P, Roze J-C . (2004). Early intestinal bacterial colonization and necrotizing enterocolitis in premature infants: the putative role of *Clostridium*. *Pediatr Res* 56: 366–370.
- Deatherage DE, Barrick JE . (2014). Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Methods Mol Biol* 1151: 165–188.

Delcher AL, Phillippy A, Carlton J, Salzberg SL . (2002). Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* 30: 2478-2483.

Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP *et al.* (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biol* 10: R85.

Edgar RC . (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26: 2460-2461.

Fossum S, Crooke E, Skarstad K . (2007). Organization of sister origins and replisomes during multifork DNA replication in *Escherichia coli*. *EMBO J* 26: 4514-4522.

Funkhouser LJ, Bordenstein SR . (2013). Mom knows best: the universality of maternal microbial transmission. *PLoS Biol* 11: e1001631.

Groer MW, Luciano AA, Dishaw LJ, Ashmeade TL, Miller E, Gilbert JA . (2014). Development of the preterm infant gut microbiome: a research priority. *Microbiome* 2: 38.

Guttman D, Dykhuizen D . (1994). Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science* 266: 1380-1383.

He M, Miyajima F, Roberts P, Ellison L, Pickard DJ, Martin MJ *et al.* (2013). Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. *Nat Genet* 45: 109-113.

Hendricks CW . (1972). Enteric bacterial growth rates in river water. *Appl Microbiol* 24: 168-174.

Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ . (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11: 119.

Joshi N, Fass J . (2011), Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.29) [Software].

Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M . (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* 42: D199-D205.

Köser CU, Holden MTG, Ellington MJ, Cartwright EJP, Brown NM, Ogilvy-Stuart AL *et al.* (2012). Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N Engl J Med* 366: 2267-2275.

La Rosa PS, Warner BB, Zhou Y, Weinstock GM, Sodergren E, Hall-Moore CM *et al.* (2014). Patterned progression of bacterial populations in the premature infant gut. *Proc Natl Acad Sci* 111: 12522-12527.

Langmead B, Salzberg SL . (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9: 357-359.

Lee H, Popodi E, Tang H, Foster PL . (2012). Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. *Proc Natl Acad Sci USA* 109: E2774–E2783.

Leinonen R, Diez FG, Binns D, Fleischmann W, Lopez R, Apweiler R . (2004). UniProt archive. *Bioinformatics* 20: 3236–3237.

Loman N, Constantinidou C, Christner M, Rohde H, Chan JZ, Quick J *et al.* (2013). A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of shiga-toxigenic *Escherichia coli* O104:H4. *JAMA* 309: 1502–1510.

Lowe TM, Eddy SR . (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25: 0955–0964.

Macdorman MF, Declercq E, Mathews TJ . (2013). Recent trends in out-of-hospital births in the United States. *J Midwifery Women's Heal* 58: 494–501.

Magne F, Abély M, Boyer F, Morville P, Pochart P, Suau A . (2006). Low species diversity and high interindividual variability in faeces of preterm infants as revealed by sequences of 16S rRNA genes and PCR-temporal temperature gradient gel electrophoresis profiles. *FEMS Microbiol Ecol* 57: 128–138.

Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Pillay M *et al.* (2014). IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res* 42: D560–D567.

McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ *et al.* (2013). The comprehensive antibiotic resistance database. *Antimicrob Agents Chemother* 57: 3348–3357.

Moles L, Gómez M, Heilig H, Bustos G, Fuentes S, de Vos W *et al.* (2013). Bacterial diversity in meconium of preterm neonates and evolution of their fecal microbiota during the first month of life. *PLoS One* 8: e66986.

Moran NA, McLaughlin HJ, Sorek R . (2009). The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science* 323: 379–382.

Morowitz MJ, Deneff VJ, Costello EK, Thomas BC, Poroyko V, Relman DA *et al.* (2011). Strain-resolved community genomic analysis of gut microbial colonization in a premature infant. *Proc Natl Acad Sci USA* 108: 1128–1133.

Morrow AL, Lagomarcino AJ, Schibler KR, Taft DH, Yu Z, Wang B *et al.* (2013). Early microbial and metabolomic signatures predict later onset of necrotizing enterocolitis in preterm infants. *Microbiome* 1: 13.

Nawrocki EP, Eddy SR . (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29: 2933–2935.

Neu J, Walker WA . (2011). Necrotizing enterocolitis. *N Engl J Med* 364: 255–264.

Peng Y, Leung HCM, Yiu SM, Chin FYL . (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28: 1420-1428.

Raveh-Sadka T, Thomas BC, Singh A, Firek B, Brooks B, Castelle CJ *et al.* (2015). Gut bacteria are rarely shared by co-hospitalized premature infants, regardless of necrotizing enterocolitis development. *Elife*; e-pub ahead of print 3 March 2015; doi:10.7554/eLife.05477.

Schnabl K-L, Van Aerde J-E, Thomson A-B, Clandinin M-T . (2008). Necrotizing enterocolitis: a multifactorial disease with no cure. *World J Gastroenterol* 14: q2142-2161.

ShaIron I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, Banfield JF . (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res* 23: 111-120.

Sim K, Shaw AG, Randell P, Cox MJ, McClure ZE, Li M-S *et al.* (2015). Dysbiosis anticipating necrotizing enterocolitis in very premature infants. *Clin Infect Dis* 60: 389-397.

Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Henderson DK, Palmore TN *et al.* (2012). Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. *Sci Transl Med* 4: 148ra116.

Sommer DD, Delcher AL, Salzberg SL, Pop M . (2007). Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics* 8: 64.

Ultsch A, Moerchen F . (2005). ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM. *Tech Rep Dep Math Comput Sci Univ Marburg, Ger* 46.

Wang Y, Hoenig JD, Malin KJ, Qamar S, Petrof EO, Sun J *et al.* (2009). 16S rRNA gene-based analysis of fecal microbiota from preterm infants with and without necrotizing enterocolitis. *ISME J* 3: 944-954.

Zerbino DR, Birney E . (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18: 821-829.

Zhou Y, Shan G, Sodergren E, Weinstock G, Walker WA, Gregory KE . (2015). Longitudinal analysis of the premature infant intestinal microbiome prior to necrotizing enterocolitis: a case-control study. *PLoS One* 10: e0118632.

Acknowledgements

This work was supported by NIH Grant 5R01AI092531, a Sloan foundation grant APSF-2012-10-05 and an EMBO long-term fellowship (to TRS and IS). ggKbase was supported by DOE grants DE-SC0004918 and ER65561. We thank Andrea Singh for assistance with ggKbase and Elizabeth Costello for insights from 16S rRNA-sequencing data.