Title
The Interaction of Neonatal Intensive Care Unit Microbes with the Microbiome of the Developing Preterm Infant Gut

Author
Brooks, Brandon

Publication Date
2016

Peer reviewed|Thesis/dissertation

The Interaction of Neonatal Intensive Care Unit Microbes with the Microbiome of the Developing Preterm Infant Gut

By

Brandon Brooks

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Microbiology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jillian F. Banfield, chair
Professor Steven E. Lindow
Professor William W Nazaroff

Fall 2016

Abstract

The Interaction of Neonatal Intensive Care Unit Microbes with the Microbiome of the Developing Preterm Infant Gut

by

Brandon Brooks

Doctor of Philosophy in Microbiology

University of California, Berkeley

Professor Jillian F. Banfield, Chair

Humans spend approximately 90% of their time indoors, yet we know very little about the microbial ecosystem of the built environment and how it impacts occupants. Here we use infants hospitalized in a neonatal intensive care unit (NICU) as a model system to track the exchange of microbes between room and occupants. By leveraging high-throughput sequencing and other "omics" technologies, we conducted four major studies to broadly address the composition of microbes populating NICU surfaces, how these microbes migrate to the infant gut, and once in the gut, how these microbes compete for resources. Over the course of these campaigns we collected and processed over 5,000 samples from hospital room surfaces and over 300 infant fecal samples creating the largest collection of hospital samples to be interrogated with next-generation sequencing techniques. Using an approach that reassembles the entire 16S rRNA gene from room amplicons and gut metagenomics data, we discovered several organisms on room surfaces before their detection in the infant gut. Once in the gut, we used a metaproteomics technique to investigate the metabolisms of early infant gut colonizers. Unlike the anaerobic gut environment of older children and adults, we discovered a relatively high utilization of aerobic pathways in many of the facultative anaerobes colonizing the infant gut. We also observed niche partitioning amongst closely related *Citrobacter* strains in our strain-resolved proteomics data, providing insight into how early colonizers compete in the nascent infant gut. To better understand biomass trends inside and outside the gut, we developed an assay to quantify 16S rRNA gene copies using droplet digital PCR (ddPCR). We discovered a surprising amount of variation in bacterial densities across different NICU environments. These data also allowed us to adapt a novel *in silico* data cleaning method that leverages the quantification of negative controls to provide data less impacted by the inherent noise of low-biomass amplicon workflows. Cleaner data allowed us to apply a machine learning classifier that showed each infant's room had a distinct microbial fingerprint. To validate this result, we conducted a metagenomics campaign on pooled room samples from six different infants. After assembly and binning, we were able to recover hundreds of high quality genomes. Utilizing genomes from this dataset and previously isolated genomes from our lab, we discovered the same strains in the room as in infants. Further, we found several taxa frequently isolated from infant gut samples in this NICU are the same strains in the NICU room metagenomes. Overall,

the analysis from this work suggests that where a premature infant is born and the history of room occupancy can impact its gut microbiome development.

TABLE OF CONTENTS

FIGURES INDEX

TABLES INDEX

Dedication

To the hardest working man I know, my father, Randy Brooks. Inheriting your work ethic is the only way a scientist named “Bubba” could survive grad school.

To my loving mother, Sue Brooks. You tolerated my curiosities for years. It seems to have paid off.

To my creative inspiration, my wife, Emma Brooks. You challenge me to do better. Thank you.

Introduction

Humans spend approximately 90% of their time indoors (Klepeis *et al.*, 2001), yet we know very little about the microbial ecosystem of the built environment and how it impacts occupants. Understanding room and occupant interaction has minor human health implications in quotidian settings, like discomfort from allergens in an office or classroom, to more life threatening outcomes in hospital intensive care wards. While microbiome studies from a variety of building types have been conducted (e.g., in classrooms (Qian *et al.*, 2012; Meadow *et al.*, 2014), offices (Chase *et al.*, 2016), clean rooms (Mahnert *et al.*, 2015; Weinmaier *et al.*, 2015), homes (Ruiz-Calderon *et al.*, 2016; Barberan *et al.*, 2015; Lax *et al.*, 2014; Adams *et al.*, 2013), and the international space station (Checinska *et al.*, 2015)), no studies have implemented next-generation sequencing technologies to directly link colonization of occupants by strains sourced from the built environment. Two main reasons drive this knowledge gap. One, humans older than 2.5 years old have a fully developed gut microbiome (Koenig *et al.*, 2011). Any innocuous strain from the room is not likely to perturb a mature microbiome, since the adult microbiome is relatively stable over time and resistant to perturbation (David *et al.*, 2014). This feature makes detection of a room strain difficult due to its low abundance relative to the mature, established microbial community. Two, recovering genomes from complex adult microbiome samples has only recently been achieved (Di Rienzi *et al.*, 2013). This limitation in earlier research is notable as strain-level resolution is essential in determining if the same room microbe is colonizing occupants.

From a microbial standpoint, newborn infants represent a "blank canvas" that is highly susceptible to environmental influences and perturbations of the colonization process. The colonization process is critical (Cahenzli *et al.*, 2013; Costello *et al.*, 2012; Arrieta *et al.*, 2015; Sim *et al.*, 2013). Recent evidence suggests that aberrant patterns of microbiome development in newborns are linked to adverse short-term complications such as sepsis (Madan *et al.*, 2012) and necrotizing enterocolitis (Morowitz *et al.*, 2010b) and long-term complications such as asthma (Huffnagle, 2010) and atopic skin disease (Kong *et al.*, 2012). Here, we implement a series of studies that leverage recently developed 'omics' techniques to characterize the microbes of the neonatal intensive care unit (NICU) built environment and link these microbes to the colonization of preterm infant occupants.

At the time our first study was designed (Brooks *et al.*, 2014), only four published studies using molecular techniques to characterize microbes in intensive care unit wards had been conducted (Bokulich *et al.*, 2013; Hewitt *et al.*, 2013; Poza *et al.*, 2012; Oberauner *et al.*, 2013). Briefly, these prior studies found major differences between different regions of the hospital. The general access hallways have much higher microbial diversity relative to restricted access intensive care units. Microbes in a NICU largely resemble human-associated taxa and cleaning practices maintained low levels of biomass in the NICU. We designed an experiment to test if we could find 16S rRNA genes on NICU room surfaces before they were detected in the gut of infants. We found several instances where this was true (Brooks *et al.*, 2014), indicating NICU surfaces could be a reservoir that continually seeds the nascent infant gut. We then scaled up a second experiment to include more infants, more time points, and deeper sequencing, ultimately with the goal of linking the same room strain with strains recovered in metagenomic data from infant gut samples.

We also investigated the metabolisms of early infant gut colonizers. Early in infant gut colonization there is often a shift from an aerobic to anaerobic state (Penders *et al.*, 2006). Since many organisms commonly detected in the early preterm infant gut are facultative anaerobes, it was unclear if the infants in our cohorts were utilizing the aerobic or anaerobic repertoire of their metabolic potential, or both. To address this matter, we coupled strain-resolved community metagenomics data with mass spectrometry-based proteomics to resolve growth mode and to compare activity levels during colonization of a preterm infant (Brooks *et al.*, 2015). The results showed utilization of both lifestyles, highlighting niche partitioning in the infant gut. This study was the first to differentiate expression profiles between two closely related strains in the gut, two *Citrobacter spp*, using an untargeted metaproteomics technique.

We developed a wet lab workflow to circumvent major problems with processing of the vast number of low biomass samples collected during the NICU room-occupant campaign. Popular iTag sequencing workflows are notorious for generating spurious OTUs and are highly influenced by contamination (Lazarevic *et al.*, 2016). We leveraged template counting via droplet digital PCR and developed an *in silico* method to generate cleaner data and mediate common pitfalls in cleanroom microbiology. The results from this effort were compiled in a study that quantified biomass in the NICU over time. Surprisingly there is a wide range of bacterial density in the NICU, with occupancy driving much of this signal. We also implemented a machine learning algorithm which highlighted how each infant's NICU room contains a unique microbial fingerprint. This fingerprint echoes much of the successional patterns in the infant's gut. This emerging information led us to propose a model in which microbes bloom in the infant gut, are shuttled to the room via healthcare providers, they then tolerate the room environment and colonize downstream infants, resetting the cycle. This model was inferred from 16S rRNA gene amplicon data obtained from room samples and metagenomics data obtained from infant fecal samples. To validate this model, room metagenomics was needed.

Since our main NICU study collected over 3,700 samples, it offered the opportunity to conduct metagenomics on room samples. NICU room sample are extremely low biomass relative to most studies that attempt to recover genomes from metagenomic data. Fortunately, we had quantified biomass in these samples via ddPCR and were able to extrapolate the amount of genomic DNA (gDNA) in these samples in order to reach the minimum recommend amount for Illumina library construction. After a massive sample pooling effort and deep sequencing on an Illumina 4000 sequencing instrument, we successfully recovered hundreds of genomes from NICU surfaces. This outcome represents the first time that researchers have recovered genomes of this quality from indoor surfaces. The success of our effort was attributable in part to *a priori* knowledge of biomass per sample, the amount of sequencing allocated, and our lab's experience in genome recovery from complex microbial samples. Perhaps the most exciting finding from this effort, and possibly the most exciting finding from this dissertation, was confirmation of our hypothesis. Specifically, we found several genomes in room samples, resolved to the strain level, before they were detected in infant gut samples. Many of these strains have been consistently recovered from infant gut samples years apart in this NICU (Raveh-Sadka *et al.*, 2016).

Overall results from this work suggest a cycle of room to occupant interaction. Preterm infants enter the NICU with a relatively sterile gut microbiome. They then source a subset of microbes from the room. Once in the gut, microbes bloom and are dispersed to the immediate environment. The relative abundance of these dispersed microbes varies with cleaning, but despite efforts to maintain rooms in a clean and hygienic state, it is never possible to sterilize rooms (Hu *et al.*, 2015). We now have evidence this cycle continues for years, maintaining a

subset of taxa that appear to specialize at preterm infant gut colonization and also possess a tolerance for hospital surfaces. Knowing that one cannot remove or kill these persistent organisms despite careful efforts to maintain room hygiene and cleanliness at a high state, perhaps a more holistic approach to hospital hygiene should be explored.

Chapter 1:

1 Microbes in the neonatal intensive care unit resemble those found in the gut of premature infants

Brandon Brooks[1], Brian A. Firek[3], Christopher S. Miller[2,4], Itai Sharon[2,6], Brian C. Thomas[2], Robyn Baker[5], Michael J. Morowitz[3], Jillian F. Banfield[2]

1 – Department of Plant and Microbial Biology, University of California, Berkeley, CA
2 – Department of Earth and Planetary Sciences, University of California, Berkeley, CA
3 – University of Pittsburgh School of Medicine, Pittsburgh, PA
4 – Department of Integrative Biology, University of Colorado Denver, Denver, CO (current address)
5 – Division of Newborn Medicine, Children's Hospital of Pittsburgh of UPMC, Pittsburgh, PA
6 – Department of Computer Science, Tel-Hai College, Safed, Israel (current address)

This material was published in an open access journal and is freely available here (Brooks *et al.*, 2014): http://www.microbiomejournal.com/content/2/1/1

1.1 Abstract

Background

The source inoculum of gastrointestinal tract (GIT) microbes is largely influenced by delivery mode in full-term infants, but these influences may be decoupled in very low birth weight (VLBW, <1,500 g) neonates via conventional broad-spectrum antibiotic treatment. We hypothesize the built environment (BE), specifically room surfaces frequently touched by humans, is a predominant source of colonizing microbes in the gut of premature VLBW infants. Here, we present the first matched fecal-BE time series analysis of two pre-term VLBW neonates housed in a neonatal intensive care unit (NICU) over the first month of life.

Results

Fresh fecal samples were collected every three days and metagenomes sequenced on an Illumina HiSeq2000. For each fecal sample, approximately 33 swabs were collected from each NICU room from six specified areas: sink, feeding and intubation tubing, hands of healthcare providers and parents, general surfaces, and nurse station electronics (keyboard, mouse, and cell phone). Swabs were processed using a recently developed EMIRGE amplicon pipeline in which full-length 16S rRNA amplicons were sheared and sequenced using an Illumina platform, and short reads reassembled into full-length genes. Over 24,000 full-length 16S rRNA sequences were produced, generating an average of approximately 12,000 OTUs (clustered at 97% nucleotide identity) per room-infant pair. Dominant gut taxa, including *Staphylococcus epidermidis*, *Klebsiella pneumoniae*, *Bacteroides fragilis,* and *Escherichia coli*, were widely distributed throughout the room environment with many gut colonizers detected in more than half of samples. Reconstructed genomes from infant gut colonizers revealed a suite of genes that confer resistance to antibiotics (e.g., tetracycline, fluoroquinolone, and aminoglycoside) and sterilizing agents, which likely offer a competitive advantage in the NICU environment.

Conclusion

We have developed a high-throughput culture-independent approach that integrates room surveys based on full-length 16S rRNA gene sequences with metagenomic analysis of fecal samples collected from infants in the room. The approach enabled identification of discrete ICU reservoirs of microbes that also colonized the infant gut and provided evidence for the presence of certain organisms in the room prior to their detection in the gut.

1.2 Introduction

From birth to death, humans spend approximately 90% of their time indoors (Klepeis *et al.*, 2001). This realization, coupled with advancements in DNA sequencing technologies, has spawned a new interest in studying buildings as ecosystems. Pioneering efforts have revealed a built environment (BE), a term used here to collectively describe both the biotic and abiotic features of a building structure, that is far more complex than originally imagined (Tringe *et al.*, 2008; Rintala *et al.*, 2008). Diverse microbial communities have been uncovered in a variety of BEs (Kelley and Gilbert, 2013) and, surprisingly, from sites engineered to be sterile or near-sterile, such as NASA clean rooms (La Duc *et al.*, 2007, 2012) and high-risk hospital wards (Perkins *et al.*, 2009; Poza *et al.*, 2012; Oberauner *et al.*, 2013; Hewitt *et al.*, 2013). Additionally,

recent studies characterizing different building types have revealed general trends suggesting a room's function or architecture influences the BE's microbiome (Poza *et al.*, 2012; Kembel *et al.*, 2012). Intra-building experiments in hospitals have corroborated this concept, showing that general use areas, like waiting rooms and lobbies, have a markedly different microbial community compared to more restrictive hospital zones such as intensive care units (Poza *et al.*, 2012). The exchange between the BE microbiome and the human microbiome communities remains unclear; however, the observation that human pathogens are enriched in hospital settings is of obvious concern (Kembel *et al.*, 2012). Here we aimed to characterize the interaction between the BE's microbiome and the human microbiome through study of very low birth weight (VLBW, < 1500 g) infants housed in a NICU as our model system.

Infants housed in a NICU are well suited to studies that aim to characterize interactions between the BE and occupants. *In utero*, infants are canonically thought to exist in a sterile or near-sterile environment (Penders *et al.*, 2006). Acquisition of the microbiome starts at birth and is strongly influenced by mode of delivery (Dominguez-Bello *et al.*, 2010). Patterns of colonization in full-term infants tend to follow a well documented trajectory affected by diet, host genotype, and a limited set of other variables, with the infant gut converging on an adult-like state around 2.5 years of life (Palmer *et al.*, 2007; Trosvik *et al.*, 2010). In VLBW infants, early gut succession is characterized by extremely limited diversity, chaotic flux in community composition, and an abundance of opportunistic pathogens (Morowitz *et al.*, 2010a; Wang *et al.*, 2009; Morowitz *et al.*, 2010b; Mshvildadze *et al.*, 2010). It is possible that a high rate of caesarean deliveries and the routine use of broad-spectrum antibiotics during the first week of life serve to decouple VLBW infants from source inoculum introduced during the birthing process. These influences likely render premature infant microbiomes especially susceptible to environmental influences.

There is strong evidence suggesting that the ICU serves as a reservoir of clinically relevant pathogens. "Outbreaks" of disease in ICUs are relatively common, and a recent study estimated at least 38% of all ICU outbreaks could be attributed to microbial sources within the ICU environment, such as equipment, or personnel (Gastmeier *et al.*, 2007). In addition, upward of 63% of extremely preterm infants develop life-threatening infections (Stoll *et al.*, 2010). Epidemiologic investigations indicate environmental sources of infective agents in air (Adler *et al.*, 2005), infant incubators (Singh *et al.*, 2005; Touati *et al.*, 2009), sink drains (Bonora *et al.*, 2004), soap dispensers (Buffet-Bataillon *et al.*, 2009), thermometers (Van Den Berg *et al.*, 2000), and baby toys (Naesens *et al.*, 2009). Clearly there is a growing need for comprehensive ecological surveys of the hospital BE to better understand the overall process of microbe migration and establishment on and in the bodies of occupants. Here, we performed the first matched time series characterization of the NICU and infant gut. Our analysis used two main approaches: (a) metagenomic sequencing of microbial community DNA extracted from fecal samples to evaluate the metabolic potential of gut colonizing microorganisms and (b) a recently developed EMIRGE amplicon protocol to profile the microbial community composition of BE samples collected from six environment types (Miller *et al.*, 2013). Our protocol was aimed at addressing the hypothesis that the BE, specifically room surfaces frequently touched by humans, is a predominant source of colonizing microbes in the GI tract of premature infants.

1.3 Methods

1.3.1 Sample collection

Fecal samples were collected every third day, starting on the third day of life, for one month from two infants. Infants were enrolled in the study based on the criteria that they were <31 weeks gestation, <1250 g at birth, and were housed in the same physical location within the NICU during the first month of life. A summary of health-related metadata including antibiotics exposure is provided in Table 1-1. Fecal samples were collected using a previously established perineal stimulation procedure and were stored at -80 °C within 10 minutes (Morowitz *et al.*, 2010a). All samples were collected after signed guardian consent was obtained, as outlined in our protocol to the ethical research board of the University of Pittsburgh (IRB PRO11060238). This consent included sample collection permissions and consent to publish study findings.

All samples were obtained from a private-style NICU at Magee-Womens Hospital of the University of Pittsburgh Medical Center. Room samples were collected concurrently with fecal samples and spanned four time points on days of collection (9:00, 12:00, 13:00, and 16:00). Most frequently touched surfaces were determined by visual observation and health care provider interviews in the weeks leading up to sample collection. Microbial cells were removed from surfaces using foam tipped swabs (BBL CultureSwab EZ Collection and Transport System, Franklin Lakes, NJ, USA) and a sampling buffer of 0.15 M NaCl and 0.1% Tween20. Six frequently touched areas were processed per infant room: sink, feeding and intubation tubing, hands of healthcare providers and parents, general surfaces, access knobs on the incubator, and nurse station electronics (keyboard, mouse, and cell phone). All samples were placed in a sterile transport tube and stored within 10 minutes at -80 °C until further processing.

1.3.2 DNA extraction and PCR amplification

Frozen fecal samples were thawed on ice and 0.25 g of thawed sample added to tubes with pre-warmed (65 °C) lysis solution from the PowerSoil DNA Isolation Kit (MoBio Laboratories, Carlsbad, CA, USA). The incubation was conducted for five minutes and the manufacturer's protocol followed thereafter. Swab heads followed the same procedure, except heads were cut with sterilized scissors into the extraction tube before starting the protocol.

DNA extracted from swabs was pooled such that the four time points sampled in one day per environment were consolidated to one. Pooled DNA was used as template for amplification of the full-length 16S rRNA gene with 27F (5'-AGAGTTTGATCCTGGCTCAG-3') and 1492R (5'-GGTTACCTTGTTACGACTT-3') primers (Stackebrandt and Goodfellow, 1991). To limit PCR bias, gradient PCR was performed with 5 units μl^{-1} of *TaKaRa Ex Taq*™ (Takara Bio Inc., Otsu, Japan) across 7 different annealing temperatures with the following reaction: 1 min at 94 °C; 35 cycles of 1 min at 94 °C, 30 s at 48–58 °C (7 temperature gradient) and 1 min at 72 °C; and a final extension for 7 min at 72 °C. Amplicons were combined across gradients and cleaned with the QIAquick PCR Purification Kit (Qiagen, Hilden, Germany) as directed by the manufacturer. Cleaned amplicons were quantified via Qubit and input into an Illumina library preparation pipeline.

1.3.3 Sequencing preparation and sequencing

Illumina library construction followed standard protocols at the University of California Davis DNA Technologies Core Facility (http://dnatech.genomecenter.ucdavis.edu) as previously described (Miller *et al.*, 2013). Briefly, amplicons were fragmented to an average size of 225 bp using the Bioruptor NGS (Diagenode), and sheared fragments were used in a robotic library preparation protocol using the Appollo 324 robot (Integenx) following the manufacturer's instructions. Each sample was tagged with unique barcodes consisting of 6 nucleotides internal to the adapter read as a separate indexing read, and ligated to each fragment. Twelve cycles of PCR were enriched for adapter-ligated fragments before library quantification and validation. Fecal samples underwent the same preparation with two exceptions: (1) Genomic DNA was used and (2) DNA was fragmented to 550 bp. Libraries were added, in equimolar amounts, to the Illumina HiSeq 2000 platform. Paired-end sequences were obtained with 100 cycles and the data processed with Casava version 1.8.2. Raw read data have been deposited in the NCBI Short Read Archive (accession numbers SRP033353).

1.3.4 EMIRGE assembly of full-length 16S rRNA gene amplicons

EMIRGE is an iterative template-guided assembler that relies on a database of 16S rRNA gene sequences to probabilistically generate full-length 16S rRNA gene sequences and provide the relative abundance of these sequences in the assayed consortia (Miller *et al.*, 2011). For the reference database, we used version 108 of the SILVA SSU database, filtered to exclude sequences < 1200 bp and > 1900 bp (Pruesse *et al.*, 2007). To remove closely related sequences, we clustered the database at 97 % identity with USEARCH (Edgar, 2010). One million paired-end reads from each barcoded library were sampled randomly without replacement to accommodate computational restrictions associated with use of the full dataset. Reads from the subsample from each library were stringently trimmed using Sickle (Joshi, 2011) for quality scores > 30 and length > 60 bp. Trimmed reads were input into an amplicon-optimized version of EMIRGE (Miller *et al.*, 2013) for assembly using default parameters. Eighty iterations were performed for each subsample. EMIRGE-reconstructed sequences without Ns (ambiguous bases) and with an estimated abundance of 0.01 % or greater were kept for analysis. Putative chimeras were removed by using the intersection between two chimera detection programs, DECIPHER (Wright *et al.*, 2012) and UCHIME v6.0 (Edgar *et al.*, 2011) searched against the 2011 Greengenes database (McDonald *et al.*, 2012). Finally, reconstructed sequences from a spike-in control experiment (data not shown) were removed for downstream analysis. Sequences used in the analysis are publicly available as a project attachment at http://ggkbase.berkeley.edu/NICU-Micro/organisms.

1.3.5 Metagenomic EMIRGE assembly of 16S rRNA gene

Metagenomic sequencing of 16 fecal samples on one lane of an Illumina HiSeq 2000 produced ~350 Mbp of 101 bp paired-end reads. Trimmed reads were input into EMIRGE and default parameters run for 80 iterations using the aforementioned database. After the final iteration, 153,980 reads, spanning all samples, were used in reconstructing fecal 16S rRNA

sequences. Downstream filtering and analysis of reconstructed 16S rRNA gene sequences from fecal samples followed that of the room samples.

1.3.6 Community analysis of room and fecal samples

For community analysis, EMIRGE-reconstructed sequences were input into the standard QIIME 1.5.0 workflow (Caporaso *et al.*, 2010b). For presence/absence analyses, representative OTUs were clustered at the > 97% identity level using USEARCH (Edgar, 2010) and an OTU table was constructed using QIIME's pick_otus_through_otu_table.py script. An adjusted OTU table that incorporated EMIRGE generated abundances was constructed using an in-house script (Miller *et al.*, 2013) and is publicly available as a project attachment at http://ggkbase.berkeley.edu/NICU-Micro/ . OTUs were aligned to the Greengenes (DeSantis *et al.*, 2006) reference alignment (gg_97_otus_4feb2011.fasta) using the PyNAST aligner (Caporaso *et al.*, 2010a) and a phylogenetic tree built using FastTree v.2.1.3 (Price *et al.*, 2010) with default parameters. Beta diversity was calculated from similar trees using Fast UniFrac scores and visualized with principal coordinates analysis (PCoA) (Hamady *et al.*, 2010). Taxonomy was assigned to each OTU at the genera and/or species level using the RDP classifier (Wang *et al.*, 2007) at a confidence interval of 0.8 and trained with the same Greengenes database. OTUs were visualized across room-infant pairs in a spring weighted, edge-embedded network plot by using QIIME's make_otu_network.py script (Caporaso *et al.*, 2010b) with the modified OTU table as input.

1.3.7 Metagenomic assembly and gene prediction

Assemblies were constructed using idba_ud (Peng *et al.*, 2012) and an iterative implementation of Velvet (Zerbino and Birney, 2008; Sharon *et al.*, 2013). For idba_ud assemblies, trimmed reads were assembled using default parameters. For the Velvet assemblies, sequence coverage bins representing major genomes in the dataset were identified by first running the program with permissive parameters in which the k-mer size covered the whole range of observed coverages. We summed the k-mer coverages for all contigs generated by this assembly to define the coverage bins (each of which contains one or more genomes). This provided bin-specific expected coverage, k-mer size, coverage cutoff, and coverage collection threshold parameters for the iterative assembly. After each iteration targeting a specific bin, the bin-specific reads were removed from the dataset.

Time-series-coverage-based emergent self-organizing maps (ESOMs) were used to bin scaffolds generated by metagenomic assembly (Ultsch and Mörchen, 2005). Genes were predicted and translated into protein sequences using Prodigal (Hyatt *et al.*, 2010). Functional annotation was added with an in-house pipeline (Sharon *et al.*, 2013). Genome completeness was determined based on the number of single-copy genes and other conserved genes (Sorek *et al.*, 2007; Wu and Eisen, 2008) identified in each bin. The relative abundance of each organism in each sample was calculated by mapping reads to unique regions on the assembled genomes. Metagenomic assemblies along with their annotations are publicly available at http://ggkbase.berkeley.edu/NICU-Micro/.

1.3.8 Enterococcus faecalis concatenated ribosomal protein phylogeny

For phylogenetic resolution beyond the 16S rRNA gene, 32 highly conserved, single copy ribosomal proteins were used from infant 1 and 2's assemblies (RpL10, 13, 14, 16, 17, 18, 19, 2, 20, 21, 22, 24, 27, 29, 3, 30, 4, 5, and RpS10, 11, 12, 13, 15, 16, 17, 18, 19, 20, 5, 6, 7, 8). The same genes from recently sequenced *Enterococcus faecalis* genomes, in addition to genes from more distantly related taxa, were obtained from the JGI IMG database. Together, each gene set was aligned using MUSCLE 3.8.31 (Edgar, 2004a, 2004b) and manually curated to remove ambiguously aligned regions and end gaps (Hug *et al.*, 2013). The curated alignments were concatenated to form a 32 gene, 39 taxa, 4,101-position alignment. A maximum likelihood phylogeny for the concatenated alignment was conducted using PhyML under the LG + α + γ model of evolution with 100 bootstrap replicates.

1.4 Results

1.4.1 Stability of NICU room samples over time and space

After sample preparation, 57 and 36 room samples amplified successfully and were subsequently analyzed for Infant 1 and Infant 2, respectively (Table 1-2). EMIRGE generated approximately 12,000 full-length 16S rRNA sequences and OTUs for each room-infant pair (clustered at the 97% nucleotide identity level). Broadly speaking, species richness decreased from electronics > sinks > surfaces > incubators > hands > tubes, a finding that was corroborated with several alpha diversity indexes (Table 1-3). Nearly 300 genera were detected in the NICU. To broadly visualize temporal stability of environments across time and space, the phylum level classifications are plotted in Figure 1-1. Actinobacteria, Firmicutes, and Proteobacteria dominate the sampled environments, with areas most exposed to human skin deposition having the most variation over time. At lower taxonomic levels, similar trends are observed. Based on the twenty most abundant families, frequently touched surfaces are distinct from infrequently touched surfaces (Figure 1-1). UniFrac distance-based community composition PCoA reveals four discernible ecosystem types (skin associated communities, sinks, tubes, and feces) and confirms clustering of samples prone to skin deposition via touching (Figure 1-2).

1.4.2 Time-series characterization of fecal samples

More than 94% of the reads from Infant 1's samples mapped to scaffolds generated by the idba_ud assembly. Consequently, this assembly was accepted for further analysis. In comparison, the initial idba_ud assembly of metagenomic data from Infant 2 was highly fragmented, and less than 40% of reads could be mapped to the assembled scaffolds. Subsequent reassembly of metagenomic data from Infant 2's samples using the iterative Velvet-based assembly approach (Sharon *et al.*, 2013) generated a significantly better result. As > 90% of reads could be mapped to the scaffolds generated by the Velvet assembly, this assembly was chosen for further analysis.

The *de novo* assemblies reconstructed a majority of the genomes for four of the five and eight of the eleven most abundant bacterial colonists from Infant 1 and Infant 2's metagenomes, respectively. For Infant 1, time-series organism abundance patterns in the sample sets analyzed via ESOM (Figure 1-3) defined 5 major genome bins for which between 37 and 99% of the single copy genes were identified, based on standard analyses of the single copy gene inventory (Table 1-4). For Infant 2, time-series organism abundance patterns in the sample sets analyzed

via ESOM (Figure 1-3) defined 11 major genome bins for which between 27 and 99% of the single copy genes were identified (Table 1-4).

Infant 1 and Infant 2's GIT microbial communities are distinctly different. Infant 1's colonization pattern echoes the canonical observation in infant GIT succession that facultative anaerobes dominate early phase colonization whereas late stage colonizers are primarily obligate anaerobes (Penders *et al.*, 2006). This shift is observed on day of life 12 in Infant 1, but is not observed in Infant 2, in whom facultative anaerobes were observed throughout the study period. The metagenomic EMIRGE analyses corroborated the binning-based compositional analyses in that no sequences for new taxa were assembled for scaffolds included in the ESOM. Some 16S rRNA genes were identified in the metagenomic assemblies and match EMIRGE generated sequences with ~100% identity. The *Enterococcus faecalis* sequence from Infant 1 was not identified by EMIRGE due to low abundance, but was extracted from the assembly using RNAmmer for the phylogenetic analysis (Lagesen *et al.*, 2007).

1.4.3 Highly connected BE microbes

The distribution of shared OTUs across sampled sites was visualized through a spring-weighted edge-embedded network plot. To limit the noise from infrequently detected microorganism types, we restricted the plot to OTUs occurring in two or more samples from each infant (Figure 1-4). The spring weight is derived from EMIRGE generated abundances, and the distribution of OTUs in the plot is governed both by frequency of occurrence and abundance. In Figure 1-4, the circular white nodes (representing OTUs) found in many environment types (more edges) are pulled closer to the middle of the network whereas OTUs shared by only two samples (fewer edges) are positioned closer to the periphery of the network. The top 5% of most frequently occurring OTUs aggregate in a central cluster in the middle of the network. Similar to the PCoA plot, general clustering is observed based on environment type (i.e. skin-associated sites cluster together, as do sink samples). When restricting the network for OTUs only found in fecal samples (Figure 1-4 zooms), one can visualize the OTU distribution across the sampled NICU environments. Three highly connected OTUs are present in fecal samples, two of which are in the top 5% most frequently occurring OTUs in Infant 1's room samples. Several of the OTUs in Infant 2's fecal samples fall within the top ten most frequently occurring OTUs in the room environment. Interestingly, Infant 2's most abundant gut colonists, *Staphylococcus sp.* and *Enterococcus faecalis*, are the two most frequently occurring OTUs in the room environment.

1.4.4 The NICU as a reservoir for gut colonists

Figure 1-5 summarizes the gut colonizing organisms found in room samples at the genera level. Typically, for both infants, electronics had the lowest relative abundance of organisms detected in the gut whereas tubing had the highest. Temporal variation of gut genera was extreme in most environments.

The use of Bayesian microbial source tracking software (Knights *et al.*, 2011), with the perspective of room samples as the source and fecal samples as the sink, produced mixed results in terms of finding likely gut reservoirs (Figure 1-6). In Infant 1, tubing, surfaces, and electronics had the highest probabilities as sources, but the bloom of *B. fragilis*, from a source not detected by our sampling regime, lowered the probability of sampled source environments for the latter half of the sampling period. Infant 2's samples showed the opposite pattern in that early gut

colonists migrated from an unknown reservoir, whereas later in sampling, incubator, tubing, surfaces, and hands were the most probable reservoirs.

1.4.5 Shared gut colonizers

The infant cohort shared only one gut colonizer, *Enterococcus faecalis*, which contained 100% 16S rRNA gene level sequence identity. A higher resolution analysis using a concatenated alignment of 32 highly conserved, single-copy genes show the strains differ by only two amino acids across the 4,101 positions. These two *E. faecalis* strains phylogenetically cluster most closely to each other, but are very closely related to other *E. faecalis* strains (Figure 1-7).

To further explore similarity of shared strains, reads from Infant 1 were mapped to Infant 2's assembled contigs. Infant 1's reads covered 95% of the length of Infant 2's assembly at an average of 4.66X coverage. Read mapping revealed two distinct SNP profiles for Infant 1's reads, a major strain divergent from Infant 2's assembly and a minor strain identical to the strain in Infant 2. Seventy seven percent (77%) of the length of Infant 2's *E. faecalis* assembly is covered by Infant 1's reads mapped as mate pairs with no mismatches. This outcome suggests that Infant 1's *E. faecalis* minor strain is the same strain dominating Infant 2's gut. Pheromone-responsive plasmids were found in both infants. The plasmid from Infant 2 occurs in low abundance in Infant 1 (as expected based on the low representation of *E. faecalis* in Infant 1), but with high sequence identity.

1.4.6 Genes relevant to adaptation to the NICU environment

Analysis of reconstructed genomes for gut microorganisms can lend clues as to how organisms detected in the GIT and room environment are able to persist in the NICU, which is subjected to regular cleaning and sterilization. Numerous antibiotic resistance genes were found in genomes of microorganisms in fecal samples of both infants. A large portion of these were efflux pumps, with representatives from all four families of multidrug transporters: major facilitator superfamily (MFS), small multidrug resistant (SMR), resistance-nodulation-cell division (RND), and multidrug and toxic compound extrusion (MATE) proteins (van Veen, 2010). Particularly interesting are genes encoding the QacA/B MFS, SugE SMR, and MexA/B RND proteins, which are a growing concern in hospitals due to co-selection through the practice of combining two or more types of antibiotic treatments (Fernández and Hancock, 2012). Resistance to multiple types of antibiotics can arise from a single resistance mechanism such as efflux pumping (Buffet-Bataillon *et al.*, 2012). In addition to antibiotics, these pumps can expel quaternary ammonium compounds (QACs), the active biocide in the detergent used to clean hospital surfaces during the study. Other notable observations were the presence of biofilm forming genes in most colonizers, which can be induced by exposure to aminoglycosides (Hoffman *et al.*, 2005), a suite of genes that confer resistance to starvation, and the presence of antibiotic resistance genes encoded on several phage and plasmid genomes, as well as microbial genomes.

1.5 Discussion

Increasing throughput, decreasing cost, and rapid development of informatics and sequencing pipelines has reshaped the field of microbial ecology, allowing researchers to survey

a breadth of new environments (Mackelprang *et al.*, 2011; Dick *et al.*, 2013; Joshi, 2011; Fierer *et al.*, 2010). Recently, the first ICU survey to utilize next generation sequencing technology was published (Poza *et al.*, 2012) and showed a surprising amount of bacterial diversity for an environment under constant attack via aggressive sanitation and antibiotic treatment efforts. The consortia were generally diverse, but some consortia contained a high representation of members of the family Enterobacteriaceae, typically considered to be gut microbes. Subsequently, a study characterizing a snapshot of surfaces and sinks in two NICU rooms corroborated high proportions of fecal coliform bacteria on surface samples (Hewitt *et al.*, 2013). Certainly the NICU has the capacity to retain enteric microbes, but their propensity to migrate to the gut remains unclear.

Next-generation sequencing surveys in the ICU have reported high levels of community diversity. Poza et al found 1145 distinct OTUs in an ICU in Spain (Poza *et al.*, 2012) and subsequent studies reported 1621 and 3925 OTUs in a NICU in the US and in an Austrian ICU, respectively (Hewitt *et al.*, 2013; Oberauner *et al.*, 2013). While comparing these studies is difficult due to differences in sample size and protocols, we can begin to appreciate the need to better understand why so many types of bacteria can be found in a regularly cleaned environment. Our study, the first time series survey of an ICU using next-generation sequencing technologies, unveiled over 20,000 OTUs across two NICU rooms occupied by different infants with partial time overlap. Our study is distinct from prior NICU surveys in that it used amplicon-EMIRGE, a16S rRNA gene assembly software which can be more sensitive in OTU detection (Miller *et al.*, 2013) and can provide increased confidence when making lower taxonomic level classifications (Ong *et al.*, 2013). The increase in OTUs from study to study might be attributed to increases in sequencing read lengths and, in this study, increased information from reassembled, full-length genes, but the biological relevance of this increase is unclear. Notably, of the over 20,000 OTUs characterized here, only 984 were found in two or more samples. Further surveys are needed, integrating time-series sampling and samples from multiple surface types from different hospitals, to better characterize the expected number of OTUs in an ICU and the implications of this number for ICU occupants.

The increased sensitivity provided by EMIRGE was helpful when evaluating temporal patterns, especially pertaining to source-sink characterization. Similarly, our source-sink analyses benefited from the increased number of samples and time points relative to prior studies (Poza *et al.*, 2012; Hewitt *et al.*, 2013; Oberauner *et al.*, 2013), which did not attempt to identify source-sink relationships. The SourceTracker results suggest the most probable room reservoir for gut colonists is tubing followed by surfaces, incubators, and hands (Figure 1-6). The tubing area sampled, the hub of the silastic nasogastric feeding tube, is the closest in proximity to the infant and, since SourceTracker is not bidirectional, it is difficult to tease out the directionality in this exchange (Knights *et al.*, 2011). Incubators from both infants also appear to mirror successional patterns in the infant's GIT, but without finer scale temporal sampling it is difficult to differentiate between the source and the sink. The observation that hands tend to show a variable amount of potential fecal colonist is likely due to the variability in sampling and hand hygiene, as hand samples were taken both before and after infants received care from healthcare providers. A good example of this is Infant 1's DOL 27 hand sample in which the large spike in *Escherichia* likely came from a swab collected directly after contact with the infant (Figure 1-5).

Given the large inventory of sequences and the time-series dataset, it was possible to identify likely reservoirs of microorganisms in the room environment, prior to their appearance in the GIT (e.g., the asterisked OTUs in Figure 1-4). Many of these sequences had perfect or near

perfect identity between room and GIT 16S rRNA genes. Two notable examples include the *K. pneumoniae* in Infant 1 and *F. magna* in Infant 2, whose fecal to room sequence best hits averaged 99.4% and 99.6% identity respectively. Infant 1's *K. pneumoniae* is first detectable in the gut on DOL 9, but NICU samples first detect the organism on electronic and sink samples starting at DOL 3, our earliest sampling point. Interestingly, the *K. pneumoniae* is outcompeted in the gut, yet is reintroduced on two separate occasions. This observation could be a byproduct of our detection limits, but Klebsiella's relatively high abundance in many NICU samples and its availability at all time points, suggests the opportunity for reinoculation from multiple room reservoirs. The *F. magna* in Infant 2's samples exhibit similar patterns in that it is initially a high ranking taxa that is out competed by other Firmicutes, but is reintroduced later in the time series.

If the environment is a reservoir for gut colonizing microbes in our cohort, then it is likely infants housed in close proximity will share the same strain. The 16S rRNA gene survey shows the availability of reservoirs of colonizing populations (likely with multiple strain variants) in the infant's immediate environment. However, it cannot discriminate at the strain level, so the mere existence of a phylotype in the room prior to gut colonization is not a direct measure of BE to infant transfer. The current work resolves this, by using extensive genome sequence comparison of *E. faecalis* from the gut of two infants housed in the same ward to establish that environment to room occupant transfer occurs in the NICU. The mode of acquisition of Infants 2's abundant strain by Infant 1 is unclear, but nosocomial infection by enterococci is not uncommon.

Enterococci are particularly difficult to classify due the plasticity of their genomes. Upwards of 25% of *E. faecalis* genomes may be comprised of mobile or acquired elements (Arias and Murray, 2012). Recent experiments attribute this genome flexibility partially to the ability to produce transconjugant hybrid strains in which several 100 kb fragments can be transferred between donor and recipient strain (Manson *et al.*, 2010). Transfer of these genome fragments is dependent on pheromone-responsive plasmids, which were found in all strains studied here. The ability to form hybrids not only confounds the ability to confirm identical strains, unless the entire genome has been recovered, it also provides a competitive advantage in the hospital BE where enterococci have been problematic for decades (Murray, 1990; Arias and Murray, 2012). Enterococci are notoriously hardy and are able to persist on medical equipment and hospital surfaces for long periods (Bradley and Fraise, 1996; Arias and Murray, 2012). They are able to withstand chlorine, heat, some alcohol treatments, and possibly most concerning, several types of antibiotics (Arias and Murray, 2012). Their genome plasticity and ability to easily acquire new genes from other strains make them particularly well suited to thrive in the hospital environment.

Gut colonists must withstand selective pressures both inside and out of the gut. Two obvious forms of selection in the NICU come from hospital cleaning and the broad use of antibiotics. All rooms were cleaned daily using wet solutions containing QACs and all infants were administered multiple types of antibiotics. Incorrect administration of biocides, through misuse or unintended mixing with existing fluids (i.e. water from sink samples or removing sanitizing agents via water rinsing), could enrich for resistance genes (Condell *et al.*, 2012). Even if used to factory standards, if surface-dried cells or biofilms remain, biocide activity could be ineffective and contribute to cross-resistance to biocides and antibiotics (Weiss-Muszkat *et al.*, 2010). Biofilm forming communities can be upwards of 1000 times more resistant to QACs than their planktonic forms (Romanova *et al.*, 2007) and biofilm formation can be triggered by the types of antibiotics administered in this study (Hoffman *et al.*, 2005). These characteristics

may be a contributing factors as to why a recent study found enteric microbial communities to be relatively unaltered before and after routine NICU surface cleaning (Bokulich *et al.*, 2013). Certain types of biofilms in many *Enterobacteriaceae*, including those studied here, contain amyloid fibers, called curli. Curli have been implicated in adhesion to abiotic surfaces, like polystyrene, Teflon, and stainless steel, and contribute to adhesion to host epithelial cells and invasion by *E. coli* in the gut (Barnhart and Chapman, 2006). This type of dual-purpose adaptation may allow enteric organisms to persist on NICU surfaces until transmission to a more favorable environment like the gut. Efflux pumps are another multi-purpose adaptation conferring competitive advantages inside and out of the gut. Numerous pumps from every major class of efflux pump were identified here and, collectively, can function to pump out QACs and administered antibiotics. Previous studies have positively correlated high QAC MICs with increased antibiotic resistance markers in enteric microbes (Buffet-Bataillon *et al.*, 2011), indicating biocide efflux may be an important function for microbes in the ICU. Efflux and biofilm formation are two of many possible explanations as to how colonizers combat both biocides administered during NICU cleaning and host-administered antibiotics.

1.6 Conclusion

Through a time series analysis using full-length rRNA gene sequences, we have established that organisms that appear in the GI tract in the early phase of colonization have reservoirs in the room environment. The findings point to a scenario in which gut microbes are introduced from room sources, thrive in the gut, and are disseminated to the immediate environment, creating a cycle of room to infant colonization. The research also highlights the value of extensive genome comparisons to link colonists from different individuals, an approach that in the future may also target populations sampled directly from room reservoirs.

1.7 Abbreviations

BE built environment
DOL day of life
ESOM emergent self-organizing map
GIT gastrointestinal tract
ICU intensive care unit
MATE multidrug and toxic compound extrusion
MFS major facilitator superfamily
MIC minimum inhibitory concentration
NICU neonatal intensive care unit
OTU operational taxonomic unit
PCoA principal coordinates analysis
QAC quaternary ammonium compound
RND resistance-nodulation-cell division
SMR small multidrug resistant
VLBW very low birth weight

1.8 Competing Interests

The authors declare that they have no competing interests.

1.9 Authors' Contributions

JFB, MJM, and BB conceived of the project. RB organized cohort recruitment and sample collections. BAF conducted nucleic acid extractions and BB conducted the amplification reactions. IS and BB conducted the metagenomic assemblies and BCT provided annotations. CSM and BB conducted the 16S rRNA gene reconstructions. BB and JFB wrote the final manuscript. All authors have read and approve the manuscript.

1.10 Acknowledgements

Funding was provided through the Alfred P. Sloan Foundation and the National Science Foundation's Graduate Research Fellowship Program.

Figure 1-1: Taxonomic classification of Infant 1 and 2's NICU room microbes

Phylum- (top) and family- (bottom) level classifications were assigned using the RDP classifier on assembled full-length 16S rRNA genes. Day of life (DOL) is plotted on the *x*-axis and relative abundance, generated by EMIRGE, is plotted on the *y*-axis.

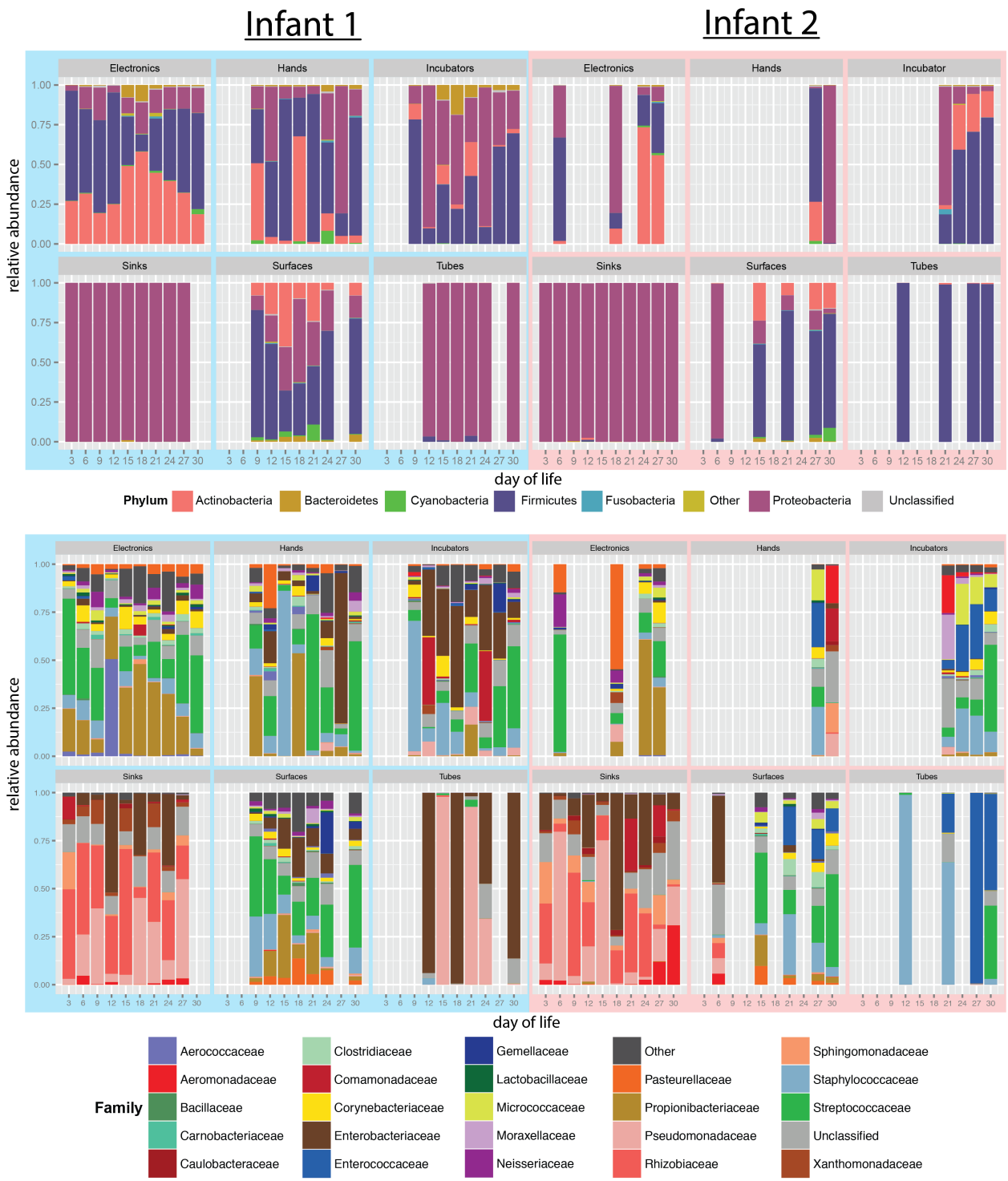


Figure 1-2: Principal coordinates analysis (PCoA) based on UniFrac scores of room and gut microbes

This figure reveals four discernible ecosystem clusters: skin associated communities, sinks, tubes, and feces.

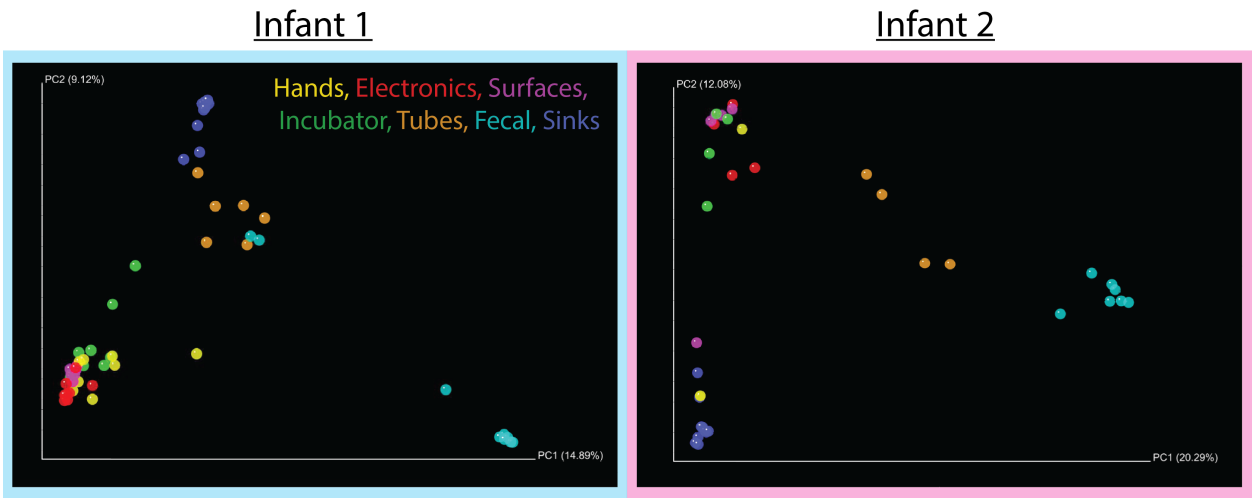
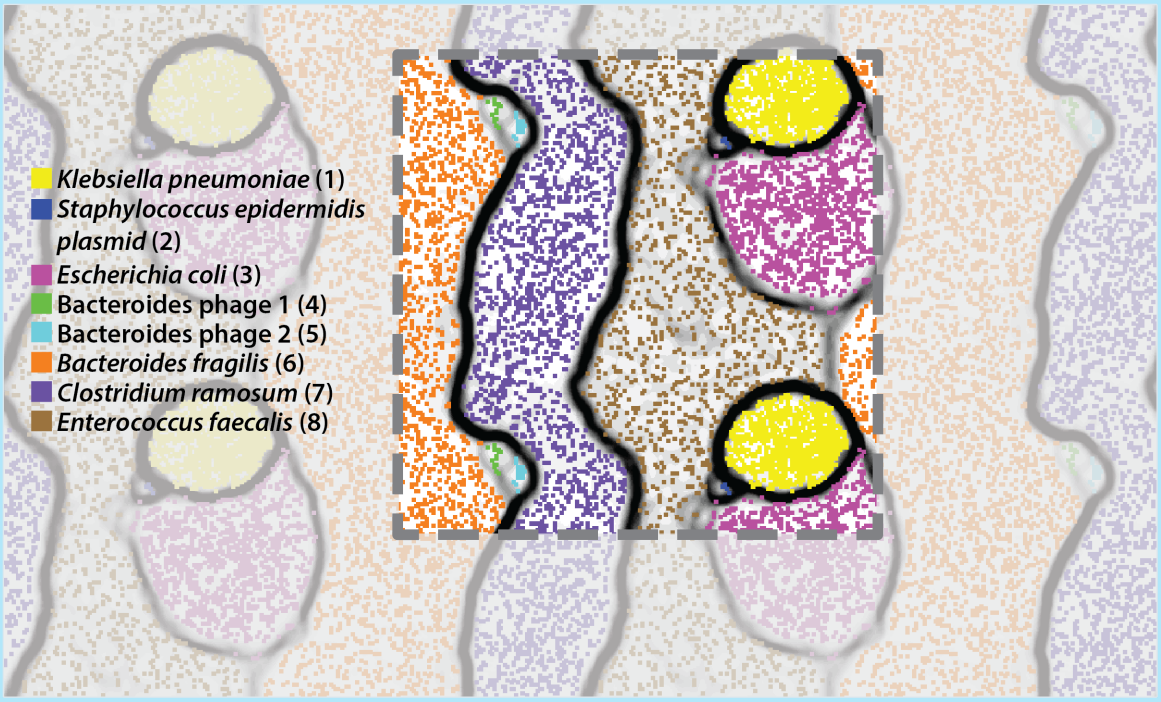


Figure 1-3: Time-series coverage emergent self-organizing maps (ESOMs) reveal discrete genome bins for each infant's dataset

The underlying ESOMs are shown in a tiled display with each data point colored by its taxonomic assignment. Labels to the left are colored to match their respective data points and numbers in parentheses correspond to the bin number in Table 4.

Infant 1



Infant 2

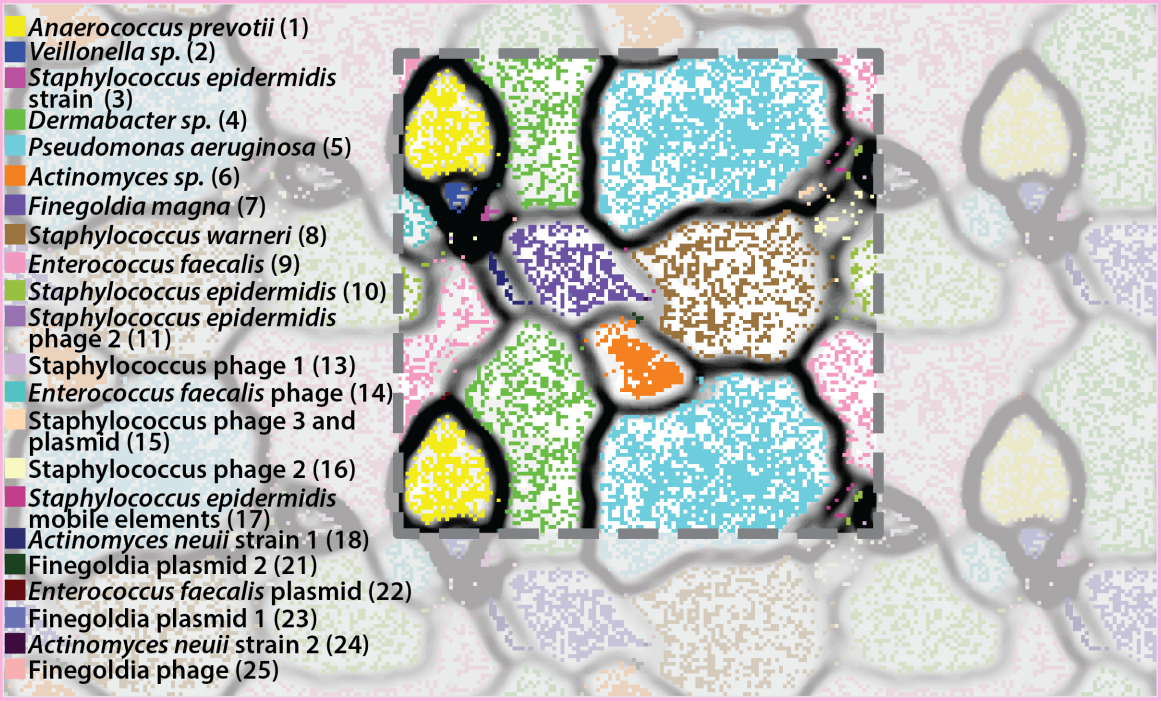
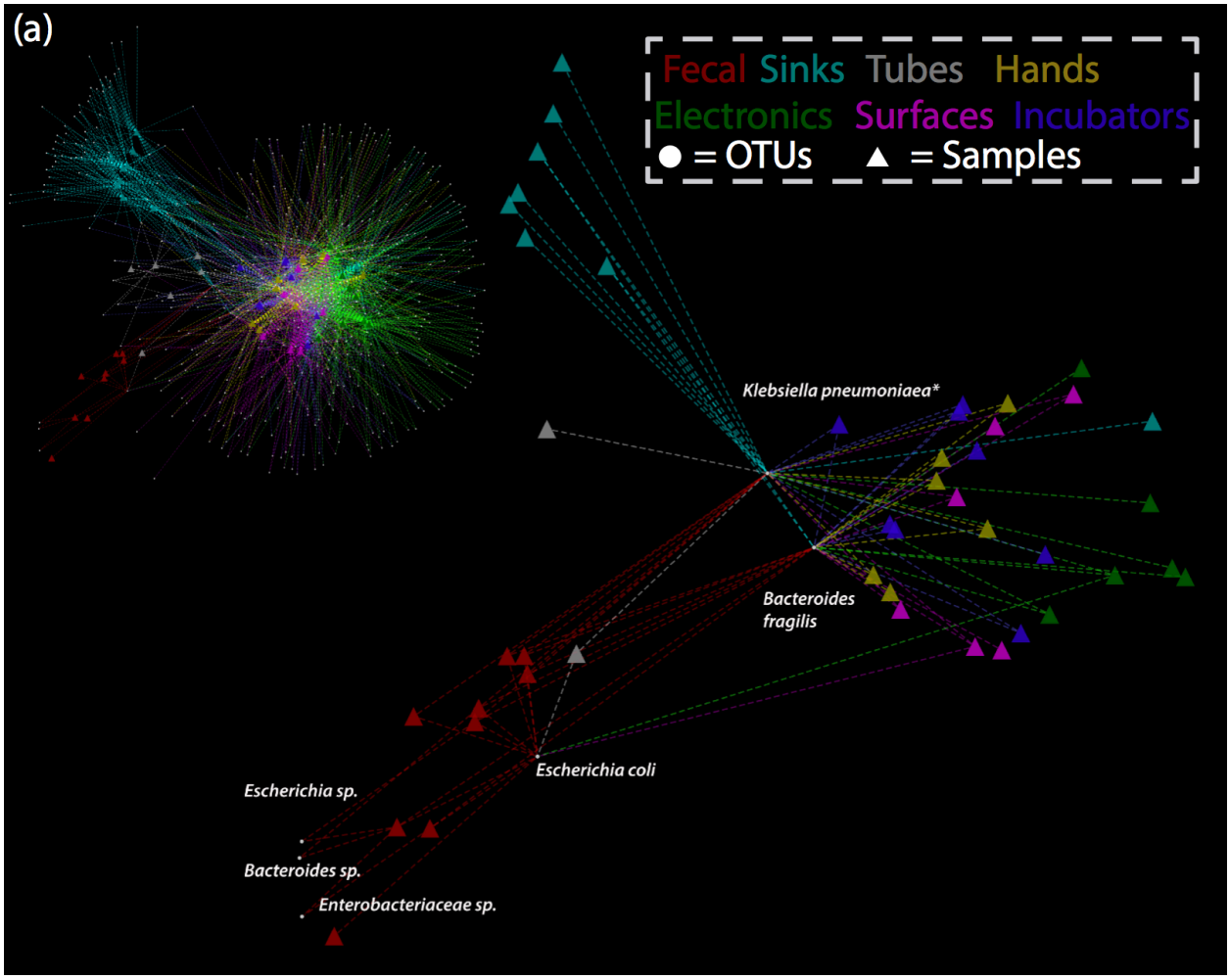


Figure 1-4: Spring weighted edge-embedded network plots of room and fecal OTUs

Spring weighted edge-embedded network plots of room and fecal OTUs found in two or more samples (Infant 1 (a), Infant 2 (b)). Left, the entire network is displayed. To better visualize the distribution of gut colonizers across room samples, only room samples sharing fecal OTUs are shown in the excerpt (right). Triangles represent samples and circles represent OTUs. Spring weight is derived from EMIRGE generated abundances and edges are colored by environment type. Each OTU has a taxonomic label and asterisks indicate OTUs detected in room samples before detection in the gut.



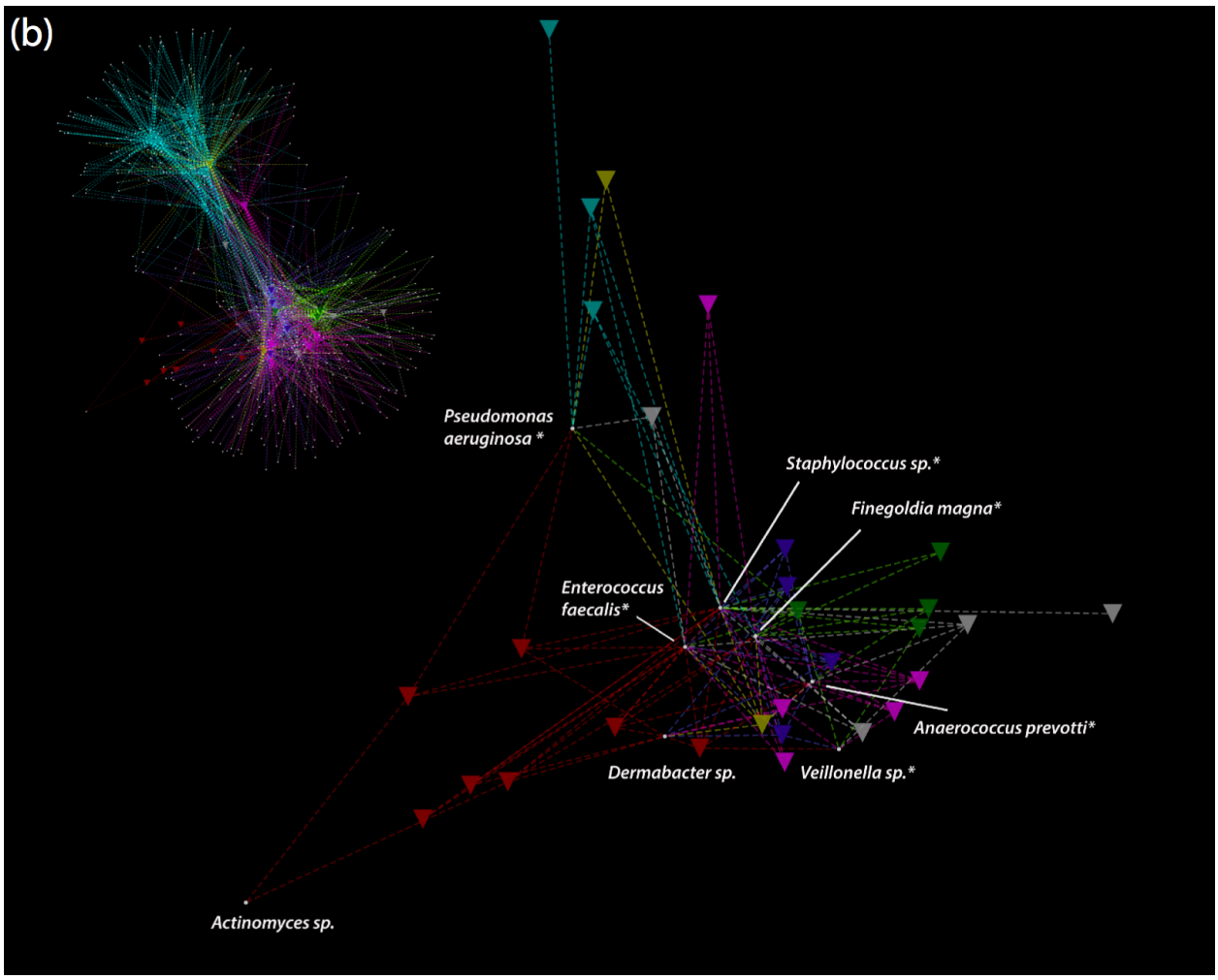

(b)
Pseudomonas aeruginosa *
Staphylococcus sp.*
Finegoldia magna*
Enterococcus faecalis*
Anaerococcus prevotti*
Dermabacter sp.
Veillonella sp.*
Actinomyces sp.

Figure 1-5: Community composition of gut colonizing microbes and room microbes through the first month of life

Time-series characterization of the fecal microbial community (left) and fecal microbes concurrently collected from the room (right) display discrete reservoirs of gut colonizers in the NICU.

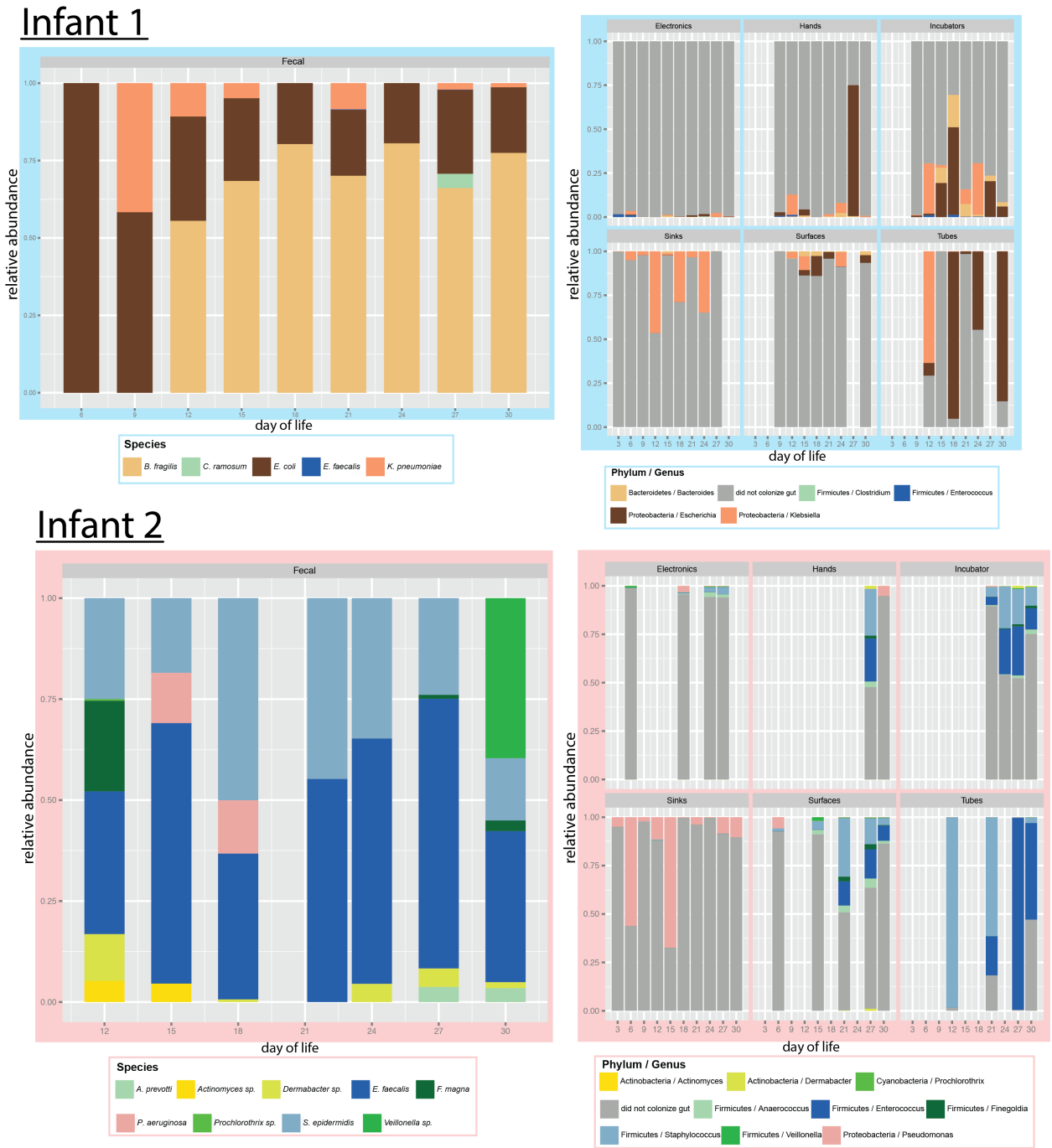


Figure 1-6: The most probable source of gut colonizing microbes

The most probable source of gut colonizing microbes was generated using the source-sink characterization software, SourceTracker (Knights *et al.*, 2011). NICU room sequences were designated as putative sources and fecal sequences sinks.

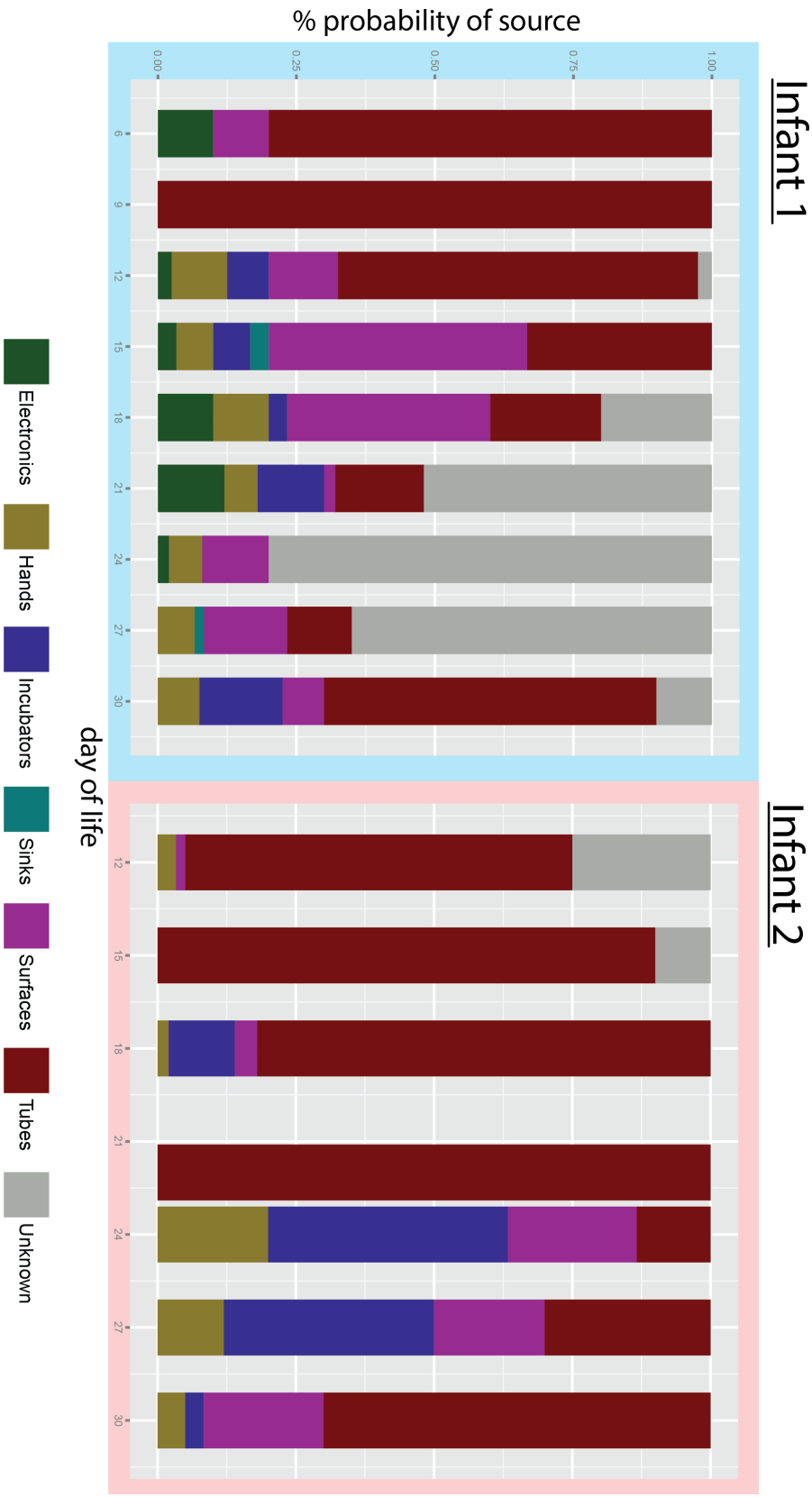

% probability of source
0.00
0.25
0.50
0.75
1.00
Infant 1
Infant 2
day of life
6
9
12
15
18
21
24
27
30
Electronics
Hands
Incubators
Sinks
Surfaces
Tubes
Unknown

Figure 1-7: *Enterococcus faecalis* phylogeny using 32 concatenated ribosomal proteins reveals closely related strains

Maximum likelihood phylogeny of *E. faecalis* strains was based on a concatenation of single-copy, highly conserved ribosomal proteins from our data set and available reference genomes. Bootstrap values greater than 100 are shown. An excerpt of the *E. faecalis* clade is shown to the right.

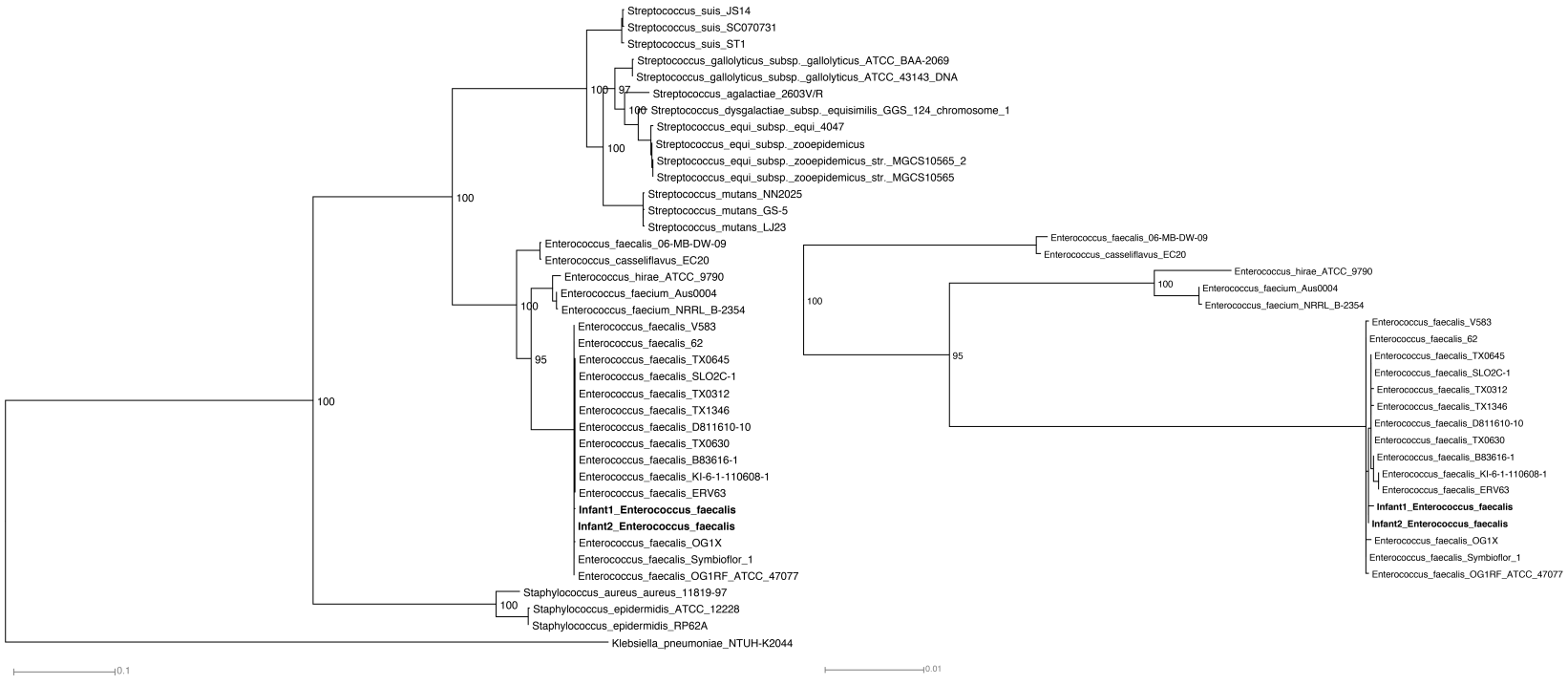

Streptococcus_suis_JS14
Streptococcus_suis_SC070731
Streptococcus_suis_ST1
Streptococcus_gallolyticus_subsp._gallolyticus_ATCC_BAA-2069
Streptococcus_gallolyticus_subsp._gallolyticus_ATCC_43143_DNA
Streptococcus_agalactiae_2603V/R
Streptococcus_dysgalactiae_subsp._equisimilis_GGS_124_chromosome_1
Streptococcus_equi_subsp._equi_4047
Streptococcus_equi_subsp._zooepidemicus
Streptococcus_equi_subsp._zooepidemicus_str._MGCS10565_2
Streptococcus_equi_subsp._zooepidemicus_str._MGCS10565
Streptococcus_mutans_NN2025
Streptococcus_mutans_GS-5
Streptococcus_mutans_LJ23
Enterococcus_faecalis_06-MB-DW-09
Enterococcus_casseliflavus_EC20
Enterococcus_hirae_ATCC_9790
Enterococcus_faecium_Aus0004
Enterococcus_faecium_NRRL_B-2354
Enterococcus_faecalis_V583
Enterococcus_faecalis_62
Enterococcus_faecalis_TX0645
Enterococcus_faecalis_SLO2C-1
Enterococcus_faecalis_TX0312
Enterococcus_faecalis_TX1346
Enterococcus_faecalis_D811610-10
Enterococcus_faecalis_TX0630
Enterococcus_faecalis_B83616-1
Enterococcus_faecalis_KI-6-1-110608-1
Enterococcus_faecalis_ERV63
Infant1_Enterococcus_faecalis
Infant2_Enterococcus_faecalis
Enterococcus_faecalis_OG1X
Enterococcus_faecalis_Symbioflor_1
Enterococcus_faecalis_OG1RF_ATCC_47077
Staphylococcus_aureus_aureus_11819-97
Staphylococcus_epidermidis_ATCC_12228
Staphylococcus_epidermidis_RP62A
Klebsiella_pneumoniae_NTUH-K2044
100
97
100
100
100
100
95
100
100
95
100
0.1
0.01

Table 1-1: Health profile of premature infant cohort

	Infant 1	**Infant 2**
Gestational age	26 3/7 wk	28 2/7 wk
Weight	951 g	1148 g
Multiple gestation	No	Twin
Delivery mode	vaginal	vaginal
Chorioamnionitis	yes	yes
DOL 1-7 antibiotics	ampicillin, gentamycin	ampicillin, gentamycin
Other antibiotics	No	DOL 14-16, vancomycin, cefotaxime
Feeding initiated	DOL 3, maternal milk	DOL 8, artificial formula
Survive to discharge	yes	yes

Table 1-2: Sample collection summary and summary of the number of 16S rRNA genes assembled

	Infant 1	**Infant 2**
Number of Samples		
Electronics	10	4
Surfaces	7	5
Incubator	8	4
Sink	9	10
Hands	8	2
Tubes	6	4
Fecal	9	7
Total	57	36
Number of EMIRGE Sequences		
Electronics	3359	1298
Surfaces	2440	2205
Incubators	2270	1751
Sinks	2936	4766
Hands	1783	812
Tubes	272	198
Fecal	33	32
Total	13093	11062
Number of OTUs		
Electronics	3353	1293
Surfaces	2436	2197
Incubators	2264	1749
Sinks	2933	4762
Hands	1781	812
Tubes	271	198
Fecal	33	32
Total	13071	11043
Shared OTUs	3822	
Number of unique OTUs		
Electronics	2486	1202
Surfaces	2211	2015
Incubators	2048	1606
Sinks	2756	4453
Hands	1603	801
Tubes	256	185
Fecal	11	11
Total	10371	10273

Table 1-3: Alpha diversity indexes from NICU room and fecal samples

	Shannon		**Simpson**		**Chao1**	
Infant	1	2	1	2	1	2
Surfaces	8.43	8.76	0.997	0.998	4.3×10^4	4.7×10^4
Electronics	8.36	8.28	0.997	0.997	4.6×10^4	3.4×10^4
Incubators	8.11	8.76	0.996	0.998	3.0×10^4	7.6×10^4
Sinks	8.29	8.83	0.997	0.998	4.1×10^4	9.7×10^4
Hands	7.56	8.61	0.993	0.997	2.9×10^4	8.9×10^4
Tubes	5.06	5.21	0.962	0.964	1.8×10^3	1.8×10^3
Fecal	1.71	2.10	0.641	0.748	9.7	13.7

Table 1-4: Genome Summaries

Infant 1							
Taxa	**Bin #**	**bp**	**Contigs**	**N50**	**%GC**	**Cvg**	**%SCG**
Bacteroides fragilis	6	4,551,095	39	249,654	43.3	1,930.3	99
Bacteroides phage1	4	205,842	1	205,842	41.9	2,221.4	0
Bacteroides phage2	5	144,903	1	144,903	42.0	2,060.8	0
Enterococcus faecalis	8	2,649,897	93	40,945	37.8	7.6	99
Clostridium ramosum	7	3,630,043	63	78,436	31.4	23.5	99
Escherichia coli	3	5,035,302	53	218,574	50.5	1,254.1	57
Klebsiella pneumoniae	1	5,447,442	78	189,741	57.3	345.0	37
Staphylococcus epidermidis plasmid	2	20,739	2	11,095	31.5	14.5	0
Infant 2							
Taxa	**Bin #**	**bp**	**Contigs**	**N50**	**%GC**	**Cvg**	**%SCG**
Actinomyces neuii strain 1	18	1,580,717	37	280,583	56.9	15.6	27
Actinomyces neuii strain 2	24	2,375,188	27	179,095	56.7	17.6	70
Actinomyces sp.	6	2,666,449	11	345,356	59.3	55.4	99
Anaerococcus prevotii	1	1,599,845	13	225,571	33.1	39.2	99
Caudovirales bacteriophage	26	18,308	1	18308	29.5	1169.7	0
Dermabacter sp.	4	2,040,279	12	289,797	62.8	51.9	90
Enterococcus faecalis	9	3,011,019	26	499,183	37.1	147.3	99
Enterococcus faecalis phage	14	335,286	39	12,896	34.8	103.7	0
Enterococcus faecalis plasmid	22	8,514	2	4,866	30.4	90.6	0
Finegoldia magna	7	1,729,913	42	78,482	32.0	93.0	99
Finegoldia phage	25	3,168	1	3,168	32.3	138.5	0
Finegoldia plasmid 1	23	7,589	2	3,969	33.0	103.4	0
Finegoldia plasmid 2	21	28,958	3	15,674	55.4	10.9	0
Pseudomonas aeruginosa	5	6,755,599	64	212,603	66.0	51.5	99
Staphylococcus epidermidis	10	1,902,759	82	40,484	33.0	65.4	7
Staphylococcus epidermidis mobile	17	55,503	10	6,452	31.7	54.5	43
Staphylococcus epidermidis phage 2	11	19,082	2	12,983	29.4	84.3	0
Staphylococcus epidermidis strain	3	81,754	9	14,965	29.4	67.1	0
Staphylococcus phage 1	13	216,785	13	8,080	29.5	45.7	0
Staphylococcus phage 2	16	198,742	14	20,782	0.3	79.3	0
Staphylococcus phage 3 and plasmid	15	137,609	12	19,343	29.3	67.8	0
Staphylococcus warneri	8	2,363,750	22	198,467	32.8	33.9	53
Veillonella sp.	2	2,281,484	223	12,637	37.8	56.2	70

Chapter 2:

2 Strain-resolved microbial community proteomics reveals simultaneous aerobic and anaerobic function during early stage gastrointestinal tract colonization

Brandon Brooks[1], Ryan S. Mueller[2,3], Jacque C. Young[4,5,7], Michael J. Morowitz[6], Robert L. Hettich[4,5], Jillian F. Banfield[2]

1 – Department of Plant and Microbial Biology, University of California, Berkeley, CA
2 – Department of Earth and Planetary Sciences, University of California Berkeley, Berkeley, CA
3 – Department of Microbiology, Oregon State University, Corvallis, OR (current address)
4 – Graduate School for Genome Science and Technology, University of Tennessee, Knoxville, TN
5 – Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN
6 – Department of Surgery, University of Pittsburgh School of Medicine, Pittsburgh, PA
7 – Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA (current address)

This material was published in an open access journal and is freely available here (Brooks *et al.*, 2015): http://journal.frontiersin.org/article/10.3389/fmicb.2015.00654/abstract

2.1 Abstract

While there has been growing interest in the gut microbiome in recent years, no studies have coupled strain-resolved community metagenomics with high-throughput mass spectrometry-based proteomics to study functional changes in the microbial community during infant gut colonization. Here, we implemented this approach to characterize fecal samples collected on days of life 5-21 from an infant born at 28 weeks gestation. Genomic sequences were manually curated for strain-level resolution of populations of *Citrobacter*, an abundant organism present during the later stage of colonization. Proteome extracts from fecal samples were processed via a nano-2D-LC-MS/MS and searched against the metagenomic database. The results show a microbial community dominated by facultative anaerobes. We document the utilization of both aerobic and anaerobic metabolisms throughout the time series, likely indicating growth in distinct niches within the gastrointestinal tract. Additionally, we uncover differences in the physiology of coexisting *Citrobacter* strains, including differences in motility and chemotaxis functions, and resolve a specific community-essential role in vitamin metabolism and a predominant role in propionate production for this organism. Finally, we detect differences between genome abundance and activity levels for the dominant populations, underlining the value in layering proteomic information over genetic potential.

2.2 Introduction

The human gastrointestinal tract (GIT) harbors a complex ecosystem of microorganisms, the microbiome, whose cell count outnumbers the cells of the human body by nearly ten to one (Smith, 2014). The genes of the microbiome encode byproducts critical for host health and development (Groer *et al.*, 2014). Recent excitement in the field has been generated from findings implicating the microbial community in a variety of dysbioses from gut associated diseases like obesity and malnutrition (Turnbaugh *et al.*, 2006; Smith *et al.*, 2013), inflammatory bowel disease (Hold, 2014), and celiac disease (Nistal *et al.*, 2012) to neurological disorders like depression (Park *et al.*, 2013), anxiety (Diaz Heijtz *et al.*, 2011) and autism (Hsiao *et al.*, 2013). While significant contributions have been made to understand developed microbial communities in healthy and diseased adults, large gaps remain in understanding the acquisition of the human microbiome at birth, especially among preterm infants (Groer *et al.*, 2014).

In utero, infants have a sparse microbiome (Ardissone *et al.*, 2014), with the first major microbial inoculum encountered during the birthing process. Delivery mode, i.e. vaginal versus caesarean section, can play a significant role in how a baby is colonized (Dominguez-Bello *et al.*, 2010) as can dietary input, breast milk versus formula (Guaraldi and Salvatori, 2012), and exposure to antibiotics (Groer *et al.*, 2014). For example, infants born vaginally acquire a community more similar to the mother's vaginal and fecal microbiota whereas infants born by caesarean section have a microbiome that is more similar to those of skin and hospital environments (Brooks *et al.*, 2014; Dominguez-Bello *et al.*, 2010). Caesarean section infants appear to have lower microbial richness and diversity relative to vaginally born infants at four months of age (Song *et al.*, 2013). Throughout the first year of life the microbial community increases in diversity, reaching an adult-like state around 2.5 years of life (Koenig *et al.*, 2011). The long term health effects of different colonization paths remains to be determined, but with the many direct and indirect effects of the microbiome, it is likely to play a critical role in the development of many diseases. Understanding dynamics that govern colonization, and ultimately

defining a healthy colonization trajectory is critical, especially for preterm infants that are susceptible to numerous infections and developmental issues.

Very low birth weight (VLBW) infants accounted for approximately 35% of all infant deaths in 2009 (Groer *et al.*, 2014). These infants have an increased risk for cardiorespiratory, hematological, gastroenterologic, infectious, and neurological disorders (Groer *et al.*, 2014). Most spend several months of their early lives in the neonatal intensive care unit (NICU), where administration of antibiotics is commonplace. Among VLBW infants, incidence rates of sepsis and necrotizing entercolitis (NEC) remain high (Bizzarro *et al.*, 2014). Both diseased and non-diseased VLBW infants are characterized by low bacterial diversity, abrupt shifts in community composition (which can be phage mediated (Sharon *et al.*, 2013)), and an abundance of opportunistic pathogens (Sharon *et al.*, 2013), relative to their full term counterparts. Most opportunistic pathogens in VLBW infants are facultative anaerobes. Typically during the first weeks of life, there is a shift from facultative to obligate anaerobes (Penders *et al.*, 2006). Because facultative anaerobes are capable of growth with and without oxygen, their mode of growth cannot be determined from genomic sequence information alone. Further, organisms may be abundant but characterized by low activity levels, or vice versa. Here, we coupled strain-resolved community metagenomics data with mass spectrometry-based proteomics to resolve growth mode and to compare activity levels during colonization of a preterm infant. The samples collected during the first month of life for this VLBW infant were ideal for metaproteomic study because curated genomes for the dominant organisms were available and because the communities contained a limited number of highly abundant organisms (Morowitz *et al.*, 2010a), enabling deep proteomic analysis. We identified differences in metabolic potential and protein abundance levels in closely related strains, determined that both aerobic and a variety of anaerobic pathways were operational, and confirmed differences between genome abundance and metabolic activity.

2.3 Materials and Methods

2.3.1 Infant description and sample collection

The female infant was delivered by caesarean section at 28 weeks gestation after premature rupture of membranes. The infant received antibiotics (ampicillin/gentamicin) for the first seven days of life (DOL). Breast milk enteral feeding was administered on DOL 4-9 but was stopped on DOL 9-13 because of abdominal distension. Enteral feeding was slowly resumed on DOL 13 with artificial formula (Similac Special Care 20 calories per fluid ounce; Abbott Nutrition). Additionally, parenteral nutrition was provided until caloric intake from enteral nutrition was adequate (DOL 28). Fresh fecal samples were collected on DOL 5-21 as available using a previously described technique (Morowitz *et al.*, 2010a). Informed parental consent was obtained before patient enrollment and research protocol approved by the Institutional Review Board of The University of Chicago (protocol # 15895A).

2.3.2 Protein extraction, digestion, and Nano-2D-LC-MS/MS

Complete details of protein extraction, digestion, and nanospray-two dimensional liquid chromatography coupled with tandem mass spectrometry (nano-2D-LC-MS/MS) are reported elsewhere (Young *et al.*, 2015). Briefly, fecal material was boiled in Tris-Cl containing SDS and

DTT, and underwent bead beating for 30 minutes to lyse cells and denature proteins. The supernatant was collected, boiled again, spun down, and precipitated overnight. Protein pellets were washed, re-solubilized, and sonicated to break up the pellet. Iodoacetamide was added to block disulfide bond reformation. Proteins were then diluted, and enzymatically digested using sequencing grade trypsin (Promega). Peptides were diluted, a second dose of trypsin added, and digestion continued overnight. An acidic salt solution was used to clean up the peptides, which were then spun through a 10 kDa cutoff spin column filter (VWR).

Peptides were then loaded onto a split-phase fused silica column containing reverse phase (C18) and strong cation exchange (SCX) materials. Samples were washed and placed in line with a nanospray emitter (New Objective) packed with reverse phase material then separated on-line using high performance two-dimensional liquid chromatography. Peptides were eluted, ionized via nanospray (200 nl/min) (Proxeon, Cambridge MA), and analyzed using an LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific, San Jose, CA). The mass spectrometer was run in data-dependent mode with the top ten most abundant peptides in full MS selected for MS/MS, and dynamic exclusion enabled (repeat count=1, 60 s exclusion duration). Full MS scans were collected in the Orbitrap at 30K resolution. Two microscans were collected in centroid mode for both full and MS/MS scans. Technical duplicates were run for all samples.

2.3.3 Database composition and peptide matching

A search database was generated from manually curated genomes assembled from metagenomic reads previously published (Morowitz *et al.*, 2010a). Four taxa dominate these samples: *Serratia* (*UC1SER*), major and minor strain *Citrobacter* (*UC1CIT* and *UC1CITii*), and *Enterococcus* (*UC1ENC*). These taxa represent approximately 75% of the community composition, with the remaining 25% apportioned to several low ranking taxa (Morowitz *et al.*, 2010a). Low ranking taxa were excluded from the database to focus on organisms with higher peptide coverage. Strain-level variation between *Citrobacter* strains was resolved manually using Strainer version r-34 (Eppley *et al.*, 2007). MS/MS spectra were searched against the concatenated database using MyriMatch version 2.1.111. The protein database is publicly accessible at http://ggkbase.berkeley.edu/UC1/, and the MS raw files have been uploaded to ProteomeXchange Consortium with the dataset identifier number px-submission PXD000114.

2.3.4 Pathway analysis

To summarize metabolic potential, we compiled lists of genes using annotation search terms implemented via the ggKbase lists function. List search terms were manually compiled, and made use of EC numbers, KO numbers, and other search terms to describe pathways. ggKbase is an online tool for genome binning, metabolic pathway curation and community composition analysis. The current dataset is available at http://ggkbase.berkeley.edu/UC1/. To easily visualize both metabolic potential and protein expression, an expression versus potential ratio was plotted across all available lists and a subset of curated ggKbase lists (Figure 2-3 and Figure 2-4, respectively). This ratio is the non-redundant count of features per list that were identified via proteomics, divided by the count of features per list. ggKbase lists are dynamic, so a static version linking genes to lists is available at http://journal.frontiersin.org/article/10.3389/fmicb.2015.00654/full.

2.4 RESULTS AND DISCUSSION

2.4.1 General proteome description

We characterized fecal proteome extracts from seven fecal samples via nano-2D-LC-MS/MS (DOL 13-21) and uniquely identified 1149 to 2636 microbial proteins per sample based on 4300 to 15370 distinct spectra. Additionally, many human proteins were detected. A detailed analysis is described in Young *et al*. (2015). In total across all samples, we detected approximately 1000 proteins per organism for each of the three most abundant organisms. These quantities represent coverage of 22% of the predicted proteomes of these organisms (Table 2-1). On average, approximately 550 proteins per organism were identified with unique peptides in each sample. DOL 13 samples exhibited extremely low peptide detection with unique spectral matches totaling 190 and proteins with unique spectral matches totaling 79. This sample was excluded from most analyses unless explicitly stated.

2.4.2 Microbial community profile and general functional characterization

To survey the microbial community, we compared results from read and peptide mapping to the metagenomics derived database. Read mapping results confirmed the dominance of *UC1SER*, with lower abundances of *UC1CITs* and *UC1ENC* (Morowitz *et al.*, 2010a). Even when the *UC1CITs* are taken together, *Citrobacter* reads are less abundant than *Serratia*. Interestingly, the matched proteomic data indicate that, in combination, the *Citrobacter* account for the largest proportion of the proteome, suggesting that the activity level of these organisms is higher than that of *UC1SER* and *UC1ENC*. The apparent difference in cell abundance compared to activity, based on differences between read mapping and peptide mapping, is most pronounced on DOL 18. For this sample, the read count data indicate *UC1SER* comprised ~ 60% of the community but its proteins only accounted for 35% of the community proteome (Figure 2-1).

2.4.3 Aerobic and anaerobic respiration

During the latter phase of colonization, the infant was supplied with infant formula. Lactose, an abundant constituent in infant formula, can be respired aerobically or fermented. During the period of formula feeding, the community was dominated by *Serratia* and *Citrobacter* strains. These species can grow both aerobically and anaerobically, and capacities for both growth modes are encoded in the genomes of the *Citrobacter* and *Serratia* strains studied here (Figure 2-2). Evidence for respiration-based metabolism includes detection of proteins from the essentially complete TCA cycle (Figure 2-3). A strong indicator of aerobic respiration is the identification of most enzymes of the electron transport chain, including cytochrome c proteins and multiple terminal oxidases (Figure 2-4). The aerobic growth pathway was operational throughout this colonization phase in both organisms. Given that the likely source of O_2 is the intestinal tissue, and that the O_2 gradient decreases toward the (Albenberg *et al.*, 2014), it seems likely that *UC1CITs* and *UC1SER* growing aerobically are localized towards the mucosa.

Products of glycolysis could also be respired anaerobically, given the presence of pathways for nitrate, nitrite, and sulfate reduction in the genomes (Figure 2-2). The mass

spectrometry measurements identified many *Citrobacter* enzymes likely involved in anaerobic respiration, including proteins from all of these pathways, excluding sulfate reduction (Figure 2-4). The *Serratia* proteome also included proteins associated with many of these functions. Also expressed are genes involved in the anaerobic reduction of dimethyl sulfoxide and formate. Nitrate reductase proteins were particularly abundant across all time points; this finding was more pronounced for *UC1CITs* but still detectable in *UC1SER*. The source of nitrate is likely the host's immune response through various inflammatory pathways. Nitrate availability in the gut has been shown to give Enterobacteriaceae a fitness advantage over obligate anaerobes (Winter *et al.*, 2013). Citrobacter also expresses nitric oxide dioxygenase, which is involved in aerobic detoxification of NO, presumably protecting the bacterium (and the community) from various toxic nitrogen compounds.

2.4.4 Fermentation pathways

As with the human milk oligosaccharides they mimic, formula oligosaccharides can be fermented to short chain fatty acids (SCFAs). The genomes of all microorganisms present in the third phase of colonization encode a variety of fermentation pathways and there is clear proteomic evidence for fermentation-based metabolism in both *Citrobacter* strains, the *Serratia* strain and *Enterococcus faecalis*. These pathways generate SCFAs that are likely absorbed by the infant.

A particularly abundant pathway in *Citrobacter* for which proteins were detected was for the fermentation of fucose to propionate (Figure 2-4 and Figure 2-5). L-fucose isomerase, the first enzyme needed to degrade L-fucose to L-fuculose, was identified in two samples. The adjacent gene, L-fuculokinase, responsible for conversion of 1-fucose 1-phosphate, was not identified. L-fucose aldolase, also encoded in this region, converts 1-fucose 1-phosphate to L-lactaldehyde; this protein was also not identified, but the adjacent fructose operon regulator was identified in one sample.

Although both *UC1CITs* have pathways for the anaerobic degradation of rhamnose as well as fucose, the genes for rhamnose degradation (and transport) were not detected.

The protein encoded by the next gene in the anaerobic fucose degradation pathway converts L-lactaldehyde to 1,2-propanediol (1,2-PD). This protein was identified in all samples. *Citrobacter* also can convert 1,2-PD to propionyl-CoA, and probably does so within a well-characterized organelle (a microcompartment), which prevents the accumulation of toxic aldehyde intermediate (Kerfeld and Erbilgin, 2014). Shell proteins for this microcompartment, specifically shell protein PduA, were consistently identified in samples from most days. Propionyl-CoA is likely degraded to propionyl phosphate then to propionate (a SCFA) as propionate kinase, the final enzyme that converts propyionyl phosphate to propionate, was consistently identified in both *UC1CITs*. Propionate may then be excreted, and is likely to have been absorbed by the infant (Tan *et al.*, 2014). Interestingly, *Serratia* does not appear to have the capacity to ferment either fucose or rhamnose, which may be the metabolic basis of their niche separation.

Another prominent fermentative pathway found in the *Citrobacter* proteomes involves enzymes that degrade glycerol, several of which are vitamin B_{12}-dependent. Production of vitamin B_{12} (cobalamin) is unique to bacteria and archaea, and is an essential cofactor for many forms of life. We consistently detected proteins required for the biosynthesis of vitamin B_{12}, specifically CbiG and CbiK, from *UC1CIT*. The *UC1CITs* are the only relatively abundant

organisms in the infant's gut that encode cobalamin biosynthesis genes, and consistent expression of this pathway suggests it to be a key role in the community (Figure 2-3). Notably, these cobalamin biosynthesis enzymes operate under anaerobic conditions, a further indication of anaerobic niches in the gut during this phase of colonization.

Additionally, enzymes were identified for a fermentation pathway that converts glycerol to 1,3-propanediol (1,3-PD) and other SCFA byproducts (using the glycerol dehydratase complex: EC:4.2.1.30; three subunits, all of which were identified by proteomics in all samples). *Citrobacter* is one of a small number of bacterial genera with the glycerol fermentative pathway (others include *Klebsiella*, *Clostridium*, and *Lactobacillus*). The SCFA byproducts of this pathway are acetate and sometimes butyrate (Abbad-Andaloussi *et al.*, 1996). For the latter, the enzymes required for acetyl-CoA conversion to butyrate are poorly maintained within the genomes of each of these genera, and only select strains contain them (Louis *et al.*, 2004). *Citrobacter* has the genes to convert acetyl CoA to crotonoyl-CoA (e.g., for amino acid biosynthesis), but lacks those required to form butanoate.

For glycerol breakdown, *Serratia* lacks the glycerol dehydratase complex found in *Citrobacter* (EC:4.2.1.30). However, it has glycerol kinase (EC:1.1.1.6) and glycerone kinase (EC: 2.7.1.29), allowing it to convert glycerol to glycerone phosphate, potentially for consumption via glycolysis. Both of these enzymes were identified by proteomics, although only in the day 21 sample. *Serratia* also has glycerophosphoryl diester phosphodiesterase (EC:3.1.4.46) that converts alpha glycerophosphodiester to *sn*-glycerol-3-phosphate. This is also the product of glycerol kinase (EC:2.7.1.30, which was identified by proteomics in all samples). The *sn*-glycerol-3-phosphate can be degraded by glycerol-3-phosphate dehydrogenase (EC: 1.1.5.3) to dihydroxyacetone phosphate (glycerone phosphate), and some of these proteins were identified. *Serratia* then combines *sn*-glycerol-3-phosphate with acyl CoA to form 1-acyl-*sn*-glycerol 3-phosphate (identified by proteomics in one sample). These, and other proteins, are likely redirected for use in lipid biosynthesis.

We also identified multiple *Serratia* proteins of the inositol degradation pathway, including inositol 2-dehydrogenase, myo-inositol catabolism protein IolH, inosose dehydratase, 3D-(3,5/4)-trihydroxycyclohexane-1,2-dione hydrolase, and 5-dehydro-2-deoxygluconokinase. The strong representation of these enzymes indicates a potentially important role for *Serratia* in degradation of this compound, which is an important component of both breast milk and infant formula (Sharon *et al.*, 2013). *Citrobacter* also has some enzymes for inositol degradation. However, the identification of only two proteins from the *Citrobacter* pathway may indicate that inositol is a less important substrate for this organism compared to *Serratia*.

2.4.5 Motility, toxicity, and invasion

Several pathways enable microorganisms to cope with the gut immune system, respond to administered antibiotics, and to manage compounds produced by other microorganisms. Catalase, an enzyme used to protect cells from reactive oxygen species (ROS) by degrading hydrogen peroxide to water and oxygen, is consistently found in both *UC1CITs* and *UC1SER* on most days. Hydrogen peroxide can be produced by the intestinal epithelium and neutrophils during inflammation response, along with other ROS (Winter *et al.*, 2013). Other protective antioxidant proteins such as lipid hydroperoxide peroxidase, alkyl hydroperoxide reductase, superoxide dismutase, and glutathione peroxidase were identified in samples collected on most days, and were particularly abundant in *UC1CITs* and *UC1SER*.

Often ROS exposure occurs in close proximity to the host epithelium (Winter *et al.*, 2013). The ability to move away from ROS and towards a more favorable environment seems critical for microbes in our dataset. We identified many *UC1CITii* polar flagella-related proteins in samples collected on several days; no lateral flagella proteins were identified and flagella proteins were not identified for *UC1CIT* (Figure 2-4). Proteins for twitching motility were detected for *UC1CITii*. Additionally, proteins for chemotaxis, specifically chemotaxis protein methyltransferase CheR, are detected on all days in *UC1CITii*. Chemotaxis related proteins were not detected in the dominant strain, or *UC1SER*, possibly suggesting a more planktonic state for the minor strain. Perhaps the increased motility allows the minor strain to escape ROS, as expression levels and frequency of oxidative stress related genes are lower in the minor strain (Figure 2-4).

In contrast to chemotaxis and flagellar movement, proteins for biofilm formation and many fimbrial attachment proteins are detected in *UC1CIT* and *UC1SER*, though these are also detectable in *UC1CITii*. Type-1 fimbrial proteins, associated with capacity to attach to host gut epithelium (Juge, 2012), were detected consistently across the time series, as were proteins for biofilm regulation and formation. Transcriptional regulator and periplasmic proteins csgD and csgF, involved in curli biosynthesis, were also detected on one day in *UC1CITii* (Barnhart and Chapman, 2006). It is unclear whether curli expression in *UC1CITii* is promoting adherence to one another or host adherence, as it is capable of both (Barnhart and Chapman, 2006). *Citrobacter'*s affinity for host fucosylated glycans would suggest colonization of the mucosa.

Proteins that respond to both host and inter-species bacterial attack, such as the type VI secretion system (T6SS) in *UC1SER,* were detected on several days. T6SS were first investigated for their role in pathogenesis (Jani and Cotter, 2010), but have been studied more recently for "T6SS-dueling", a mechanism in which the T6SS can kill and outcompete neighboring microbes for resources (Basler *et al.*, 2013). Proteins involved in production of bacteriocin toxins like colicin were also detected on most days, but only from *U1CITii*. Many gram-negative bacteria can produce a variety bacteriocins that can target closely related species or strains that occupy similar niches (Kleanthous, 2010).

A competitive advantage for niche space (and persistence in the hospital environment) is also gained through various antibiotic resistance mechanisms. A variety of antibiotic resistance proteins were detected for all organisms across the time series, but most frequently in the *UC1CITs*. Most consistently detected for all organisms were efflux pumps. Efflux pumps can remove various toxins, waste products, and antibiotics (Fernández and Hancock, 2012). It was speculated that the multipurpose functionality of such pumps could aid in survival, both in and out of the host, as antibiotics and various biocides are encountered in both environments (Brooks *et al.*, 2014).

2.4.6 Comparison of major and minor *Citrobacter* strains

The two *UC1CITs* strains share 98.96% and 99.23% nucleotide and amino acid identity between their orthologs, respectively. Read mapping data confirms genome abundances previously published (Morowitz *et al.*, 2010a), with a drop in the minor strain population around DOL 18 (Figure 2-1). Interestingly, metabolic activity, as reflected in the proteome composition, does not correlate closely with genome abundance, as no decrease in protein spectral counts mapping to the minor strain occurs (Figure 2-1). This observation could reflect steady state growth of the abundant *Citrobacter* strain and rapid growth of the minor strain. This result

highlights the value of proteomic data for uncovering aspects of the dynamics not apparent from genome abundance information alone. Difference between genome abundance (from 16S rRNA gene surveys or metagenomics) and microbial activity levels have been reported previously.

Different methodologies can distinguish living from dead or inactive cells. For example, Maurice et al., distinguished subgroups based on nucleic acid concentrations and lack of membrane integrity (Maurice *et al.*, 2013). However, Franzosa *et al.* showed that RNA and DNA abundances were generally well correlated, with the exception of a few select pathways (Franzosa *et al.*, 2014). Erickson *et al.* reported varying degrees of incongruence between organisms based on mapped metagenomic reads versus peptide spectral counts in patients with Crohn's disease (Erickson *et al.*, 2012). To our knowledge, no genome abundance to protein expression difference, as detected here, has been reported for gut-associated microorganisms in infants.

We looked for differences in overall proteome composition for the *Citrobacter* strains, and for evidence for the production of proteins unique to one of the strains. Spectral coverage across the genomes was relatively complete, and generally, expression of the minor strain tracked with the major stain (Figure 2-5). Notable exceptions were the flagella and chemotaxis-related proteins, as noted above, and drops in coverage often associated with phage related regions, mobile elements, and some regions associated with transport and membrane proteins. Interestingly, several unique genes, genes encoded in one strain and not the other, were expressed. There are 233 and 84 genes not shared with the other strain in *UC1CIT* and *UC1CITii* respectively, and 25 and 10 of these proteins were detected via proteomics for *UC1CIT* and *UC1CITii*, respectively. These span broad functions from transcription, translation, and metabolism related proteins. From the major strain, novel proteins were identified on most days, such as carbamoyl phosphate synthase involved in pyrimidine and amino acid metabolisms, an alcohol dehydrogenase used in ethanol production, a transporter for glutathione binding, and an aromatic amino acid aminotransferase. Carbamoyl phosphate synthase is the rate-limiting step in L-arginine production and has been linked to an increase of NEC in preterm infants (Watkins and Besner, 2013). In the minor strain, ABC transporters and a DNA translocase FtsK variant involved in cell division and chromosome separation were detected on most days. In combination, the differences in proteome composition support the inference that the *Citrobacter* strains occupy distinct niches.

2.5 Concluding Remarks

The VLBW infant gut microbiome is relatively uncharacterized and little is known about microbial metabolism during the critical first few weeks of life. The opportunity for organisms to growth via both aerobic and anaerobic respiration might be anticipated to develop over the time period in which GI tract transitions from an aerobic to anaerobic state. However, the range of aerobic and anaerobic metabolisms detected at the same time may suggest heterogeneity in the developing gut in which facultative anaerobes are likely to dominate. Different niches may be associated with sub-populations of *Serratia* and the two *Citrobacter* strains in different gut niches. Further, metabolic differences between the *Citrobacter* strains support the suggestion that the populations occupy distinct niches. The distinct differences in inferred abundances and activity levels for these strains likely reflect changing opportunities occurring during this colonization phase.

2.6 Acknowledgements

We thank Dr. David Tabb for the DTASelect/Contrast software, and Langho Lee for assistance with proteomic bioinformatics. This work was partly funded by NIH grant 1R01-GM-103600, a March of Dimes Foundation grant 5-FY10-103 (M.J.M), and an NSF Graduate Fellowship to B.B. and stipend support from the Genome Science and Technology program at the University of Tennessee, Knoxville to J.C.Y.

2.7 Conflict of Interest

The authors declare no conflicts of interest.

Figure 2-1: Microbial community composition observed via read and peptide mapping

Relative proportion of reads (A) and unique peptides (B) mapped to a database of metagenomes derived from dominant gut colonizers in a preterm infant.

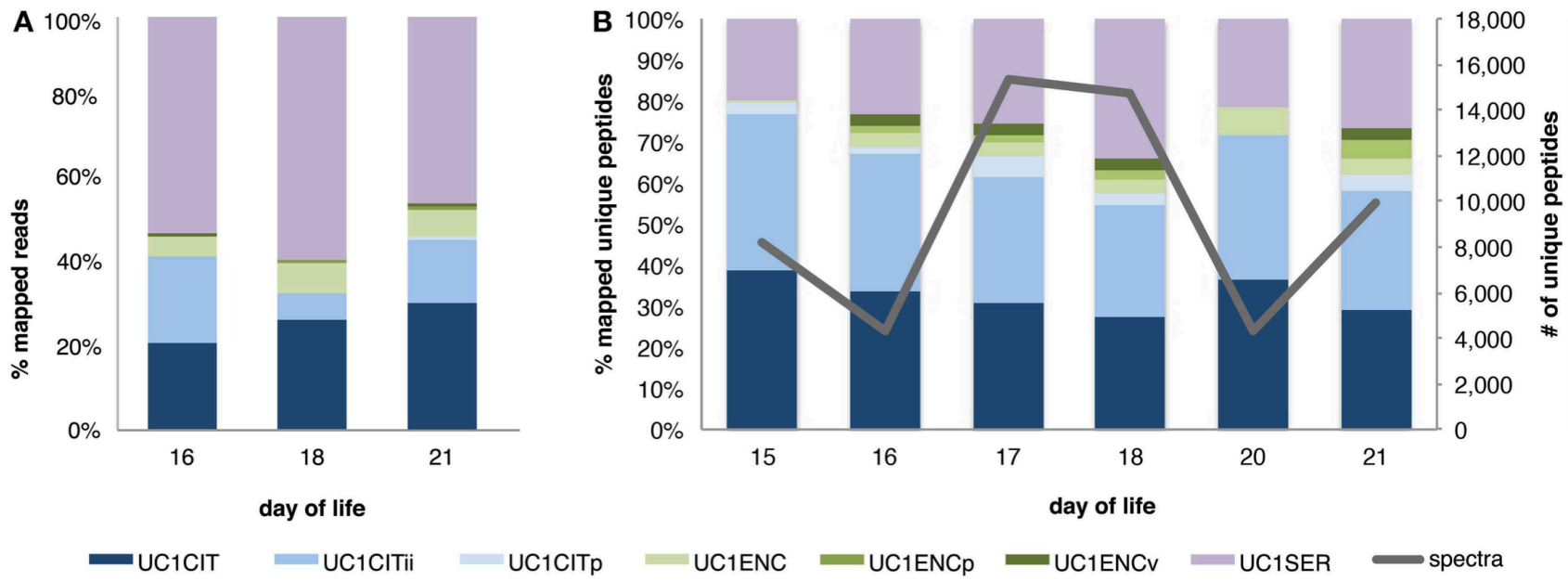

A
100%
80%
60%
40%
20%
0%
% mapped reads
16
18
21
day of life
B
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
% mapped unique peptides
18,000
16,000
14,000
12,000
10,000
8,000
6,000
4,000
2,000
0
# of unique peptides
15
16
17
18
20
21
day of life
UC1CIT
UC1CITii
UC1CITp
UC1ENC
UC1ENCp
UC1ENCv
UC1SER
spectra

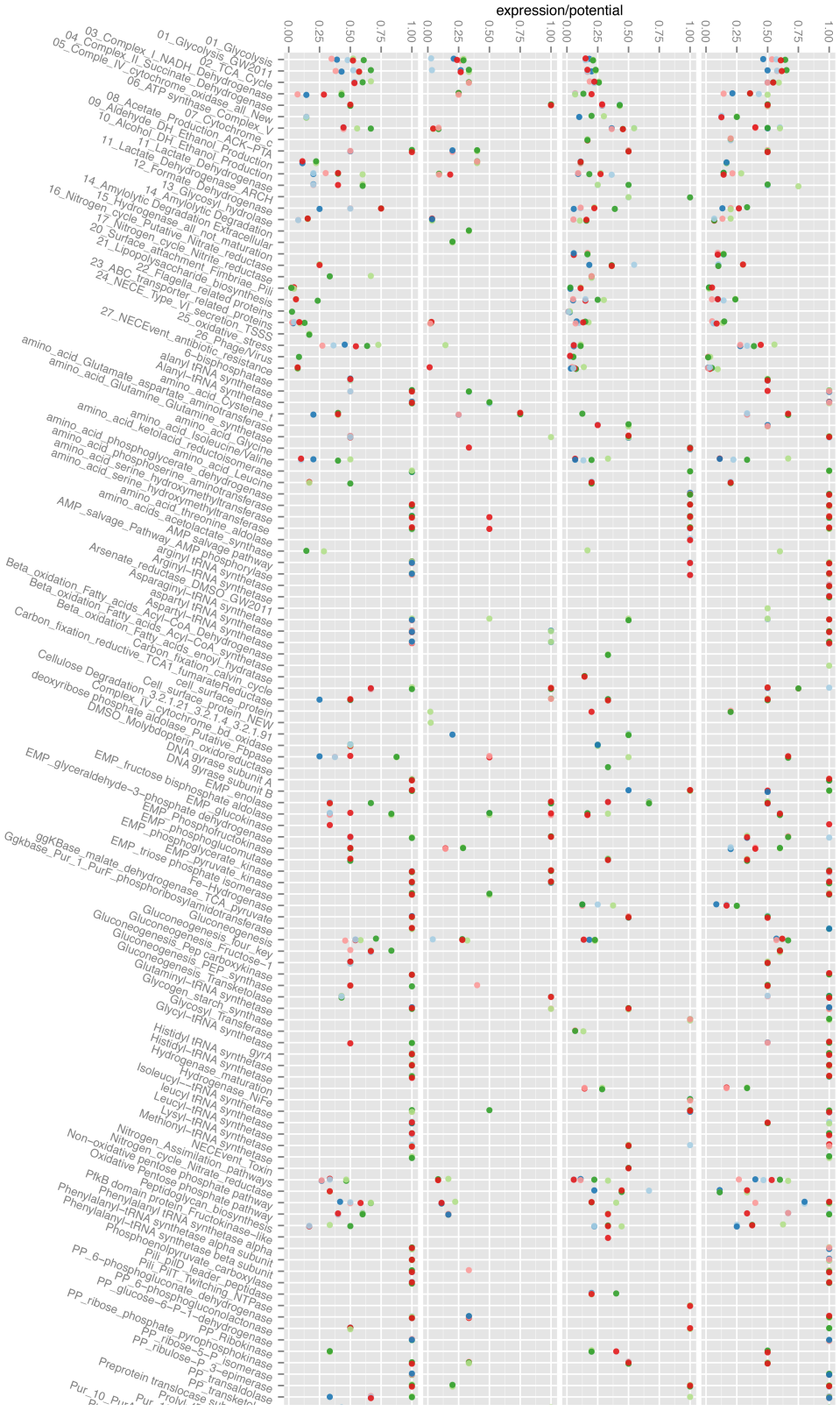
Figure 2-2: Metabolic potential of microbes colonizing a preterm infant gut

ggKbase lists illustrate the broad metabolic potential of microbes colonizing a preterm infant in the first month of life.

ETC
Fermentation
Carbon Degradation
Alt. electron acceptors
Misc. functions
Glycolysis
TCA_Cycle
Complex_I_NADH_Dehydrogenase
Complex_II_Succinate_Dehydrogenase
Comple_IV_cytochrome_oxidase_all_New COPY
ATP synthase_Complex_V
Cytochrome_c
Acetate Production (ACK-PTA)
Aldehyde_DH_Ethanol_Production
Alcohol_DH_Ethanol_Production
Lactate_Dehydrogenase
Formate_Dehydrogenase
Glycosyl_hydrolase
Amylolytic Degradation
Hydrogenase_all_(not_maturation)
Nitrogen_cycle_Putative_Nitrate_reductase
Nitrogen_cycle_Nitrite_reductase
Dissimilatory_Sulfate_Reduction
Cell_surface_Fibronectin
Surface_attachment_Fimbriae_Pili
Lipopolysaccharide_biosynthesis
Flagella_related proteins
ABC transporter related protein
NECE_Type_VI_secretion_TSSS
oxidative_stress
Phage/Virus
NECEvent_antibiotic_resistance
UC1ENC
UC1CIT
UC1CITii
UC1SER
36 3 4 1 4 23 5 5 11 4 2 29 6 3 2 1 41 21 80 202 1 7 58 69
28 22 14 4 8 10 5 4 6 14 4 15 15 3 21 10 6 3 1 2 21 80 81 1 18 59 63
31 27 15 7 10 11 4 6 9 11 4 18 19 3 18 11 5 6 1 36 20 86 114 1 18 77 95
23 15 14 4 7 9 3 2 9 10 5 4 13 2 4 3 4 3 47 17 38 140 6 11 12 69

Figure 2-3: Expression over potential ratio of infant gut microbes

A non-redundant count of the number of features identified via proteomics in a metabolic ggKbase list was divided by the number of features in that list and plotted for each organism using separate symbol colors to represent each day of life time point.


expression/potential
0.00
0.25
0.50
0.75
1.00
01_Glycolysis
01_Glycolysis_GW2011
02_TCA_Cycle
03_Complex_I_NADH_Dehydrogenase
04_Complex_II_Succinate_Dehydrogenase
05_Comple_IV_cytochrome_oxidase_all_New
06_ATP synthase_Complex_V
07_Cytochrome_c
08_Acetate_Production_ACK-PTA
09_Aldehyde_DH_Ethanol_Production
10_Alcohol_DH_Ethanol_Production
11_Lactate_Dehydrogenase
11_Lactate_Dehydrogenase_ARCH
12_Formate_Dehydrogenase
13_Glycosyl_hydrolase
14_Amylolytic Degradation
14_Amylolytic Degradation Extracellular
15_Hydrogenase_all_not_maturation
16_Nitrogen_cycle_Putative_Nitrate_reductase
17_Nitrogen_cycle_Nitrite_reductase
20_Surface_attachment_Fimbriae_Pili
21_Lipopolysaccharide_biosynthesis
22_Flagella_related proteins
23_ABC_transporter_related proteins
24_NECE_Type_VI_secretion_TSSS
25_oxidative_stress
26_Phage/Virus
27_NECEvent_antibiotic_resistance
6-fosphosphatase
alanyl tRNA synthetase
Alanyl-tRNA synthetase
amino_acid_Cysteine_t
amino_acid_Glutamate_aspartate_aminotransferase
amino_acid_Glutamine_Glutamine_synthetase
amino_acid_Glycine
amino_acid_Isoleucine/Valine
amino_acid_ketolacid_reductoisomerase
amino_acid_Leucine
amino_acid_phosphoglycerate_dehydrogenase
amino_acid_phosphoserine_aminotransferase
amino_acid_serine_hydroxymethyltransferase
amino_acid_serine_hydroxymethyltransferase
amino_acid_threonine_synthase
amino_acids_acetolactate_aldolase
AMP_salvage_Pathway_AMP_synthase
AMP salvage pathway
Arginyl tRNA synthetase
Arsenate_reductase_DMSO_phosphonylase
Asparaginyl-tRNA synthetase
Aspartyl_GW2011
Aspartyl-tRNA synthetase
Beta_oxidation_Fatty_acids_Acyl-CoA_Dehydrogenase
Beta_oxidation_Fatty_acids_Acyl-CoA_synthetase
Beta_oxidation_Fatty_acids_enoyl_hydratase
Carbon_fixation_reductive_TCA1_Carbon_fixation_calvin_cycle
Cell_cell_surface_protein
Cellulose Degradation_3.2.1.21_3.2.1.4_3.2.1.91
Complex_IV_cytochrome_bd_oxidase
deoxyribose phosphate aldolase_Putative_Fbpase
DMSO_Molybdopterin_oxidoreductase
DNA gyrase subunit A
DNA gyrase subunit B
EMP_fructose bisphosphate aldolase
EMP_glyceraldehyde-3-phosphate dehydrogenase
EMP_glucokinase
EMP_Phosphofructokinase
EMP_phosphoglucomutase
EMP_phosphoglycerate_kinase
EMP_pyruvate_kinase
EMP_triose phosphate isomerase
Fe-Hydrogenase
ggKBase_malate_dehydrogenase_TCA_pyruvate
Ggkbase_Pur_1_PurF_phosphoribosylamidotransferase
Gluconeogenesis
Gluconeogenesis_four_key
Gluconeogenesis_Fructose-1
Gluconeogenesis_Pep carboxykinase
Gluconeogenesis_PEP_synthase
Gluconeogenesis_Transketolase
Glutaminyl-tRNA synthetase
Glycogen_starch_synthase
Glycosyl_Transferase
Glycyl-tRNA synthetase
gyrA
Histidyl tRNA synthetase
Histidyl-tRNA synthetase
Hydrogenase_maturation
Hydrogenase_NiFe
Isoleucyl-tRNA synthetase
Leucyl tRNA synthetase
Lysyl-tRNA synthetase
Methionyl-tRNA synthetase
NECEvent_Toxin
Nitrogen_Assimilation_pathways
Nitrogen_cycle_Nitrate_reductase
Non-oxidative pentose phosphate pathway
Oxidative Pentose phosphate pathway
Peptidoglycan_biosynthesis
PfkB domain protein_Fructokinase-like
Phenylalanyl tRNA synthetase alpha subunit
Phenylalanyl-tRNA synthetase beta subunit
Phosphoenolpyruvate carboxylase
Pili_PilD_leader_peptidase
Pili_PilT_twitching_NTPase
PP_6-phosphogluconate_dehydrogenase
PP_glucose-6-P-1-dehydrogenase
PP_ribose_phosphate_pyrophosphokinase
PP_Ribokinase
PP_ribulose-5-P_isomerase
PP_3-epimerase
PP_transaldolase
Preprotein translocase
Prolyl
Pur_10_Pur
ggKbase List

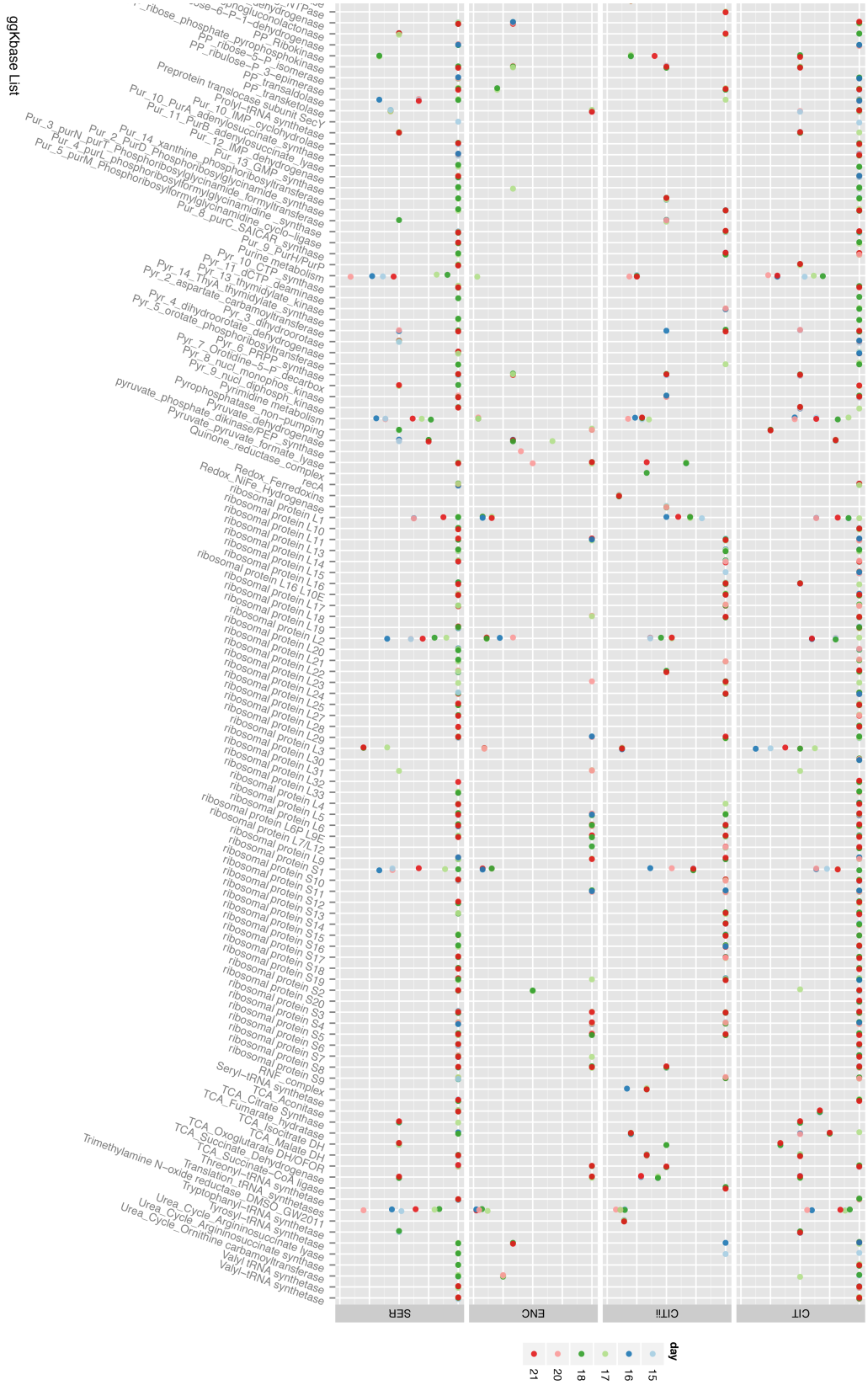

ggKbase List
SER
ENC
CITli
CIT
day
21
20
18
17
16
15

Figure 2-4: Expression over potential ratio of infant gut microbes (subset)

A non-redundant count of the number of features identified via proteomics in a metabolic ggKbase list was divided by the number of features in that list and plotted for each organism across time.

Figure 2-5: Comparison of proteomic profiles of two closely related *Citrobacter* strains

Unique peptide counts, normalized by the length of the protein, were mapped to *UC1CITs* contigs (aligned on the *x*-axis). Panels are separated by day of life. Triangles represent genes unique to the respective strain. Annotations marked with arrows indicate features of interest.

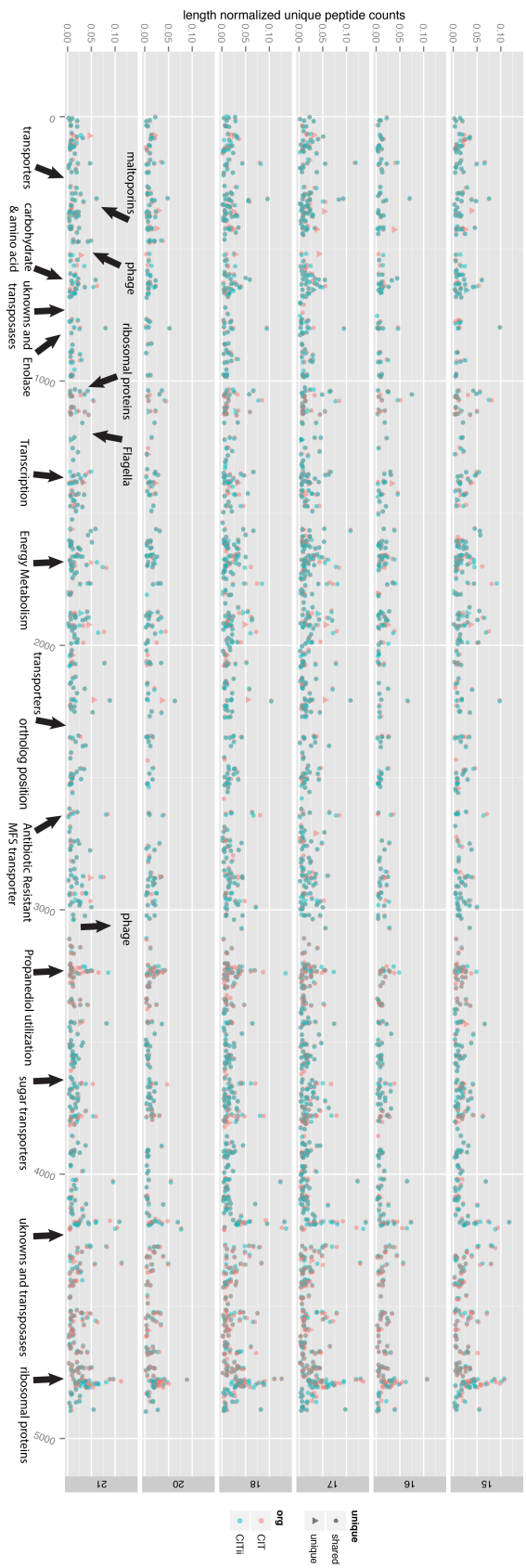

length normalized unique peptide counts
0.00
0.05
0.10
0
1000
2000
3000
4000
5000
transporters
carbohydrate & amino acid
uknowns and transposases
Enolase
maltoporins
phage
ribosomal proteins
Flagella
Transcription
Energy Metabolism
transporters
ortholog position
Antibiotic Resistant MFS transporter
phage
Propanediol utilization
sugar transporters
uknowns and transposases
ribosomal proteins
15
16
17
18
20
21
org
CiTti
CIT
unique
shared
unique

Table 2-1: Genome and proteomics summary

taxa	UC1CIT	UC1CITii	UC1CITp	UC1ENC	UC1ENCp	UC1ENCv	UC1SER
bp	4,902,348	4,901,982	59,966	2,576,397	77,038	37,230	5,027,440
contigs	10	10	2	785	2	2	9
max contig	2,550,874	2,550,962	57,067	18,409	68,691	28,900	2,360,977
genes	4,829	4,696	61	3,589	94	56	4,569
%GC	52	52	53	37	34	32	60
SCG (51 total)	50	48	0	40	0	0	45
unique proteins detected	1,049	1,017	5	195	4	1	1,021
avg unique protein per day	619	603	2	84	2	1	520
unique protein matches	20,038	19,207	26	1,636	25	17	16,129
avg unique protein matches per day	3,331	3,192	5	270	7	5	2,677

Chapter 3:

3 The developing premature infant gut microbiome is a major factor shaping the microbiome of neonatal intensive care unit rooms

Brandon Brooks[1], Brian A. Firek[3], Robyn Baker[4], Brian C. Thomas[2], Michael J. Morowitz[3], Jillian F. Banfield[2]

1 – Department of Plant and Microbial Biology, University of California, Berkeley, CA
2 – Department of Earth and Planetary Sciences, University of California, Berkeley, CA
3 – University of Pittsburgh School of Medicine, Pittsburgh, PA
4 – Division of Newborn Medicine, Children's Hospital of Pittsburgh of UPMC, Pittsburgh, PA

3.1 Abstract

The neonatal intensive care unit (NICU) contains a unique cohort of patients with underdeveloped immune systems and nascent microbiome communities. Patients often spend several months in the same room and it has been previously shown that the gut microbiomes of these infants often resemble the microbes found in the NICU. Little is known, however, about the identity, persistence and absolute abundance of NICU room-associated bacteria over long stretches of time. Here we couple droplet digital PCR (ddPCR), 16S rRNA gene surveys, and recently published metagenomics data from infant gut samples to infer how infants acquire microbes from their immediate NICU environment. Over 3,700 swabs, wipes, and air samples were collected from sixteen private-style NICU rooms housing very low birthweight (<1,500 g), premature (<31 weeks' gestation) infants. For each infant, room samples were collected daily, Monday through Friday, for one month. The first samples from the first infant and last samples from the last infant were collected 383 days apart. Twenty-two NICU locations spanning room surfaces, hands, electronics, sink basins, and air were collected. Results show a room community largely dominated by 5-10 taxa, mostly skin associated. Biomass estimates reveal 5-6 orders of magnitude difference between the least to the most dense microbial communities, air and sink basins, respectively. Biomass trends from bioaerosol samples and petri dish dust collectors suggest occupancy to be a main driver of suspended biological particles within the NICU. Using a machine learning algorithm to classify the origin of room samples, we show that each room has a unique microbial fingerprint. Several important taxa driving this model were dominant gut colonizers of infants housed within each room. Collectively, the data suggests that housed infants, in combination with their caregivers, shape the microbiome of NICU rooms.

3.2 Introduction

Hospital acquired infections (HAIs) remain a major problem in the US. One out of every twenty-five patients will experience a HAI, costing the US approximately $30 billion per year (CDC, 2016). Infants hospitalized in the neonatal intensive care units (NICU) are particularly susceptible to infection due to their underdeveloped immune systems (Arrieta *et al.*, 2015; Cahenzli *et al.*, 2013). To protect against infection, infants are often prescribed antibiotics during the first week of life. In fact, antibiotics are 3 of the 6 most commonly administered medications in the NICU (Gasparrini *et al.*, 2016). This treatment largely kills microbes acquired during the birthing process (Raveh-Sadka *et al.*, 2015) and promotes a categorically different colonization pattern in preterm infants relative to full term infants (Groer *et al.*, 2014). Preterm infants are often colonized by ESKAPE organisms (*Enterococcus spp.*, *Staphylococcus aureus*, *Klebsiella spp.*, *Acinetobacter spp.*, *Pseudomonas aeruginosa*, Enterobacteriaceae), which are also the most frequent cause of nosocomial infections (Hu *et al.*, 2015). The relatively sterile preterm infant gut microbiome and the high frequency at which infants are colonized by hospital associated microbes, creates a valuable study setting to better understand how room microbes are exchanged with occupants. Here, we conducted an experiment to quantify and characterize NICU room microbes and their role in infant gut colonization.

The source of early stage gut colonizers in preterm infants is poorly understood. In a recent pilot study, we tracked two infants over the first month of life, collecting samples from room surfaces and infant fecal samples (Brooks *et al.*, 2014). Using an amplicon-EMIRGE approach, which allows for recovery of the full-length 16S rRNA gene (~1500 b) (Miller *et al.*,

2011), as opposed to the more common hypervariable region approach (~150-400 b), we detected the same sequences in room samples before they were detected in gut samples. With genomes recovered from the infant metagenomic data, we found that the same strain of *Enterococcus faecalis* colonized both infants. The results from that study suggested that infants housed in the same NICU at approximately the same time acquire initial gut colonizers from the room environment.

Subsequent studies from our group and others offer a more nuanced perspective. While it is possible to have the same strain colonize several cohoused infants, the vast majority of strains are not shared (Raveh-Sadka *et al.*, 2015). Nearly 150 strains were recovered from 10 infants' fecal samples and only 4 of these were shared. These samples were collected within a month of each other, suggesting that a multitude of strains is available in the NICU at any given point in time, and only a few strains may be widespread. Infant fecal samples collected years apart, in this same NICU, show some strains that are consistently shared across time (Raveh-Sadka *et al.*, 2016). These "persister" strains were not found to differ significantly in virulence, antibiotic resistance, or metabolism from other non-persister strains. A recent study, however, using a different functional metagenomics approach, found 794 antibiotic resistance genes in preterm infant stool samples, 79% which had not previously been classified as associated with resistance (Gibson *et al.*, 2016). It is possible these antibiotic resistance genes not only offer a competitive advantage in the gut, but also allow tolerance to surface cleaning outside the gut (Buffet-Bataillon *et al.*, 2011).

The results from the prior studies suggest that a subset of bacteria that colonize infants are sourced from their room environment, but specific reservoirs remain elusive. To address this knowledge gap, we designed a study with more infants (n = 16), finer temporal sampling (Monday through Friday), and more sites (n = 22) relative to our pilot study. Additionally we performed droplet digital PCR (ddPCR) on all room samples to better understand how biomass varies in the NICU and used quantification of negative controls to account for background taxa. Overall, our experimental design broadly aimed to better understand room-occupant interactions, with the results furthering our understanding of infant gut colonization and the dynamics that govern how patients acquire microbes during their hospitalization.

3.3 Methods

3.3.1 Sample Collection

Infants were enrolled in the study based on the criteria that they were < 31 weeks gestation, < 1250 g at birth, and were housed in the same physical location within the NICU during the first month of life. Samples were collected Monday through Friday for days of life (DOL) 5-28. Fecal samples were collected using a previously established perineal stimulation procedure and were stored at -80 °C within 10 minutes (Morowitz *et al.*, 2010a). When fresh samples were not available, diaper samples were used (noted within the metadata). All samples were collected after signed guardian consent was obtained, as outlined in our protocol to the ethical research board of the University of Pittsburgh (IRB PRO11060238). This consent included sample collection permissions and consent to publish study findings.

All samples were obtained from a private-style NICU at Magee-Womens Hospital of the University of Pittsburgh Medical Center. Twenty-two of the most frequently touched surfaces were determined by visual observation and health care provider interviews in the weeks leading

up to sample collection. Microbial cells were removed from most surfaces using nylon FLOQSwabs (Copan Diagnostics, Brescia, Italy) and a sampling buffer of 0.15 M NaCl and 0.1% Tween20. Samples were collected by one research nurse to ensure consistent sampling technique. Ten square centimeters of each surface was sampled or, for smaller surfaces, the entire surface itself (e.g., isolette knobs and sink basin drain grill). Wipe samples were collected from the floor and exterior top of the isolette using Texwipe TX1086 wipes (Texwipe, Kernersville, NC, USA). Before collecting each wipe sample, the collector would put on latex examination gloves and clean these gloves with an isopropanol wipe. The wiped surface area was approximately forty-eight square centimeters or, for smaller surfaces, the entire surface itself (e.g., isolette top). A wipe was also used to collect microbial cells at the exterior facet of the heating, ventilation and air conditioning (HVAC) system. The wipe was suspended via airflow on the exterior (upstream) face of the MERVE 8 pleated filter, the zone in which supply and return air are mixed before thermal and humidity treatment of the airstream for four days. Features of the HVAC system are described in detail in a recently published paper (Licina *et al.*, 2016).

Air samples were collected using the NIOSH two-stage bioaerosol cyclone 251 sampler (Lindsley *et al.*, 2010) and a suspended petri dish method (Adams *et al.*, 2013). The NIOSH sampler collected samples continuously Monday through Friday, comprising approximately 96 hours of sampling at 3.5 L/minute (total volume sampled = 20 m^3). Petri dish samples were suspended approximately one meter below the drop ceiling in the corner of the room that was the furthest away from the sink. These samplers were maintained in place for the duration of the infant's stay. Petri dish "cooler" samples are plates that were taped to the top of a cooler which collected abiotic aerosol data (Licina *et al.*, 2016). At the end of the sample collection period, all samples were placed in a sterile transport tube and stored within 10 minutes at -80 °C until further processing.

3.3.2 DNA extraction and PCR amplification

Frozen fecal samples were thawed on ice and 0.25 g of thawed sample added to tubes with pre-warmed (65 °C) lysis solution from the PowerSoil-htp 96 Well DNA Isolation Kit (MoBio Laboratories, Carlsbad, CA, USA). The incubation was conducted for five minutes and the manufacturer's protocol followed thereafter. Swab heads followed the same procedure, except heads were snapped at the perforation into the extraction tube before starting the protocol. Wipe samples were stored in a sterile 250 mL tissue culture flask upon collection and thawed on ice before extraction. Cells were dislodged from wipes in a protocol adapted from Yamamoto *et al*. (Yamamoto *et al.*, 2011). Briefly, 150 mL of dislodging buffer was poured into a flask (1X PBS, 0.04% Tween 80, passed through a 0.2 µm filter), the flask was shaken vigorously for one minute, and then shaken at medium speed on a flask shaker for approximately one hour at room temperature. Supernatant was then decanted into a 250 mL disposable filter funnel with a pore size of 0.2 µm (Thermo Scientific, Waltham, MA, USA) and the filter was then placed in a MoBio PowerWater extraction tube. PowerWater extraction followed manufacturer recommendations.

Genomic DNA from room samples were subjected to 16S rRNA V3-4 MiSeq library preparation which included dual-barcoded multiplexing with a heterogeneity spacer for higher sequence quality (Fadrosh *et al.*, 2014). Two microliters of 5X concentrated gDNA template was used in the reaction and run at 35 cycles. Amplicons were purified using the Just-a-Plate PCR

normalization and purification kit (Charm Biotech, San Diego, CA, USA). Equal amounts of each sample were sent to the University of California Davis DNA Technologies Core Facility (http://dnatech.genomecenter.ucdavis.edu) and run on a MiSeq with v3 300PE chemistry.

Droplet digital PCR (ddPCR) was adapted from a method previously published on quantification of 16S rRNA templates in infant fecal samples (Raveh-Sadka *et al.*, 2015). The only deviation from the previous method was that a diluted gDNA template of 1:10 instead of 1:1000 was utilized. Both MiSeq library preparation and ddPCR were performed in 96-well plate format. Each plate had three no template PCR controls, one no template extraction control, and three positive controls containing varying concentrations of purified *E. coli* gDNA. Counts from the negative control types were averaged across type and the highest was used to correct for contaminant counts in sample data.

3.3.3 Sequencing preparation and sequencing

Illumina library construction followed standard protocols at University of California QB3 Vincent J. Coates Genomics Sequencing Core Facility (http://qb3.berkeley.edu/gsl/). Briefly, gDNA was sheared using a Covaris to approximately 600 bp and 1000 bp. Wafergen's PrepX DNA library prep kits were used in conjunction with the Apollo324 robot following factory recommendations (Integenx). Thirteen cycles of PCR were used during library construction. Libraries were added at 12 samples per lane, in equimolar amounts, to the Illumina HiSeq 2500 platform. Paired-end sequences were obtained with 150 cycles and the data processed with Casava version 1.8.2. Raw read data will be deposited in the NCBI Short Read Archive (accession numbers pending).

3.3.4 16S amplicon data processing

The LotuS 1.53 pipeline in short amplicon mode was used for quality filtering, demultiplexing, and OTU picking (Hildebrand *et al.*, 2014). LotuS was run with the following command line options: '-refDB SLV,GG -highmem 1 -p miseq -keepUnclassified 1 -simBasedTaxo lambda -threads 10.' The OTU data was rarefied to 1,000 sequences per sample, without replacement, unless explicitly stated. The final OTU table and mapping file is publicly available at the QIITA database (https://qiita.ucsd.edu/, pending).

3.3.5 Metagenomic assembly and data processing

Metagenomic sequencing of 290 fecal samples on twenty-five lanes of an Illumina HiSeq 2500 produced ~800 Gb of 150 bp paired-end reads. Reads were trimmed with Sickle (Joshi, 2011), mapped to the human genome using Bowtie2 (Langmead and Salzberg, 2012) to remove human contamination, and assembled with idba_ud (Peng *et al.*, 2012) using default parameters for all programs. Prodigal (Hyatt *et al.*, 2010) was used for gene prediction of scaffolds longer than 1 kb. Genes were annotated using USEARCH (Edgar, 2010) to search against KEGG (Kanehisa *et al.*, 2014), UniReff100 (Suzek *et al.*, 2007), and UniProt databases. Matches with bit scores greater than 60 were saved as were reciprocal best hits with scores greater than 300. rRNA sequences were identified using Infernal (Nawrocki and Eddy, 2013), and tRNAs with tRNAscan_SE (Lowe and Eddy, 1996).

3.4 Results

3.4.1 Sequencing summary and contamination removal

In all, 2584 samples were processed through a MiSeq library protocol. After quality filtering and demultiplexing, 96,876,367 reads were generated. These reads were clustered into 17,932 OTUs. Using a ratio OTU (ROTU) method that leverages biomass quantification and sequencing of negative controls (Lazarevic *et al.*, 2016), 270 OTUs and 924 samples were removed from the dataset when using an ROTU threshold of 0.001. A second *in silico* contamination cleaning method was applied (Meadow *et al.*, 2015), which removed an additional 324 OTUs and 1 sample. This indicates that approximately 4% of generated OTUs and 36% of samples present too weak of a signal to confidently distinguish them from negative control signatures.

3.4.2 Biomass and taxonomic variation across petri dish replicates

Biological and technical replicates performed for petri dish plates established the reproducibility of extraction of DNA from petri dish swabs and provided evidence for highly reproducible ddPCR measurements (Figure 3-1). The highest standard deviation in ddPCR values for biological replicates in a single room was 106,760 copies/sample (infant 6's petri plates; mean = 99,677) and for technical replicates, the largest standard deviation was 15,534 copies/sample (infant 12's petri plates, mean = 81,044). The lowest standard deviation for biological replicates was 1,981 copies/sample (infant 1's petri plates, mean = 13,785) and 737 copies/sample for technical replicates (infant 11's petri plates, mean = 32,396). Overall, this equates to a reproducibility range of 2.69 to 6.87× more reproducibility across technical ddPCR runs relative to biological replicates, with an average reproducibility ratio of 5.37× better for technical replicates.

3.4.3 Biomass varies significantly across sample type

16S rRNA gene copies were quantified for 2,584 samples (Figure 2). Samples from the HVAC system had the highest biomass of all types and bioaerosol samples had the lowest (Figure 2 a and b). Sinks had the highest biomass of the swabbed samples and hands had the lowest average median template count. Petri dishes suspended from the ceiling had the lowest biomass relative to other passive dust collectors, whereas the nurse's station dishes contained the highest bacterial load. The infant room consistently had higher template counts than the hallway bioaerosol samples. Overall, the biomass varied approximately 5-6 orders of magnitude across all sample types.

3.4.4 Skin associated taxa dominate the NICU surface environment

The microbial communities in most NICU environments were highly uneven and were dominated by 5-10 taxa (Figure 3-3). Approximately 50% of all amplicon reads belong to five of the top taxa in the NICU (Figure 3-3 and Table 3-1). Most of these taxa are human associated with many originating from the skin (*Propionibacterium*), mouth (*Streptococcus*), or nose (*Staphylococcus*). SourceTracker (Knights *et al.*, 2011) was run using skin, oral, and fecal

samples from the American Gut project as the putative source database with NICU samples labeled as "sink" samples. Skin was the most likely contributor to taxa in the NICU followed by oral and fecal samples (Figure 3-4).

Samples collected from the HVAC system had the highest bacterial diversity with 405 OTUs on average per sample, whereas bioaerosol samples had the lowest, with 13 (Figure 3-5a). The HVAC samples had the highest Shannon community evenness, followed by floor wipes, and the bioaerosol samples had the lowest Shannon diversity (Figure 3-5b). Thus, overall, the HVAC had highly even consortia with great diversity. As expected due to heavy filtration and air exchange, the NICU room air has low biomass and low diversity, and with strong dominance by members of the Aeromonadaceae.

All touched surfaces had similar numbers of OTUs per sample, although the surface monitors showed the most unevenness. These surfaces were dominated by similar groups of microbes. Although many touched surfaces were associated with skin-associated bacteria, gut associated Enterobacteriaceae OTUs also dominated environments such as the surface monitors, counter tops, and scanners. In contrast, the sink basins had comparatively low numbers of OTUs per sample (Figure 3-5a), in part due to the high dominance by four bacterial groups (Figure 3-3).

3.4.5 Biomass suggests growth patterns in sink basins

A range of 29 to 38 sink basin samples per day of the week was collected from 14 unique sink basins. When comparing biomass trends across weekdays (Figure 3-6a), a distinct pattern of decreasing biomass is apparent in sink samples relative to other swabbed environments. In comparing Shannon diversity across weekdays (Figure 3-6b), bacterial diversity in Tuesday versus Friday samples were the most distinct whereas biomass was most different in Monday versus Thursday samples (Wilcoxon rank sum, Bonferroni adjusted p = 0.4 and 0.012, respectively). Sink basins were cleaned approximately every twenty-four hours, but less frequently on the weekends so the elevated biomass at the beginning of the week may be due to regrowth of sink adapted taxa throughout the weekend (e.g., Pseudomonas, Aeromonas, and Enterobacteriaceae). The increase in Shannon diversity from Monday to Friday strengthens this inference.

3.4.6 NICU rooms harbor a unique microbial signature

Using a support vector machine (SVM) classifier with a linear kernel (Chase *et al.*, 2016), we determined that each room's microbiome contained a unique microbial fingerprint. We could predict the room origins with an overall accuracy of 51% (when we knew the room's origin but withheld that information from the classifier), which is 4.6x better than random chance (Figure 3-7). The use of ROTU over a standard QIIME pipeline achieved an increase in accuracy of approximately 12%. Typically, the most confusion occurred between samples that were collected at similar times (i.e. infant 2 and 3's samples). Important OTUs driving the SVM model are plotted and listed in Figure 3-8 and Table 3-2. Interestingly, there is an overlap between room specific OTUs that drive the SVM model and occurrence of these taxa in the gut of infant occupants. For example, the most visible signature in SVM taxa comes from a spike in *Veillonella* in infant 6's room on DOL 18 (Figure 3-8). The first major increase of *Veillonella* in infant 6's gut occurs on DOL 16 (Figure 3-9). The same pattern is seen for infant 8, and in fact,

most infants that contain *Veillonella* have strong SVM signals associated with their room. The second strongest signal from the SVM model comes from a *Clostridum* OTU. This group is present in infants 2, 3, and 8's room samples and it strongly contributes to the SVM model prediction. All three of these infants have high abundances of *Clostridium*.

3.4.7 Composition of persister taxa in the room echoes infant gut composition

To visualize the distribution of taxa that are known to persist in infants over multi-year periods (Gibson *et al.*, 2016; Raveh-Sadka *et al.*, 2016), we collapsed each study day and infant pairing by averaging all amplicon abundance data across environments (Figure 3-10, "average" panel). In this analysis, the subset of all OTUs that was classified as persister taxa was assigned a distinct color. Each color represents one family, but the same color was used multiple times if more than one OTU could be distinguished within a family. Due to high abundance, we gave OTU_5 (an *Enterobacteriaceae*) dedicated coloring. Surprisingly, persister taxa often account for > 50% of the data at many time points.

Episodes of particularly high persister abundance occurred in rooms housing infants 1, 9, 12, and 16. To better visualize which samples contributed to the averaged data (Figure 3-10, "average" panel), we also plotted data for the specific environments for which we had the most samples (armrests and sinks). Both the armrests and sinks are dominated by persister taxa during these episodes, but *Staphylococcaceae* OTUs are much more abundant in armrest samples relative to sinks. Two dominant Pseudomonas OTUs that comprised 70% and 24% of all Pseudomonadaceae (OTU_8 and OTU_15, respectively) were detected throughout the time series, but were at very low abundance in armrest samples over long time spans.

Since the room data for infant 9 had a strong persister signal, we analyzed samples from all environments separately to visualize temporal patterns (Figure 3-11a). Persister taxa dominated most of infant 9's room samples, with cellphones having the fewest and scanner and surface counter samples having the most persister groups per sample. The red lines in Figure 3-11a highlight the time point where a major increase in relative abundance of *Enterobacteriaceae* taxa occurred in infant 9's gut (Figure 3-11b). This group is present in multiple room environments prior to the increase, particularly associated with the isolette and armrest. At subsequent time points, this group becomes highly prominent in some room environments (e.g., scanner and surface counter).

OTUs belonging to the persister groups can only be resolved to the genus level and in the case of OTU_5, the family level. Since *Enterobacteriaceae* dominates the gut of infant 9, we leveraged room and fecal sample context to infer a possible identity for OTU_5. Using OTU_5's reference sequence as a query, we ran ublast (Edgar, 2010) on a database of 16S rRNA genes reassembled from infant 9's fecal metagenomic samples using the EMIRGE-like REAGO algorithm (Miller *et al.*, 2011; Yuan *et al.*, 2015). The top hit to our 429 bp query was 99.5% identical (2 mismatches) and came from several of infant 9's fecal samples. Most of the top hits have the entire 16S rRNA gene recovered from the REAGO assembly (~1,520 bp). These fecal sequences were searched against the Silva database (SLV_119_SSU) and returned identical, full-length matches to *Klebsiella pneumoniae*. While this is an extrapolation from the V3-4 region, it is possible that OTU_5 in the room is a *Klebsiella* and may be *Klebsiella pneumoniae*, the dominant bacterium colonizing infant 9.

3.5 Discussion

The first question that we aimed to answer in this study related to how biomass varies across a NICU. Using ddPCR to quantify 16S rRNA gene copy number, we show biomass density varies across NICU surfaces by 5-6 orders of magnitude (Figure 3-2). Surprisingly, the floor in front of the infant's isolette had the highest density of microbes relative to any other environment within the NICU. Naively, it may seem intuitive that the region with the most foot traffic, e.g. the floor at the main entrance of the NICU, would have the highest biomass. While the main entrance floor has a high density, it is significantly lower than the floor in front of the isolette. This finding may be due to the increased occupancy at the isolette versus the main entrance, where occupancy is more transient.

Petri dish data also suggest that higher levels of human activity drive higher amounts of microbial deposition in the room environment. Notice that the nursing station has higher petri dish-associated biomass than the infant room, followed by the hallway (Figure 3-2). This outcome occurred despite the fact that the infant room and hallway coolers collected dust at the same height (1 m), whereas the nurse station collector was at approximately double the height (1.8 m). As height above the floor increases, detection of resuspended particles from dust decreases exponentially (Luoma and Batterman, 2001; Fairchild and Tillery, 1982). This finding suggests that floor dust is not the main source of biological particles accumulated in the petri dishes, but rather the microbes are human-derived. Greater occupancy or rigor of activity (Bhangar *et al.*, 2016) at the nursing station compared to the infant room and hallway likely explains this result.

A recently published paper noted a stronger occupancy signal from the occupancy sensors in the infant room compared to the hallway (Licina *et al.*, 2016). The occupancy signal directly overlapped with the coarse particle signal (which detected particles > 10 μm in diameter). This signal was interpreted to indicate that resuspension or deposition of particles from occupants is be the largest contributor of aerosolized particles in the NICU. In the current study, our Petri dish ceiling analyses suggest a similar conclusion for settled particles, but in this case based on biological data.

If occupancy is a key feature of the NICU environment, one would expect human associated microbes to dominate in most room environments. We found that 5-10 taxa account for most of the amplicon data and a majority of these are typically skin, nose, or fecal associated (Figure 3-3). The enrichment of human associated taxa is likely due to tight control of the building envelope via HVAC treatment (Kembel *et al.*, 2012) combined with a strict cleaning schedule.

An interesting finding of this study related to the change in biomass and microbial community structure of the sink basins over the course of the week. We attribute this pattern to the room cleaning regime, which is more limited on weekend days than during the week. On Mondays, the sink biomass is highest (Figure 3-6a) and communities are relatively uneven (Figure 3-6b), presumably due to extensive growth of a few sink-associated taxa over the weekend. More intensive cleaning of the sink early in the week likely removes the majority of biomass, which is presumed to be comprised of the sink-adapted taxa. These events enable detection of small numbers of many different types of bacteria that are less well-adapted to the sink environment later in the week.

The second question addressed in our study related to the taxa that dominate NICU surfaces. To investigate this, it was necessary to adapt a method to eliminate spurious

contaminant-based signals in data from low biomass samples (Lazarevic *et al.*, 2016). The ROTU cleaning method implemented here to clean data of spurious OTUs and contaminants *in silico* was made possible due to the availability of ddPCR quantification of negative controls. This capability is particularly important for NICU studies since the rooms are cleaned regularly, causing low biomass levels to be present in many samples. Some of the bacteria that we conclude were introduced in sample processing are skin associated, although many other groups were encountered. After accounting for contamination, we conclude that human associated taxa dominate most surfaces.

Human associated taxa are likely sourced and trafficked throughout the NICU by healthcare providers (Kembel *et al.*, 2014) and many hand hygiene studies have reported as much (Luangasanatip *et al.*, 2015). Here, we implemented a machine learning classifier to address the possibility that infants and their caretakers shape the microbiome to be distinctive in each room. Our model reliably classified samples of unknown origin to their correct room-infant pair at an accuracy two times better than a recently published office microbiome study (Chase *et al.*, 2016) and achieved predictive power nearly five times better than random chance. This outcome suggests that NICU rooms are more personalized than offices. There are typically a larger variety of activities and people in office spaces and air treatment is lower (lower air exchange rates and less filtration). The combination of less frequent cleaning, increased occupancy, and more unfiltered outdoor air supply drives many of the differences between other common indoor environments and the NICU. The more unique room signal based on NICU room microbes suggests a localized source of bacteria, since a more diffuse source would cause lower prediction accuracy.

Finally, we tested for patterns of association between room-occupants and NICU room environments. We found that many taxa driving our machine learning model for the room microbiome were from groups also present in the gut of the infant occupant. Other signals came from Firmicutes and Actinobacteria not affiliated with the infant gut and that were relatively uniquely detected in certain rooms. Focusing on the subset of taxa that are gut colonizers, we show a relatively high abundance of these taxa throughout the sampling campaign (Figure 3-10). Episodes where persistent taxa increase and 2-3 OTUs comprise > 30% of the data across all environments occurred several times throughout the study (*e.g.*, in infants 9, 12, and 16). These OTUs are detected in low abundance in the room before detection in the gut (

Figure 3-11). Once in the infant gut, a far more favorable environment for growth and reproduction than on exposed hospital surfaces, bacterial density can reach nearly 10 billion cells per gram (Raveh-Sadka *et al.*, 2015). After a bloom in the gut, we see these organisms expand into the room environment, mirroring gut colonization. It is impossible to resolve room 16S rRNA amplicon data to the strain-level in order to make claims that the same gut bloom resulted in a subsequent expanded appearance in the room. Contextually, however, we linked infant 9's dominant gut *Klebsiella pneumonia* to the *Enterobacteriaceae* increase in the room. Interestingly, the same strain of *K. pneumoniae* has been detected years apart in different infants within this NICU (Raveh-Sadka *et al.*, 2016). To validate if indeed the increase of infant colonizing microbes in the room are the same strains as in the gut, whole genome recovery is needed.

Based on the current study, we conclude that two factors shape room microbiomes. First, our taxa identifications and occupancy results strengthen prior findings of a strong link between human activity levels and room microbiology (Licina *et al.*, 2016; Bhangar *et al.*, 2016; Brooks *et al.*, 2014). In fact, this connection appears to be strong enough to give rise to a relatively unique room microbiome character. Second, environmental stresses, likely associated with cleaning (Brooks *et al.*, 2014; Buffet-Bataillon *et al.*, 2011; Romanova *et al.*, 2007; Weiss-Muszkat *et al.*, 2010; Hoffman *et al.*, 2005), may be selectively shaping NICU microbiomes, primarily by selecting for microbial specialists (microbes that persist because they can both thrive in the gut and tolerate the NICU environment). While daily cleaning substantially lowers the bioburden in the NICU (Bokulich *et al.*, 2013), the harshest cleaning methods cannot sterilize hospital surfaces (Hu *et al.*, 2015). Creative new approaches to displace or prevent entrenchment of these NICU specialists, possibly through prebiotic building materials or clever probiotics, may present opportunities to break the room-occupant cycle.

3.6 Competing Interests

The authors declare that they have no competing interests.

3.7 Authors' Contributions

JFB, MJM, and BB conceived of the project. RB organized cohort recruitment and sample collections. BAF conducted nucleic acid extractions and BB conducted the sample pooling. BB conducted the metagenomic assemblies and BCT provided annotations. BB and JFB wrote the final manuscript. All authors have read and approve the manuscript.

3.8 Acknowledgments

Funding was provided through the Alfred P. Sloan Foundation and the National Science Foundation's Graduate Research Fellowship Program. This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 OD018174 Instrumentation Grant.

Figure 3-1: Variation across ddPCR replicates

16S rRNA template copy number was quantified via ddPCR for three petri dish dust collectors suspended from the drop ceiling in each infant's room. Each dot reflects the average across triplicates runs. Each infant set is labeled at the top of the plot facets.

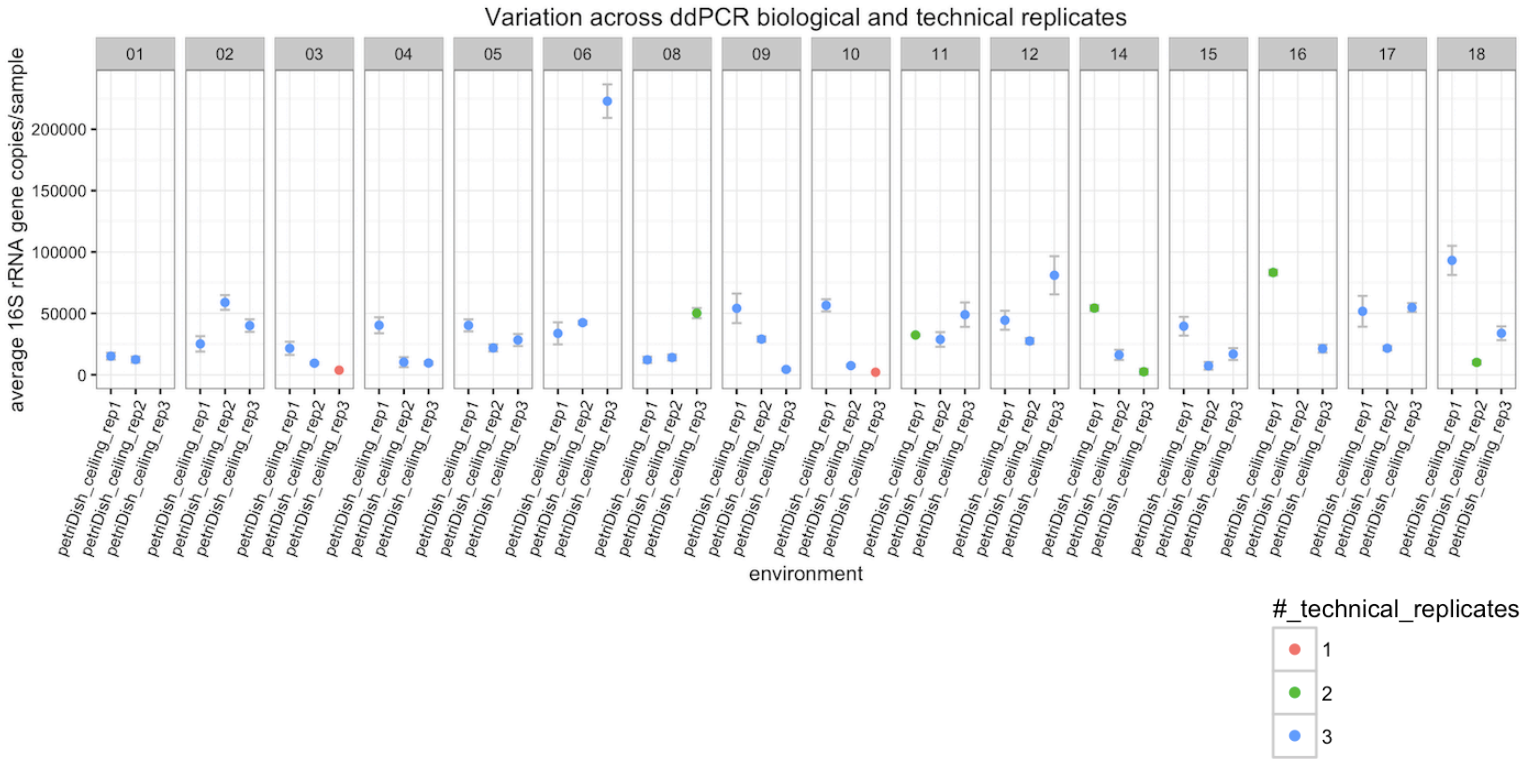
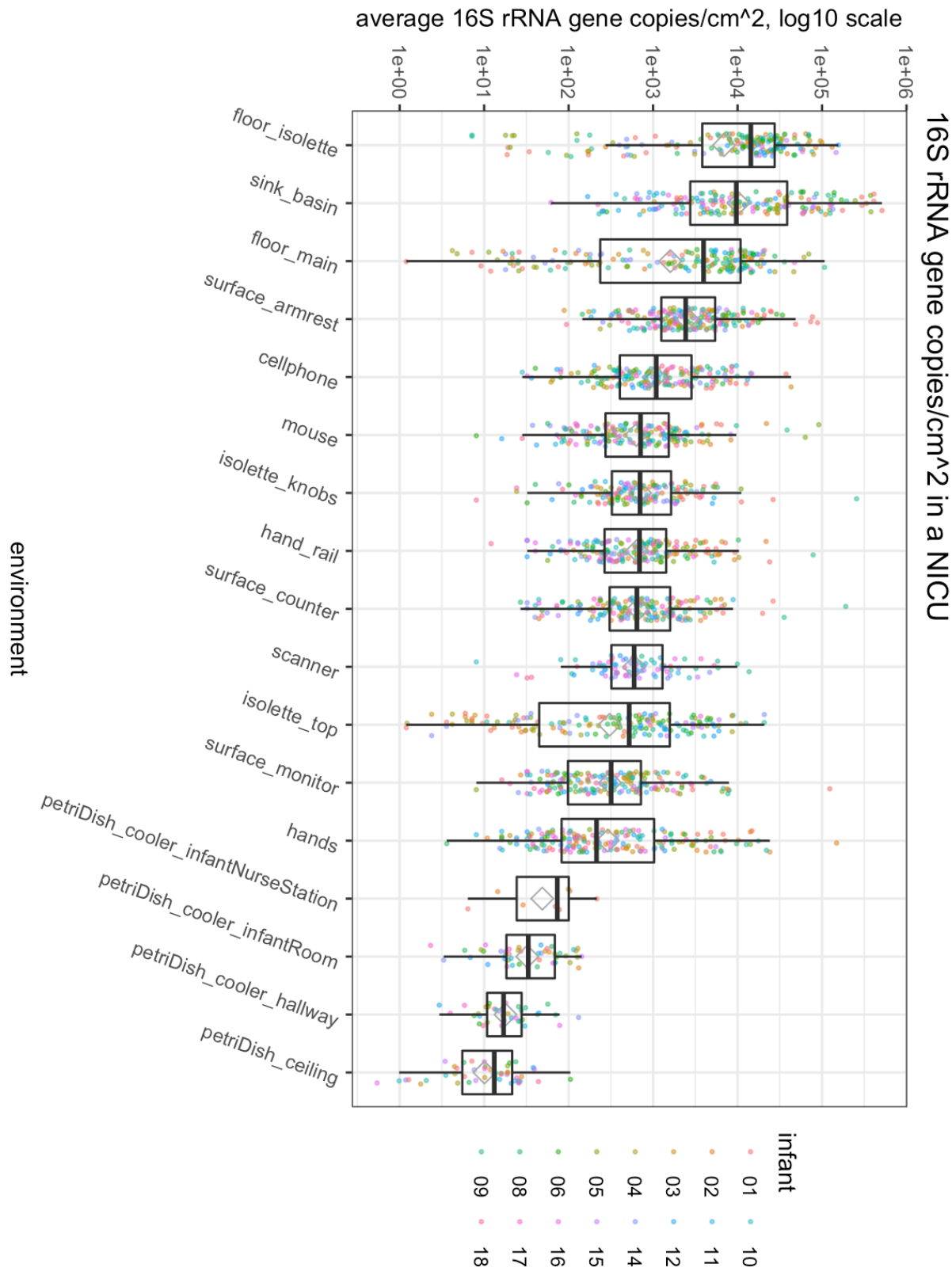
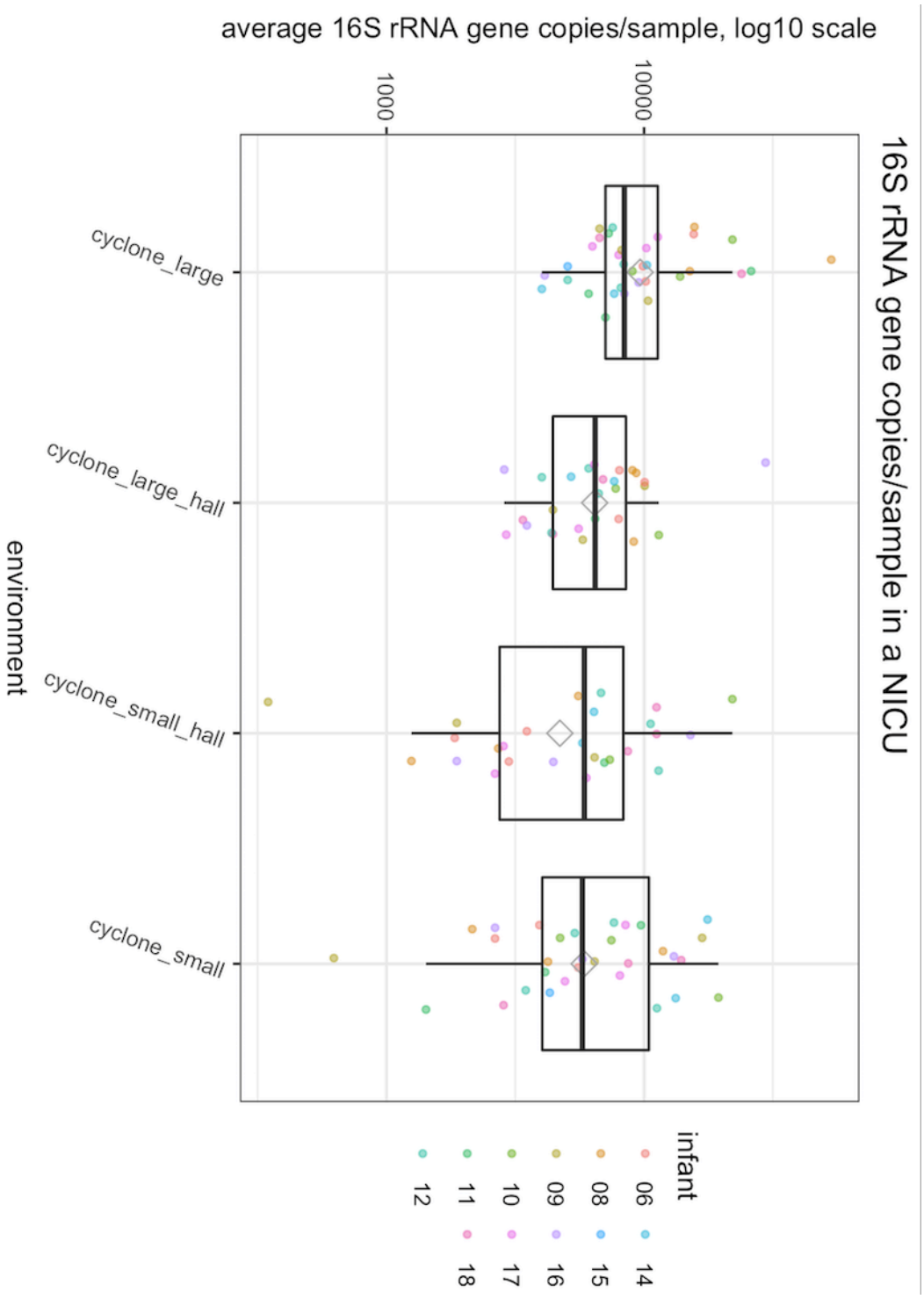

Variation across ddPCR biological and technical replicates
01
02
03
04
05
06
08
09
10
11
12
14
15
16
17
18
average 16S rRNA gene copies/sample
0
50000
100000
150000
200000
petriDish_ceiling_rep1
petriDish_ceiling_rep2
petriDish_ceiling_rep3
environment
#_technical_replicates
1
2
3

Figure 3-2: Biomass varies by 5-6 orders of magnitude in a NICU

16S rRNA template copy number was quantified via ddPCR. Each dot reflects the average across triplicates runs. Grey diamonds represent averages per environment. Figure 3-2a depicts surface samples in copies/cm^2 and Figure 3-2b and Figure 3-2c shows bioaerosol and HVAC samples in copies/sample.


16S rRNA gene copies/cm^2 in a NICU
average 16S rRNA gene copies/cm^2, log10 scale
1e+00
1e+01
1e+02
1e+03
1e+04
1e+05
1e+06
environment
floor_isolette
sink_basin
floor_main
surface_armrest
cellphone
mouse
isolette_knobs
hand_rail
surface_counter
scanner
isolette_top
surface_monitor
hands
petriDish_cooler_infantNurseStation
petriDish_cooler_infantRoom
petriDish_cooler_hallway
petriDish_ceiling
infant
01
02
03
04
05
06
08
09
10
11
12
14
15
16
17
18


16S rRNA gene copies/sample in a NICU
average 16S rRNA gene copies/sample, log10 scale
1000
10000
environment
cyclone_large
cyclone_large_hall
cyclone_small_hall
cyclone_small
infant
06
08
09
10
11
12
14
15
16
17
18

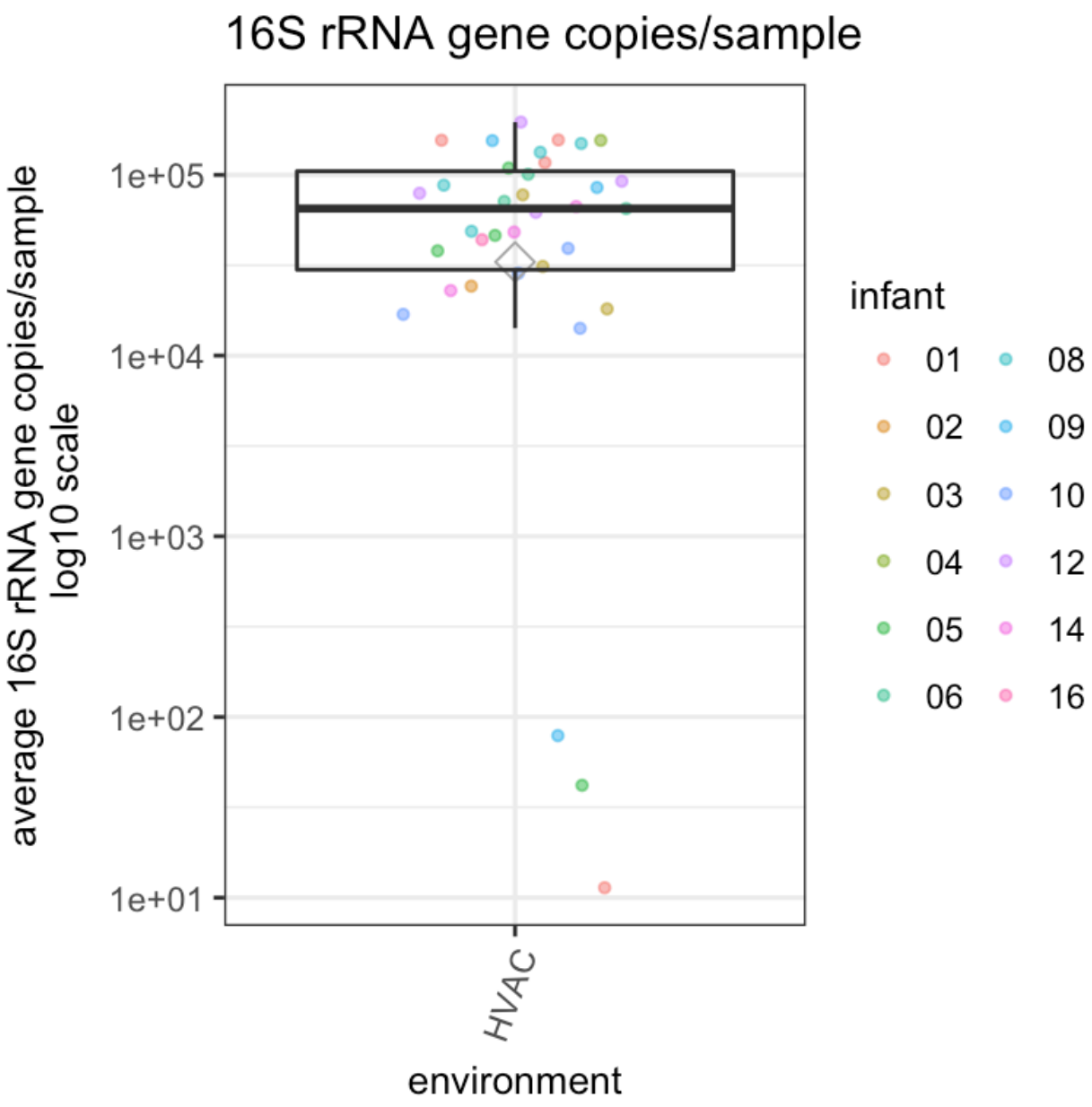

16S rRNA gene copies/sample
average 16S rRNA gene copies/sample
log10 scale
1e+05
1e+04
1e+03
1e+02
1e+01
HVAC
environment
infant
01
02
03
04
05
06
08
09
10
12
14
16

Figure 3-3: Top 10 NICU OTUs comprise > 50% of NICU taxa

Amplicon data from a 16S rRNA V3-4 workflow is plotted for each environment. Only the top 10 OTUs, determined from averages across all samples, are plotted. Each OTU is colored by its family-level classification.

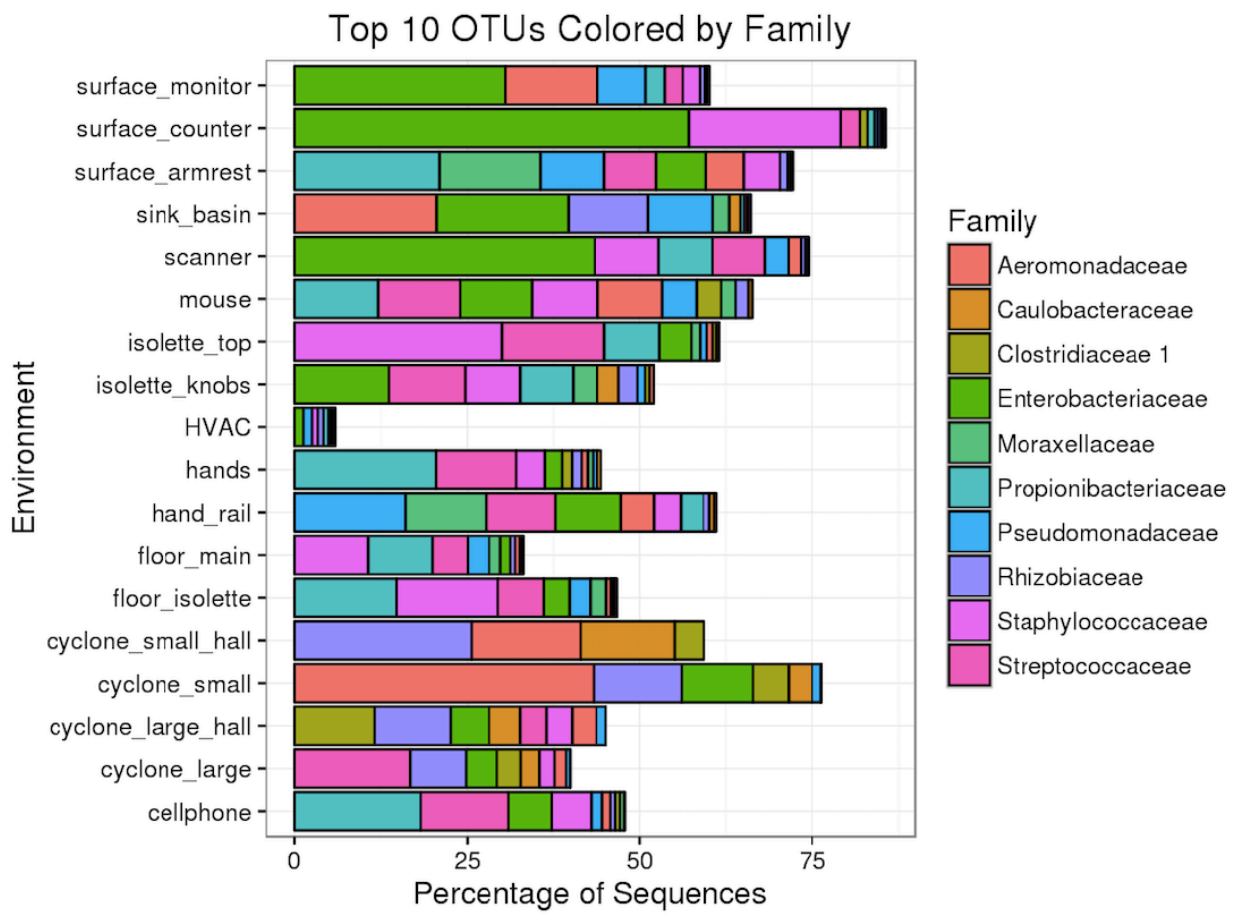


Figure 3-4: SourceTracker reveals human skin is dominant source of NICU microbes

American Gut skin, oral, and fecal samples were used as "sources" and NICU room samples were used as "sinks" and input into the SourceTracker software. Plotted on the *y*-axis is the mean relative contribution of each human-associated sources to each environmental sample.

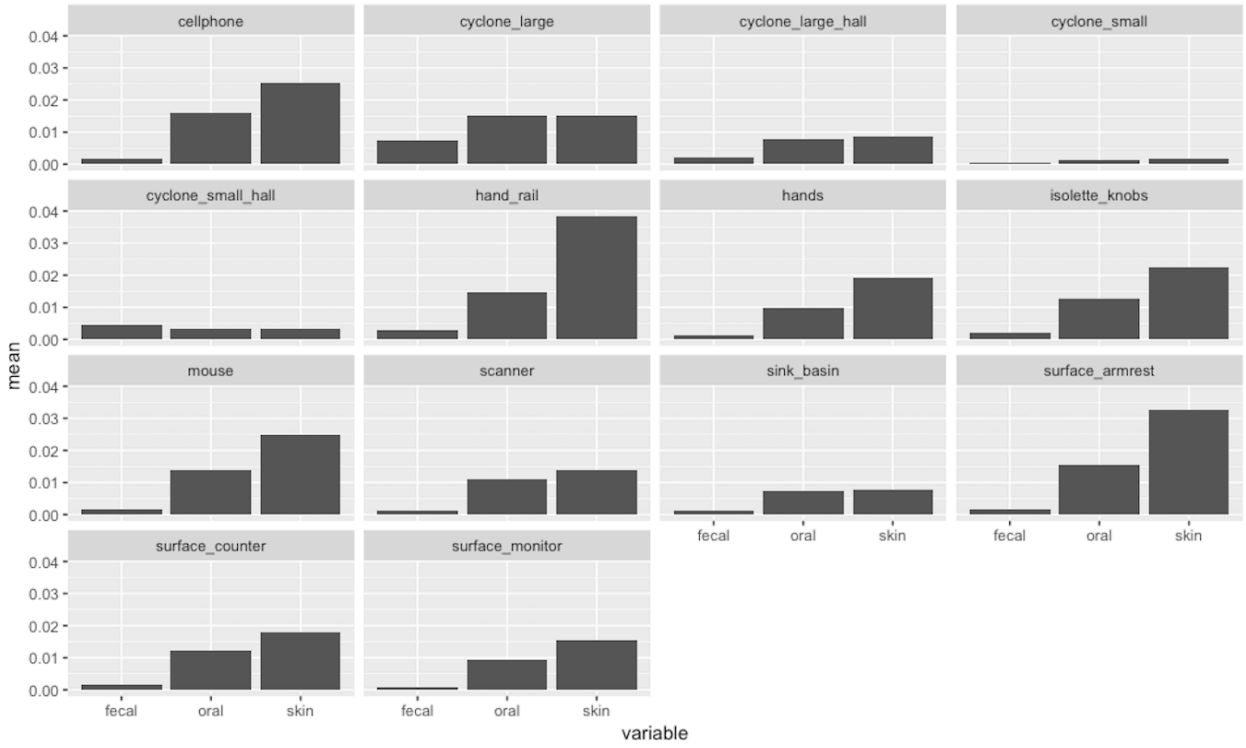
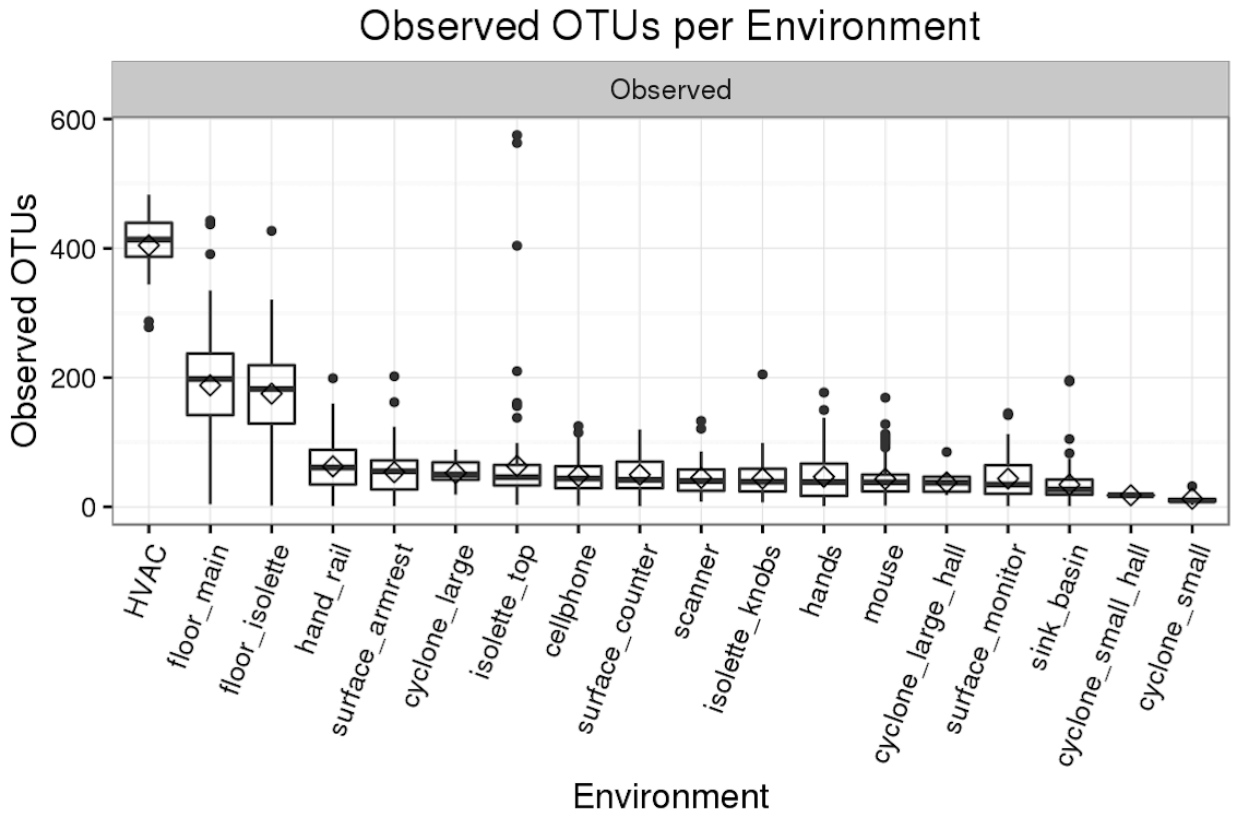


Figure 3-5: Alpha-diversity in the NICU

16S rRNA amplicon data was used to calculate number of OTUs per environment (a) and the Shannon diversity (b).



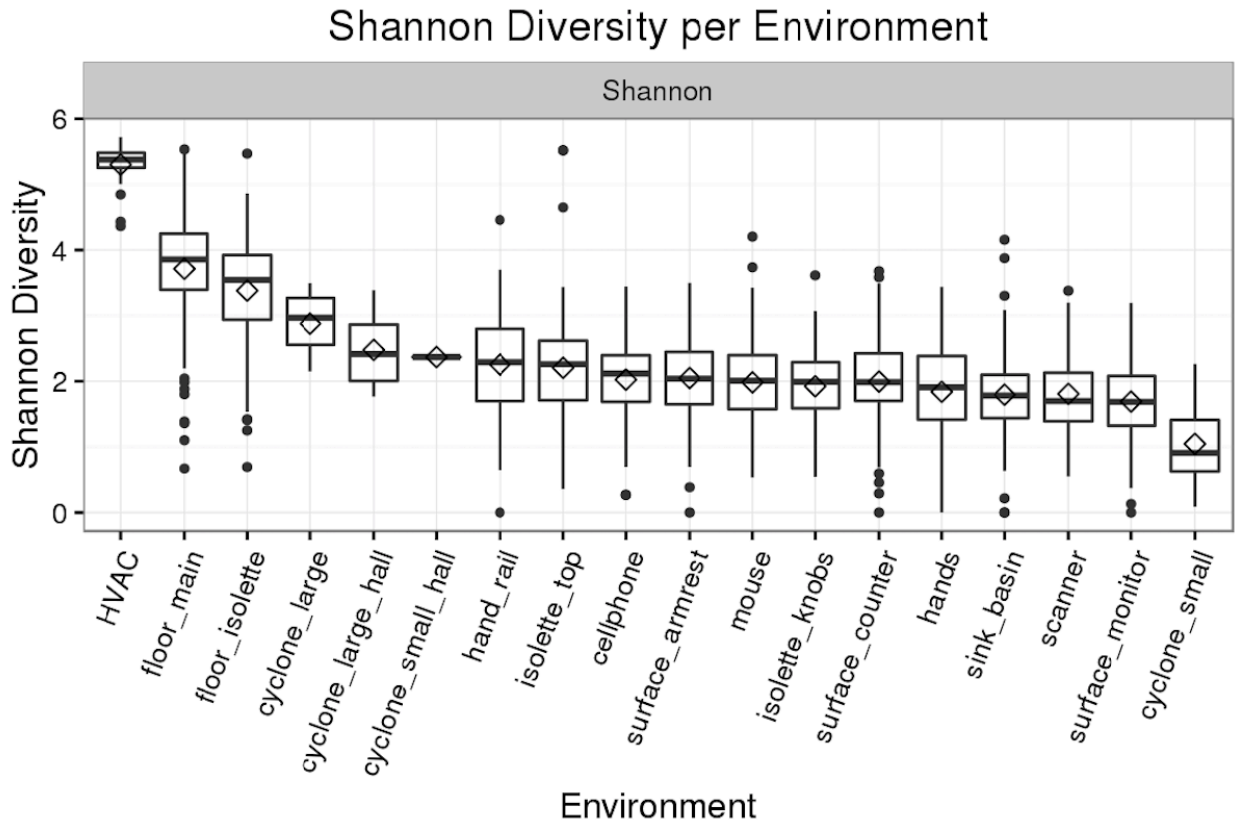
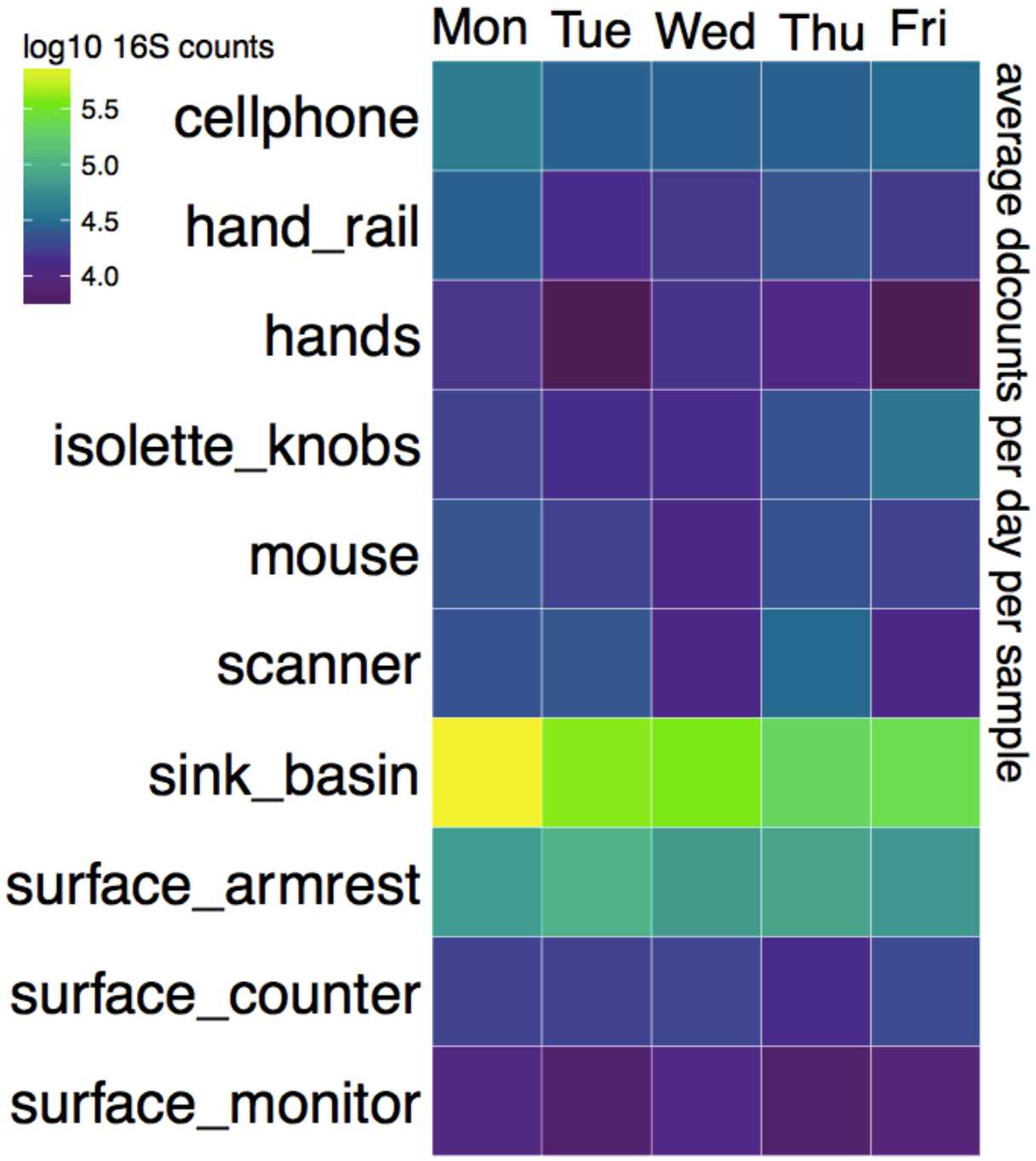

Shannon Diversity per Environment
Shannon
Shannon Diversity
6
4
2
0
HVAC
floor_main
floor_isolette
cyclone_large
cyclone_large_hall
cyclone_small_hall
hand_rail
isolette_top
cellphone
surface_armrest
mouse
isolette_knobs
surface_counter
hands
sink_basin
scanner
surface_monitor
cyclone_small
Environment

Figure 3-6: Growth detected in NICU sink samples

16S rRNA template copy number was quantified via ddPCR. Average copy number was averaged for each weekday and swabbed environment and displayed in this heatmap (a). 16S rRNA amplicon data was used to calculate number of OTUs, Shannon, and Inverse Simpson diversity metrics for sink basin samples (b). Black diamonds represent averages per weekday.



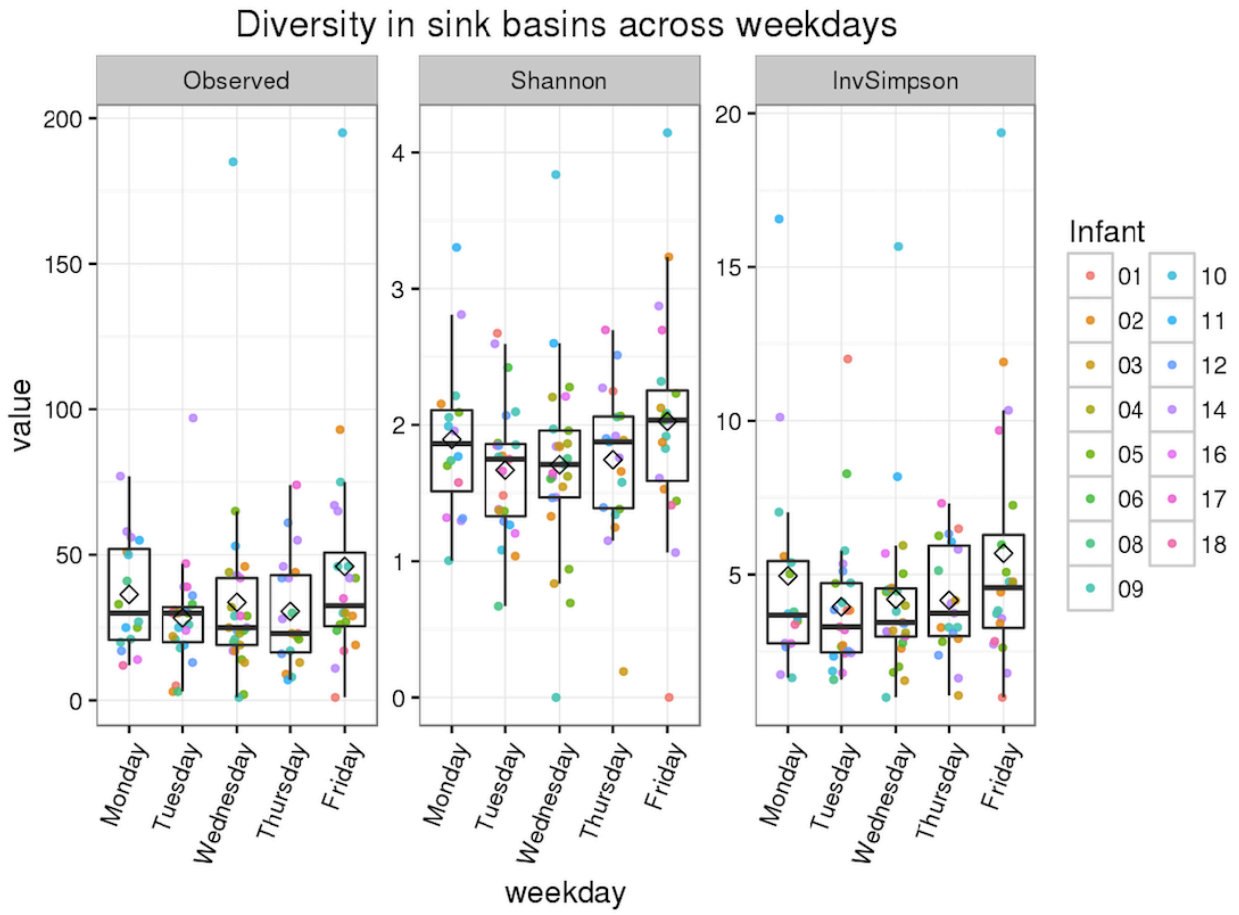

Diversity in sink basins across weekdays
Observed
Shannon
InvSimpson
value
weekday
Monday
Tuesday
Wednesday
Thursday
Friday
Infant
01
02
03
04
05
06
08
09
10
11
12
14
16
17
18

Figure 3-7: NICU rooms have a unique microbial signature

16S rRNA amplicon data was split into training, test, and validation sets to train, test, and validate a support vector machine classifier. The confusion matrix plots the accuracy of our model on the validation dataset. Percentages note the number of times a sample was predicted to belong to a room-infant pairing divided the total number of samples for that room-infant pairing. The heat coloring is based on shown percentages.

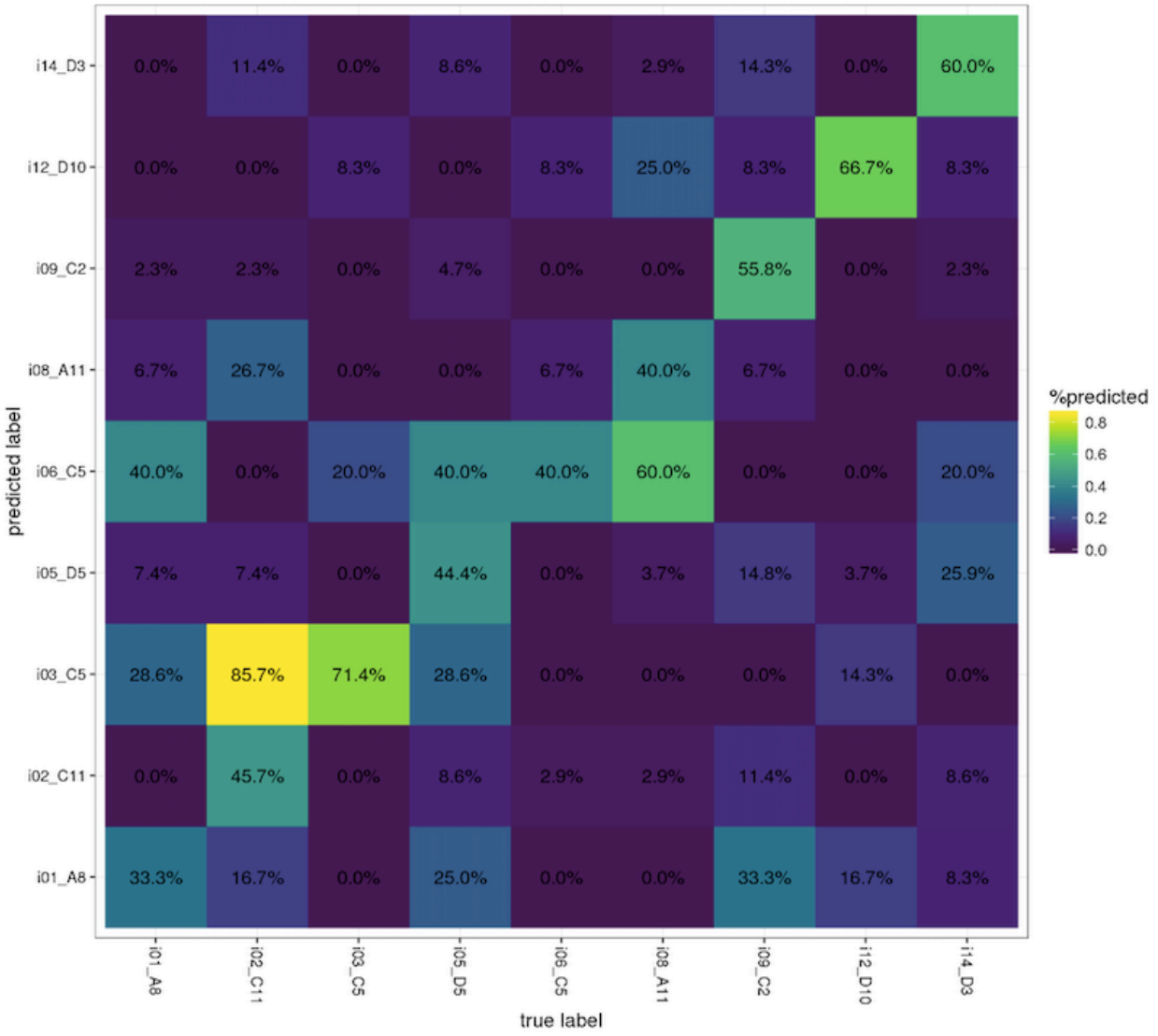
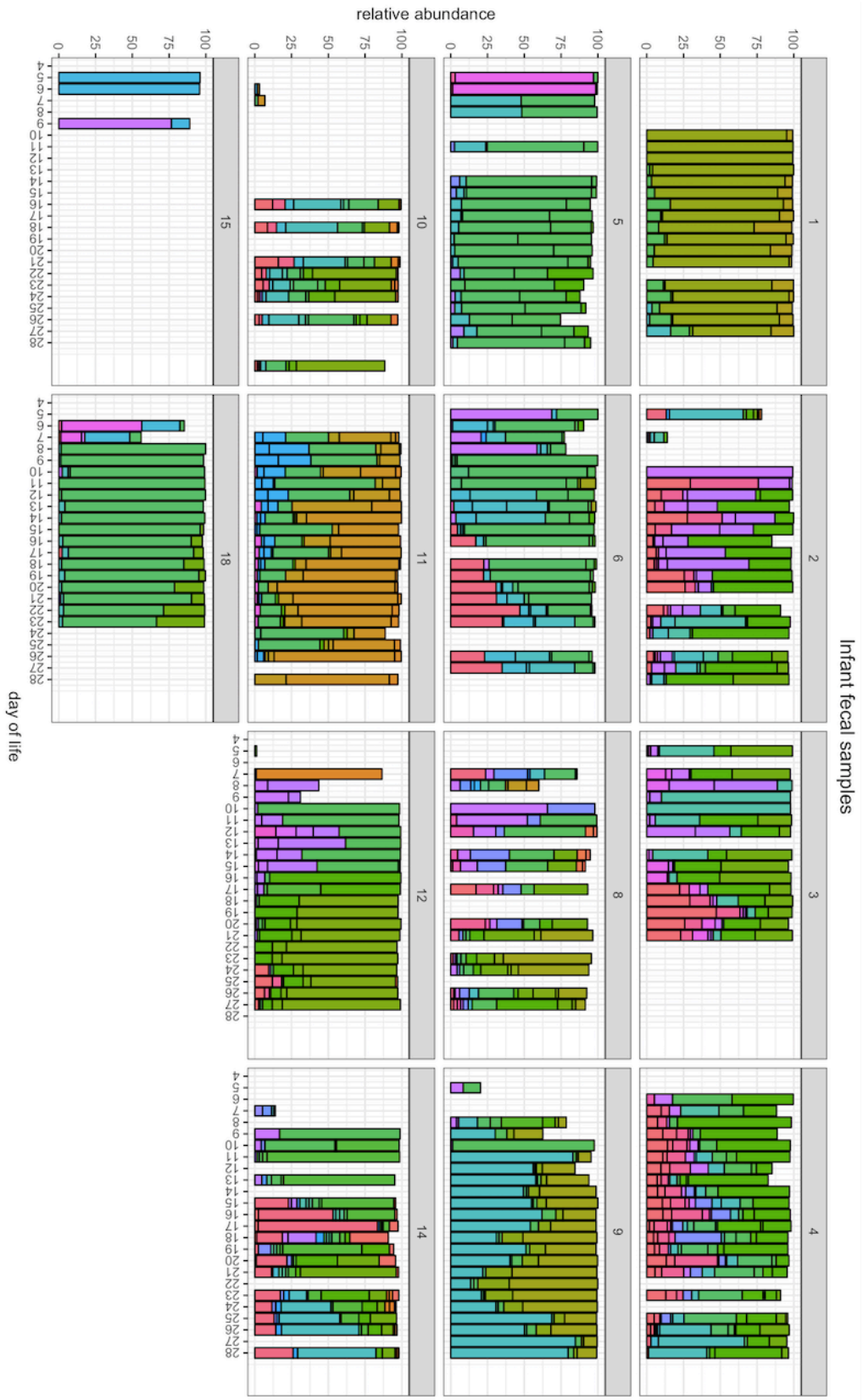


Figure 3-8: Top 10 most important taxa driving the machine learning model

The top 10 most important variables driving the SVM model are plotted for each infant. On the *y*-axis, "Abundance", notes the relative importance.

Figure 3-9: Fecal sample community composition

Plotted in each panel is the community composition of each infant's fecal samples derived from metagenomics data.


relative abundance
Infant fecal samples
day of life

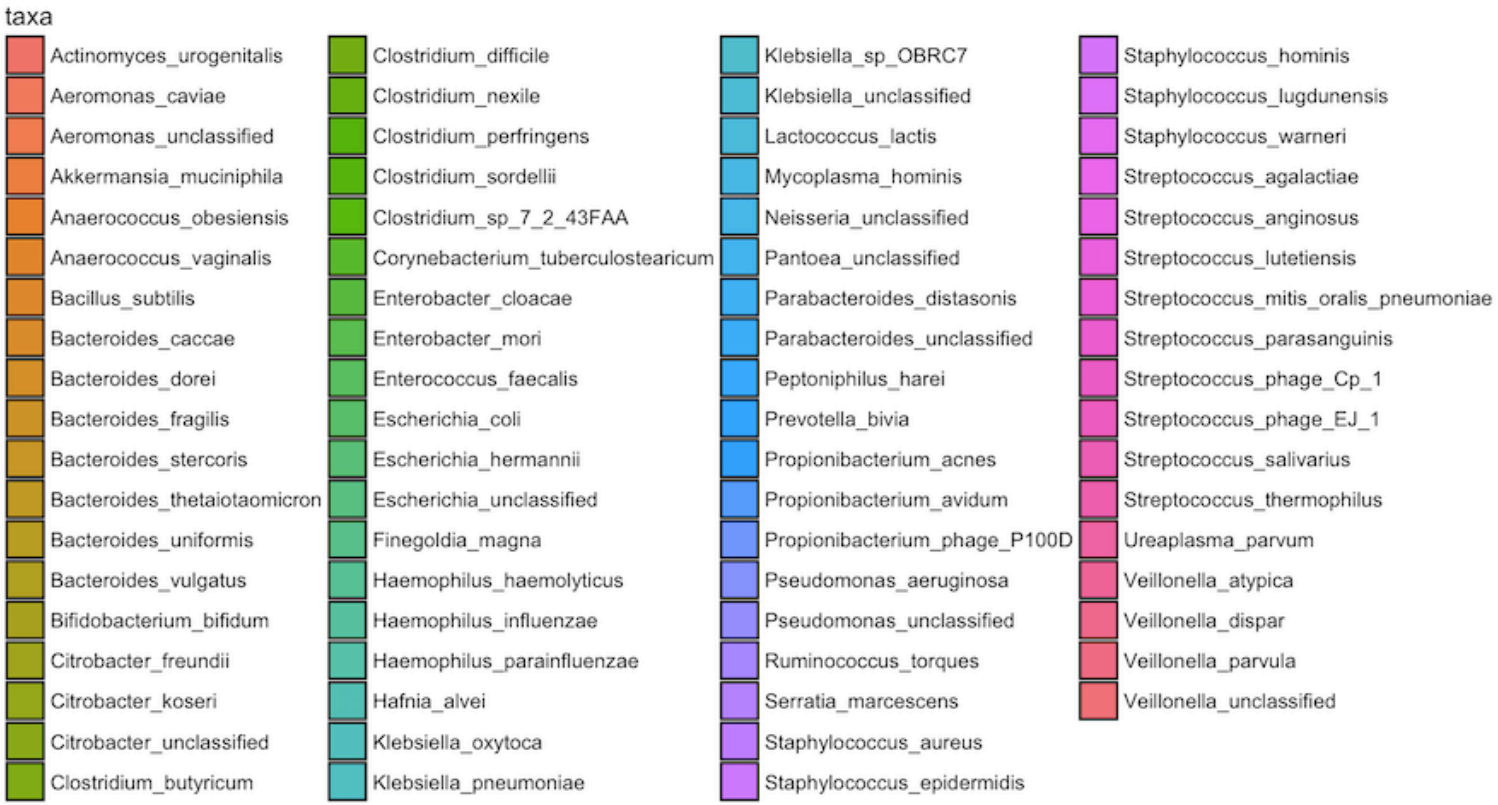
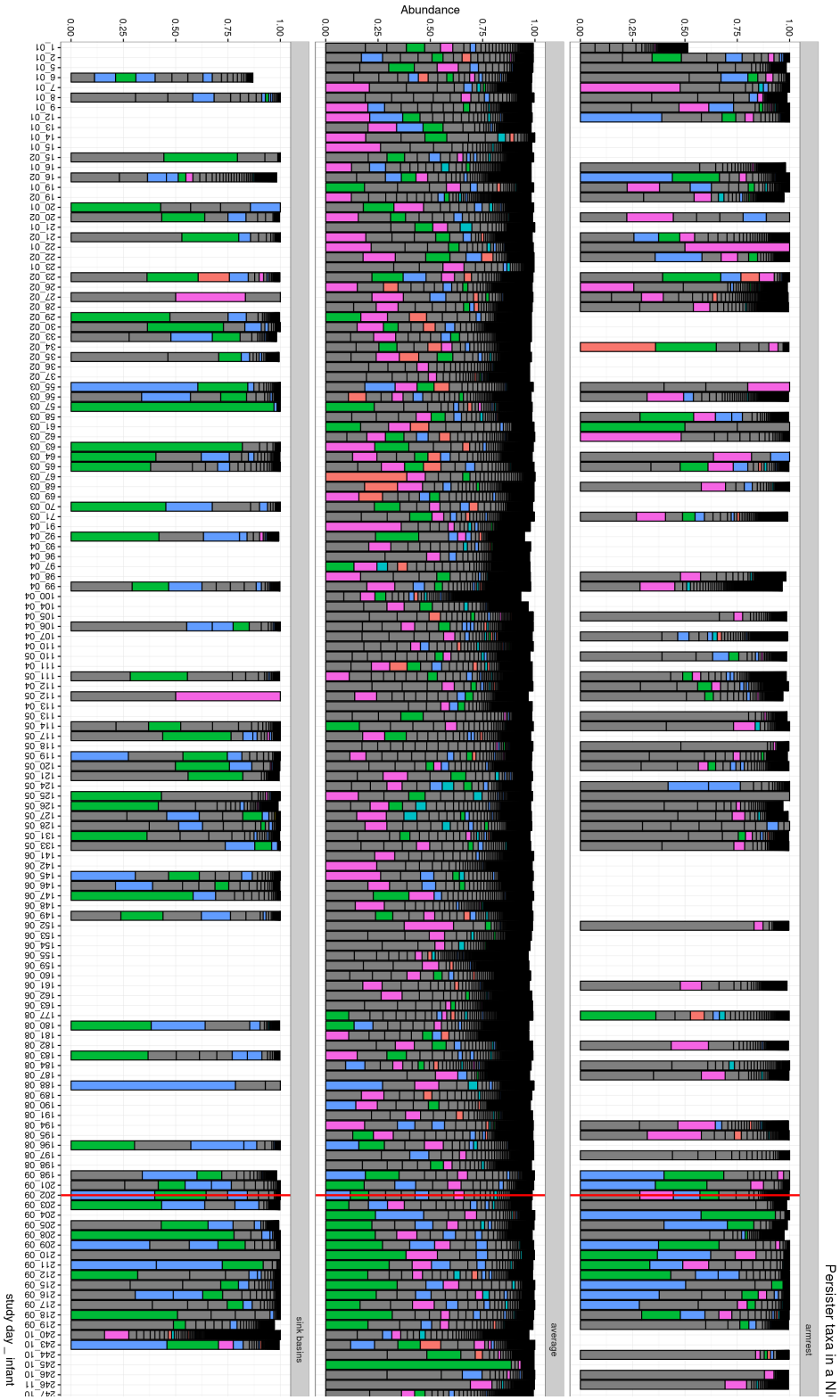

taxa
Actinomyces_urogenitalis
Aeromonas_caviae
Aeromonas_unclassified
Akkermansia_muciniphila
Anaerococcus_obesiensis
Anaerococcus_vaginalis
Bacillus_subtilis
Bacteroides_caccae
Bacteroides_dorei
Bacteroides_fragilis
Bacteroides_stercoris
Bacteroides_thetaiotaomicron
Bacteroides_uniformis
Bacteroides_vulgatus
Bifidobacterium_bifidum
Citrobacter_freundii
Citrobacter_koseri
Citrobacter_unclassified
Clostridium_butyricum
Clostridium_difficile
Clostridium_nexile
Clostridium_perfringens
Clostridium_sordellii
Clostridium_sp_7_2_43FAA
Corynebacterium_tuberculostearicum
Enterobacter_cloacae
Enterobacter_mori
Enterococcus_faecalis
Escherichia_coli
Escherichia_hermannii
Escherichia_unclassified
Finegoldia_magna
Haemophilus_haemolyticus
Haemophilus_influenzae
Haemophilus_parainfluenzae
Hafnia_alvei
Klebsiella_oxytoca
Klebsiella_pneumoniae
Klebsiella_sp_OBRC7
Klebsiella_unclassified
Lactococcus_lactis
Mycoplasma_hominis
Neisseria_unclassified
Pantoea_unclassified
Parabacteroides_distasonis
Parabacteroides_unclassified
Peptoniphilus_harei
Prevotella_bivia
Propionibacterium_acnes
Propionibacterium_avidum
Propionibacterium_phage_P100D
Pseudomonas_aeruginosa
Pseudomonas_unclassified
Ruminococcus_torques
Serratia_marcescens
Staphylococcus_aureus
Staphylococcus_epidermidis
Staphylococcus_hominis
Staphylococcus_lugdunensis
Staphylococcus_warneri
Streptococcus_agalactiae
Streptococcus_anginosus
Streptococcus_lutetiensis
Streptococcus_mitis_oralis_pneumoniae
Streptococcus_parasanguinis
Streptococcus_phage_Cp_1
Streptococcus_phage_EJ_1
Streptococcus_salivarius
Streptococcus_thermophilus
Ureaplasma_parvum
Veillonella_atypica
Veillonella_dispar
Veillonella_parvula
Veillonella_unclassified

Figure 3-10: Episodic increases of "persister" taxa in the NICU

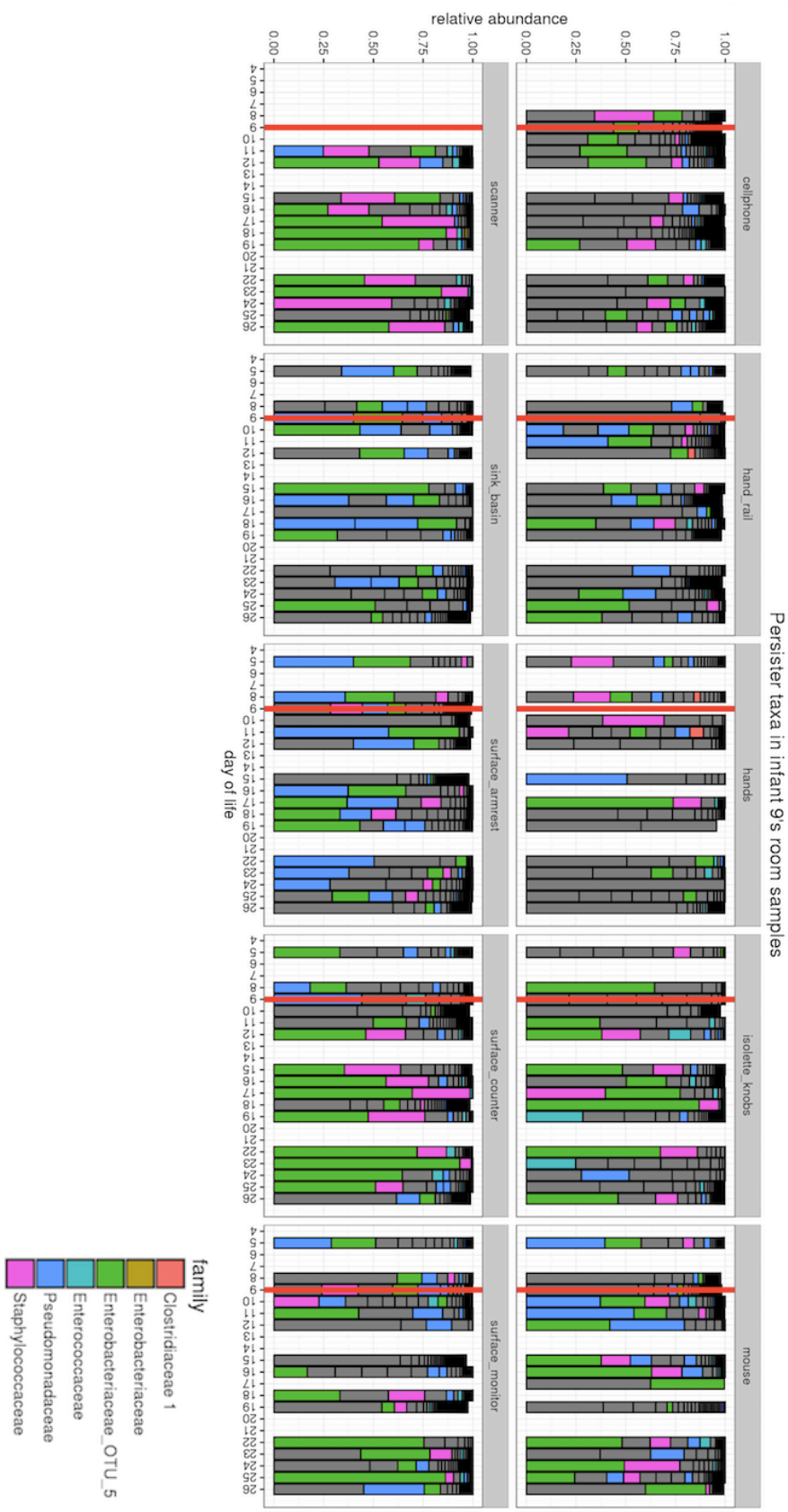
The "average" panel represents 16S amplicon data averaged across all samples at each time point per infant. The "armrest" and "sink_basins" panel is the same data but without averaging across environments. The red line highlights the time point in which an increase of *Enterbacteriaceae* was detected in infant 9's gut. Samples are plotted in chronological order on the x-axis. The plot is split across two pages for clarity.


Abundance
0.00
0.25
0.50
0.75
1.00
sink basins
average
armest
Persister taxa in a NI
study day _ infant

sister taxa in a NICU
anmrest
average
sink basins
study day _ infant
family
Clostridiaceae 1
Enterobacteriaceae
Enterobacteriaceae_OTU_5
Enterococcaceae
Pseudomonadaceae
Staphylococcaceae

Figure 3-11: Persister taxa in the room reflect composition of the infant gut

Infant 9's room amplicons are plotted for each swabbed environment (a). Colored are OTUs that belong to a persister lineage. Red lines highlight day of life 9, which coincides with an increase of several *Enterobacteriaceae* taxa in the infant gut (b). (b) is the microbial profile for fecal samples generated via genomes recovered from a metagenomics approach.


Persister taxa in infant 9's room samples
relative abundance
day of life
cellphone
hand_rail
hands
isolette_knobs
mouse
scanner
sink_basin
surface_armrest
surface_counter
surface_monitor
family
Clostridiaceae 1
Enterobacteriaceae
Enterobacteriaceae_OTU_5
Enterococcaceae
Pseudomonadaceae
Staphylococcaceae

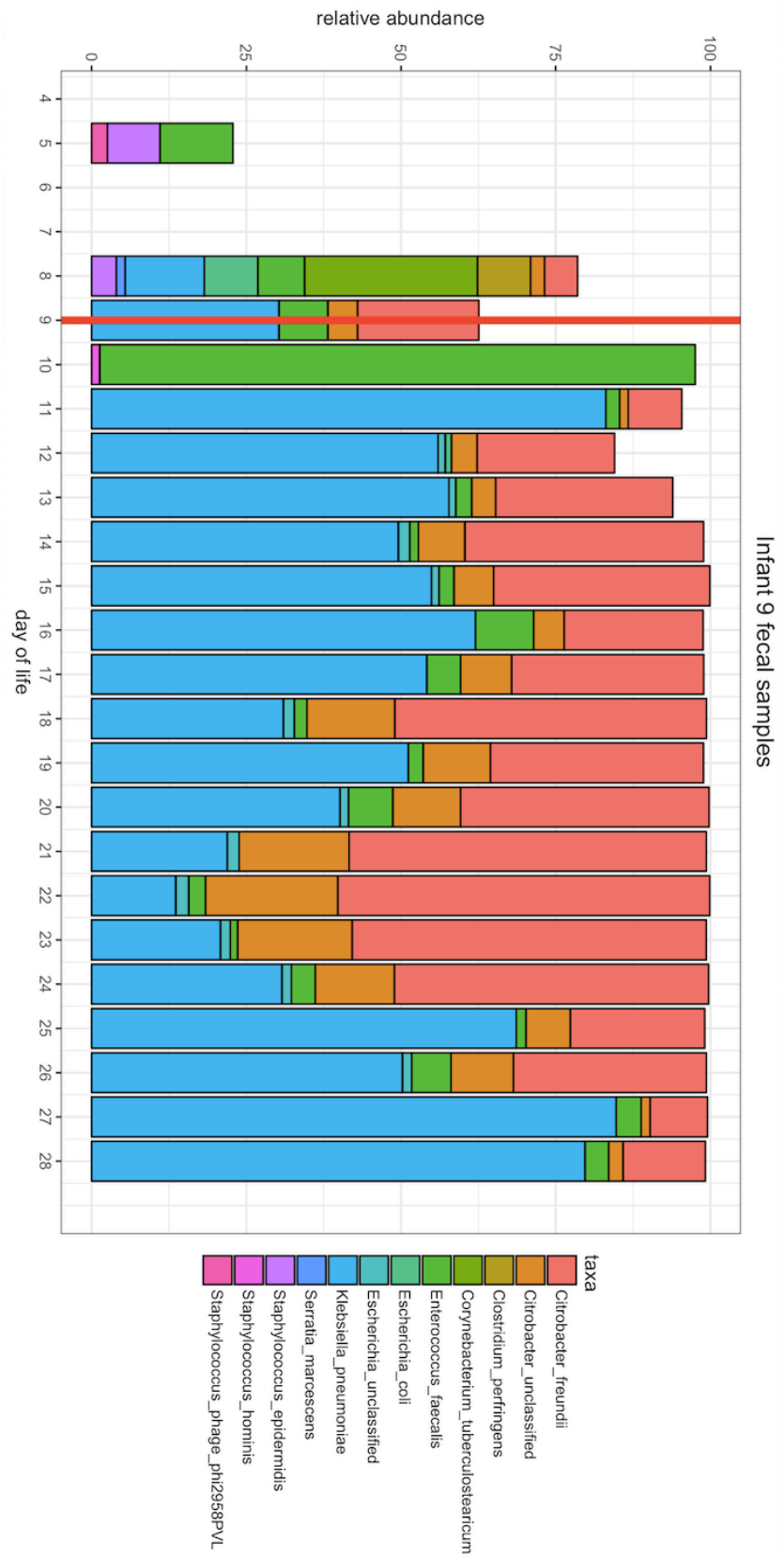

Infant 9 fecal samples
relative abundance
0
25
50
75
100
day of life
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
taxa
Citrobacter_freundii
Citrobacter_unclassified
Clostridium_perfringens
Corynebacterium_tuberculostearicum
Enterococcus_faecalis
Escherichia_coli
Escherichia_unclassified
Klebsiella_pneumoniae
Serratia_marcescens
Staphylococcus_epidermidis
Staphylococcus_hominis
Staphylococcus_phage_phi2958PVL

Table 3-1: Top 10 OTUs in the NICU

	Kingdom	Phylum	Class	Order	Family	Genus	Species	relative counts
OTU_5	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	?	?	232
OTU_6	Bacteria	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus	?	133
OTU_4	Bacteria	Actinobacteria	Actinobacteria	Propionibacteriales	Propionibacteriaceae	Propionibacterium	?	128
OTU_7	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus	uncultured organism	126
OTU_9	Bacteria	Proteobacteria	Gammaproteobacteria	Aeromonadales	Aeromonadaceae	Aeromonas	?	125
OTU_10	Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Rhizobium	?	82
OTU_8	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas	?	66
OTU_11	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter	?	42
OTU_29	Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae 1	Clostridium sensu stricto 1	uncultured organism	34
OTU_32	Bacteria	Proteobacteria	Alphaproteobacteria	Caulobacterales	Caulobacteraceae	Brevundimonas	?	33

Table 3-2: Most important variables to SVM model

	Kingdom	Phylum	Class	Order	Family	Genus	Species
OTU_29	Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae 1	Clostridium sunsu stricto 1	uncultured organism
OTU_39	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Micrococcaceae	Rothia	uncultured organism
OTU_41	Bacteria	Firmicutes	Bacilli	Bacillales	Family XI	Gemella	?
OTU_30	Bacteria	Actinobacteria	Actinobacteria	Micrococcales	Micrococcaceae	Kocuria	?
OTU_45	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Actinomyces	?
OTU_43	Bacteria	Firmicutes	Bacilli	Bacillales	Alicyclobacillaceae	Tumebacillus	uncultured Firmicutes bacterium
OTU_76	Bacteria	Firmicutes	Clostridia	Clostridiales	Family XI	Peptoniphilus	?
OTU_74	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Actinomyces	uncultured organism
OTU_28	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Veillonella	uncultured organism
OTU_66	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus	?

Chapter 4:

4 Hospitalized infants are colonized by microbes from the room environment

Brandon Brooks[1], Matthew R. Olm[1], Brian A. Firek[3], Robyn Baker[4], Brian C. Thomas[2], Michael J. Morowitz[3], Jillian F. Banfield[2]

1 – Department of Plant and Microbial Biology, University of California, Berkeley, CA
2 – Department of Earth and Planetary Sciences, University of California, Berkeley, CA
3 – University of Pittsburgh School of Medicine, Pittsburgh, PA
4 – Division of Newborn Medicine, Children's Hospital of Pittsburgh of UPMC, Pittsburgh, PA

4.1 Abstract and Introduction

Infants are born with a near sterile microbiome and acquire most of their initial colonizers during birth (Dominguez-Bello *et al.*, 2010). Extremely preterm infants, however, are often treated with antibiotics that reset the colonization process (Raveh-Sadka *et al.*, 2015). These infants exhibit categorically different colonization patterns relative to full term infants, and it is speculated that the infant's room environment may contribute to infant microbiome colonization (Brooks *et al.*, 2014; Shin *et al.*, 2015). Here, we conducted a genome-resolved metagenomics study that enabled comparison of genotypes present in the gastrointestinal tracts of infants with those of bacteria in the neonatal intensive care unit (NICU) room environment. We show that strains detected in hospitalized infants also occur in sinks and on surfaces. Interestingly, a few room-associated strains are indistinguishable from strains reported from around the world, suggesting their widespread dissemination in the human population. However, comparative genomic analyses revealed that many strains shared by the room and infants are distinct from strains reported from other systems, supporting the conclusion that the room environment is a major reservoir of bacteria that colonize hospitalized infants. Further, *Pseudomonas aeruginosa*, *Klebsiella oxytoca* and *Escherichia coli* that colonized infants were present in NICU rooms months before infant colonization occurred. Even more importantly, although *Enterococcus faecalis* is relatively uncommon in the room environment, the strains that were detected are those that are shared amongst infants over a multi-year period. Thus, we conclude that an important component of infant colonization is the cycle of room and occupant exchange in which an infant's immediate room environment shapes early stage microbiome colonization. Where a premature infant is born and the history of room occupancy can impact infant microbiome development.

4.2 Results and Discussion

Almost 10% of all births in the United States are preterm (Hamilton *et al.*, 2015). Many preterm infants are immunocompromised and are especially susceptible to hospital-acquired infections (HAI). To reduce the risk of infection, antibiotic treatment is common, and accounts for the top three of six medications administered in the NICU (Gasparrini *et al.*, 2016). Ironically, following this rigorous broad spectrum antibiotic treatment period, many preterm infants acquire an initial microbiome that resembles a collection of common hospital-associated pathogens (Gibson *et al.*, 2016). These strains are often resistant to the antibiotics administered to the infant, explaining their fitness advantage and frequency of detection. What remains unclear, however, is the source of these strains.

Previous hospital microbiome studies have implicated the room environment as an important source of strains that colonize hospitalized infants, but these studies were conducted using 16S rRNA surveys that cannot reliably distinguish different bacteria from the same taxonomic family (Tu *et al.*, 2014; Jovel *et al.*, 2016). This limitation is important. For example, there are dozens of different genera within the family *Enterobacteriacaea*, including *Escherichia*, *Klebsiella, Yersinia, Serratia,* and *Citrobacter*. Even within a particular genus, there are numerous distinct species. For example, *Escherichia fergusonii* and *E. coli* have distinct physiologies and medical implications (Luo *et al.*, 2011). Further, within *E. coli*, there are numerous strains that can differ enormously in their pathogenicity and antibiotic resistance

(Tenaillon *et al.*, 2010). Methods that can determine which strains are present are necessary to ascertain the potential medical significance of room- and infant-associated bacteria. Similarly, strain-resolved methods are essential to determine if the room environment is the actual source for organisms detected in infants or if related organisms were introduced from external sources. The taxonomic limitation of rRNA gene survey-based methods motivated us to conduct the first metagenomic study in which genomes were reconstructed from room surfaces and from the fecal samples of premature infants hospitalized in those rooms. Thus, we could test for a direct link between strains in the NICU environment and the infant gut.

Sixteen preterm infants (< 31 week's gestation, < 1500 g) housed in a NICU in the USA were studied from day of life (DOL) 5 to DOL 28. We collected 295 fecal samples from the infants (primarily obtained by perineal stimulation) at the same time that 3700 room samples were obtained using swabs and wipes. Room samples were derived from a variety of touched surfaces as well as from the interior of handwashing sink basins. Overall, 22 different hospital surface types were represented. DNA was extracted from all fecal samples. Due to extremely low biomass, DNA from multiple samples collected at different times from the same room were pooled, generating three sample types per room: swabs, wipes, and sinks. DNA from both fecal samples and room samples was sequenced, generating a total of ~1 Tb of data. Sequencing reads were trimmed, assembled and the data binned using a previously described metagenomics analysis pipeline (Olm *et al.*, 2016). For room samples, sequencing allocations were increased relative to allocations for fecal samples to compensate for the expected higher levels of microbial diversity in the room samples. Our approach generated hundreds of draft quality genomes for bacteria present in both the infant gut and room samples. The successful recovery of reasonably high quality genomes directly from room-collected samples is unprecedented.

Recent genome-based metagenomic studies suggested that most strains that colonize infants in the same NICU are not shared amongst cohoused infants, although similar species and strains were identified (Raveh-Sadka *et al.*, 2015). Of the few strains that are shared by different infants, many reoccurred in samples from infants present in the NICU several years apart (Raveh-Sadka *et al.*, 2016). A "strain" or "strain type" was considered the same if two near-complete genome bins had greater than 98% average nucleotide identity (ANI) across 95% of the bin. Currently, there is no standard convention for strain identity thresholds in bacteria. This classification likely will not be finalized soon, since the species concept of bacteria has been contested for decades (Ellegaard and Engel, 2016). Though, the ability to readily reproduce the same genomes at a high identity (> 99.9% ANI) across multiple time points from the same infant is strong support that these methods are reproducing the same "strain." Perhaps even more compelling is the genomes have the same level of identity to reference genomes which were generated via isolation and Sanger sequencing (discussed below). The results suggested to us that room reservoirs may exist for "persister" strains, and that reseeding of sequential room occupants occurs from these reservoirs. To test this hypothesis, we compared the bacterial genome sequences reconstructed from room and infant samples to determine whether the same strains occurred in both environments. The genomes were clustered at an ANI > 98% using the MASH algorithm (Ondov *et al.*, 2015). Three strains (*Escherichia coli, Enterococcus faecalis, Pseudomonas aeruginosa*) from among the five persister strains previously reported were recovered from room samples (Figure 4-1). Thus, we conclude that indeed, room habitats may host bacteria that repeatedly colonize infants housed in the NICU.

We detected a high degree of overlap in bacterial strain composition between room and infant samples (Figure 4-1). Of the 67 distinct bacterial genotypes found in more than one infant

in the studied cohorts, 17 were recovered from room samples (25%). Interestingly, Firmicutes (*Clostridium* and *Enterococcus*) were relatively rarely detected in room habitats. This observation may indicate that these obligately anaerobic bacteria are not well suited to grow in the room environment. Alternatively, they may have been under-detected, possibly because of sporulation triggered by the hostile room conditions, increasing the difficulty of DNA extraction.

The overlap in room and infant microbes cannot distinguish seeding of the room from the infant versus colonization of the infant from a room reservoir. Analysis of directionality can be undertaken using time series data. We sequenced room samples from infant 5 that were collected at different times during hospitalization. Samples collected early, middle and at late time points were pooled to ensure sufficient DNA for sequencing while still accounting for sampling time (DOL 5-12, 13-20, and 21-28, respectively). No organisms were detected in early room pools before they were detected in infant 5's fecal samples (Figure 4-2). This observation does not necessary rule out room to infant transfer, but does indicate infant to room dispersal. Following detection in the infant, all infant-associated strains, excluding one *Streptococcus*, were detected in the infant's room samples (Figure 2-2). This pattern led us to posit that infants disperse microbes into the room where the microbes can persist, and the persistent microbes may later colonize other infants housed in the same room.

To test for a cyclical pattern of infant-room-infant transfer, we further investigated genomes of bacteria that were most frequently found in both room and infant samples. While MASH clustering at > 98% ANI provides high confidence that two bacteria belong to the same species and likely the same strain, strains-level resolution can be imprecise. Thus, we implemented a highly sensitive approach (Olm *et al.*, 2016) that involved mapping of sequencing reads from room samples to a database of genomes reconstructed from infant 5's fecal samples. Near complete genomes were recovered from nearly all room samples for *Klebsiella oxytoca*, *Pseudomonas aeruginosa*, *Staphylococcus epidermidis*, and *Klebsiella pneumoniae*. We found that the bacteria represented by these room sample-derived genomes appeared in samples collected from infants months to years before and after their detection in the room. Excluding reads from infant 5's room, the highest identity match showed infant 3's sink basin reads mapping to infant 5's *Pseudomonas aeruginosa* gut genome at 99.9% ANI at 98.7% genome breadth. We also compared reads from all rooms to recently acquired skin and oral metagenomes from infant 5 and found the highest identity hit is infant 18's room swab reads to infant 5's oral *Staphylococcus epidermidis* genome. Remarkably, 8 of 14 genomes detected in infant 5's gut were identified in the room reads from other infants' rooms (ANI > 99% and genome breadth > 90%). Not detected were *Clostridium perfringens, Enterococcus faecalis, Propionibacterium sp. HGH0353, Serratia marcescens, Streptococcus mitis, Streptococcus sp. SK140*. Based on infant 5's room reads results, all of the bacteria in infant 5's gut were identified in infant 5's room except *C. perfringens* and *S. sp. SK140*. The detection of room strains before and after detection in infant 5's fecal samples supports our model of cyclic passage between room and human reservoirs. However, we cannot rule out the possibility that virtually all strains are everywhere and host selection is driving the observed colonization patterns.

We next applied the same room read mapping approach to a database of publicly available genomes and genomes recovered from infants in this NICU prior to the current study. We first focused on *Pseudomonas aeruginosa* since it was the most commonly recovered genome in room samples (14/24 samples). Seventy-two complete NCBI genomes and 10 previously assembled genomes from our lab were included in the database. Interestingly, the highest scoring match was from infant 18's sink sample reads and a genome deposited by a

group in Orsay, France in 2016 (99.9% ANI and 99.9% genome breadth). This finding may indicate that this *P. aeruginosa* strain is widely distributed in the human population.

Many of the top hits for room reads obtained in the current study were to genomes previously assembled from the same hospital. This finding confirms that "persister" strains occur in the room environment (Table 4-1). Interestingly, a high quality match occurred between the room reads that we obtained from samples collected in 2014 in Pittsburgh and the genome of a bacterium that was isolated from a burn patient in 2006 in Boston. In fact, the *P. aeruginosa* genomes only differ by 78 SNPs. The broad distribution of this strain suggests their facile dispersal and strong selective pressures for this genotype within the hospital environment.

Next we compared the genome of *Klebsiella oxytoca*, the second most frequently recovered genotype detected in the room samples (13/24 samples), to a database that includes 18 NCBI genomes and 8 previously assembled genomes from our lab. Unlike *Pseudomonas*, none of the publicly available genomes had high scoring hits. The closest strain was isolated in 2015 from a group in Jikei, Japan (ANI 99.8%, genome breadth of 87.2%). In contrast, the highest scoring match for infant 5's room swab reads was to a genome that we reconstructed from fecal samples collected in 2012. The ability to only detect high quality hits of room reads to genomes of gut-associated bacteria previously reconstructed from samples collected from infants in the same hospital makes a links between these *K. oxytoca* populations highly probable. However, a better representation of *K. oxytoca* genomes from other localities in reference databases may increase the probability that some infant-associated populations are externally derived.

Although many strains found in infants are also found in their rooms and other rooms, the opposite is not true. For example, many sink-associated bacteria (including numerous Gammaproteobacteria) do not colonize infants. Further, only a subset of species/strains in many room samples is found in infants. Thus, it is likely that environmental selection has some role in shaping even the earliest establishing microbiomes. The counterpoint is that we recently showed that identical strains colonize mouth, skin and gut habitats of premature infants (Olm *et al.*, 2016). In combination, the prior work and genome-resolved studies reported here suggest that, for those strains that can colonize infants, there is a relatively low level of host selection during initial microbiome development. Thus, it seems likely that niche occupation by closely related taxa, via the founder effect (Waters *et al.*, 2013), may be a key determinant in early infant microbiome colonization.

Preterm infants hospitalized within a NICU constitute a particularly vulnerable cohort with a high risk for hospital-acquired infections. Here, we directly link genomes from the room that later colonize infants several days to years later. Evidence supports the suggestion that the rooms are a reservoir for early stage colonizers of the infant microbiome. The results may be generalizable to other patient populations. An important implication of our findings is that the hospital in which a premature infant is born, the room that it occupies, and the infants previously housed there can shape early microbiome development. Given that gut colonization patterns during this time period are critical to human health (Cahenzli *et al.*, 2013; Costello *et al.*, 2012; Arrieta *et al.*, 2015; Sim *et al.*, 2013), further research exploring the properties of the NICU built environment and its microbiome may provide clever ways to protect these vulnerable patient populations.

4.3 Competing Interests

The authors declare that they have no competing interests.

4.4 Authors' Contributions

JFB, MJM, and BB conceived of the project. RB organized cohort recruitment and sample collections. BAF conducted nucleic acid extractions and BB conducted the sample pooling. BB conducted the metagenomic assemblies and BCT provided annotations. BB and MO conducted the strain-level read mapping experiments. BB and JFB wrote the final manuscript. All authors have read and approve the manuscript.

4.5 Acknowledgments

Funding was provided through the Alfred P. Sloan Foundation and the National Science Foundation's Graduate Research Fellowship Program. This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 OD018174 Instrumentation Grant.

Figure 4-1: Similar room strains are found in the infant gut across several cohorts and years

Each infant and its affiliated time series of fecal samples is represented along the *x*-axis. The *y*-axis labels correspond to genomes that have > 98% ANI with other genomes across infants. Genome labels on the *y*-axis are for representative genomes within that cluster. Only genomes that were found in infants across different cohorts or are shared with a genome in the room are displayed. The blue panel highlights samples sourced from the room. Year of collection is provided in the key.

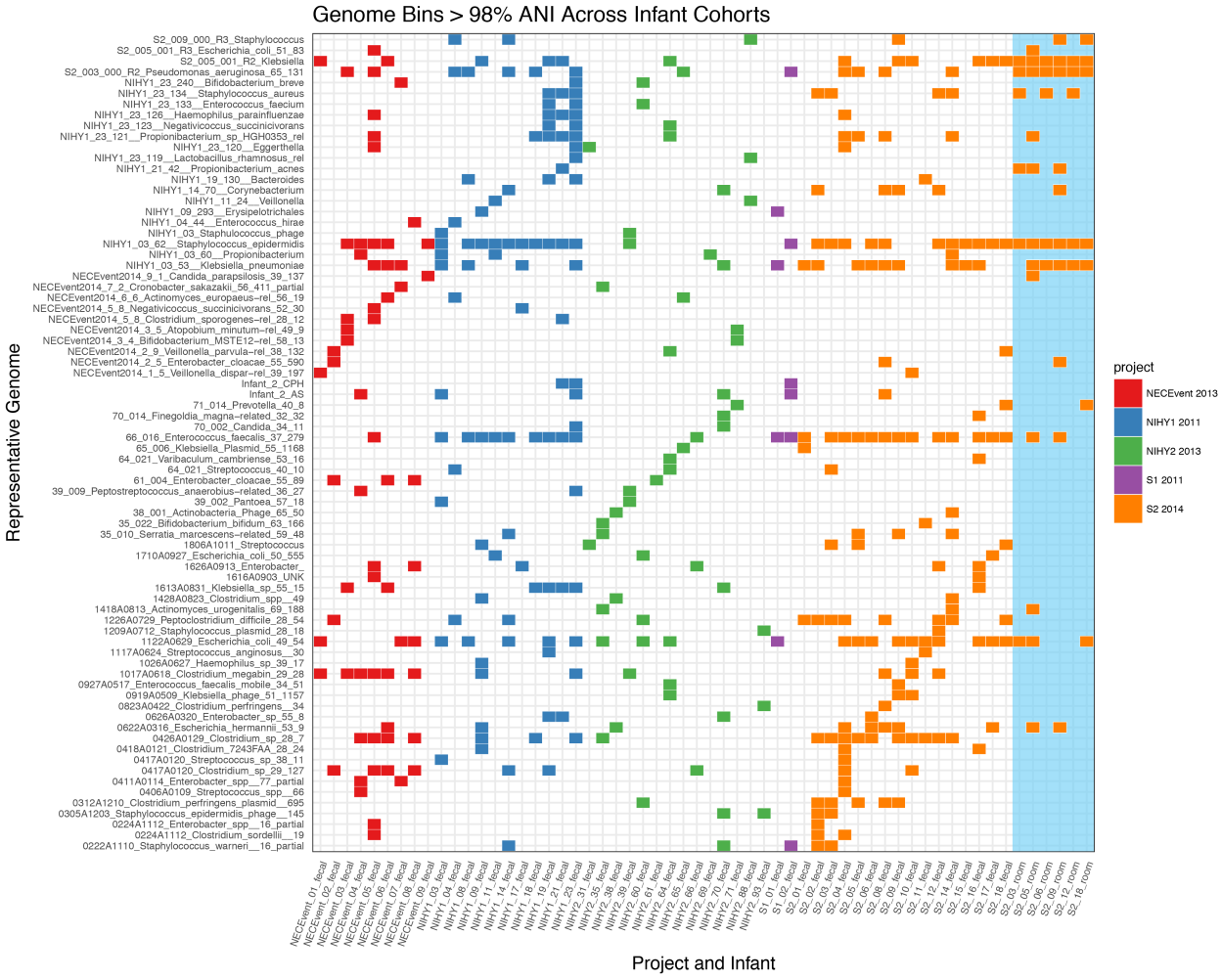


Figure 4-2: Time series room metagenomes reveal infant to room directionality

Infant 5's time series fecal and room samples are plotted on the *x*-axis in chronological order by time of collection. The *y*-axis labels correspond to genomes that have > 98% ANI with other genomes across samples. Genome labels (*y*-axis) are for representative genomes within that cluster. The blue panels highlight samples sourced from the room.

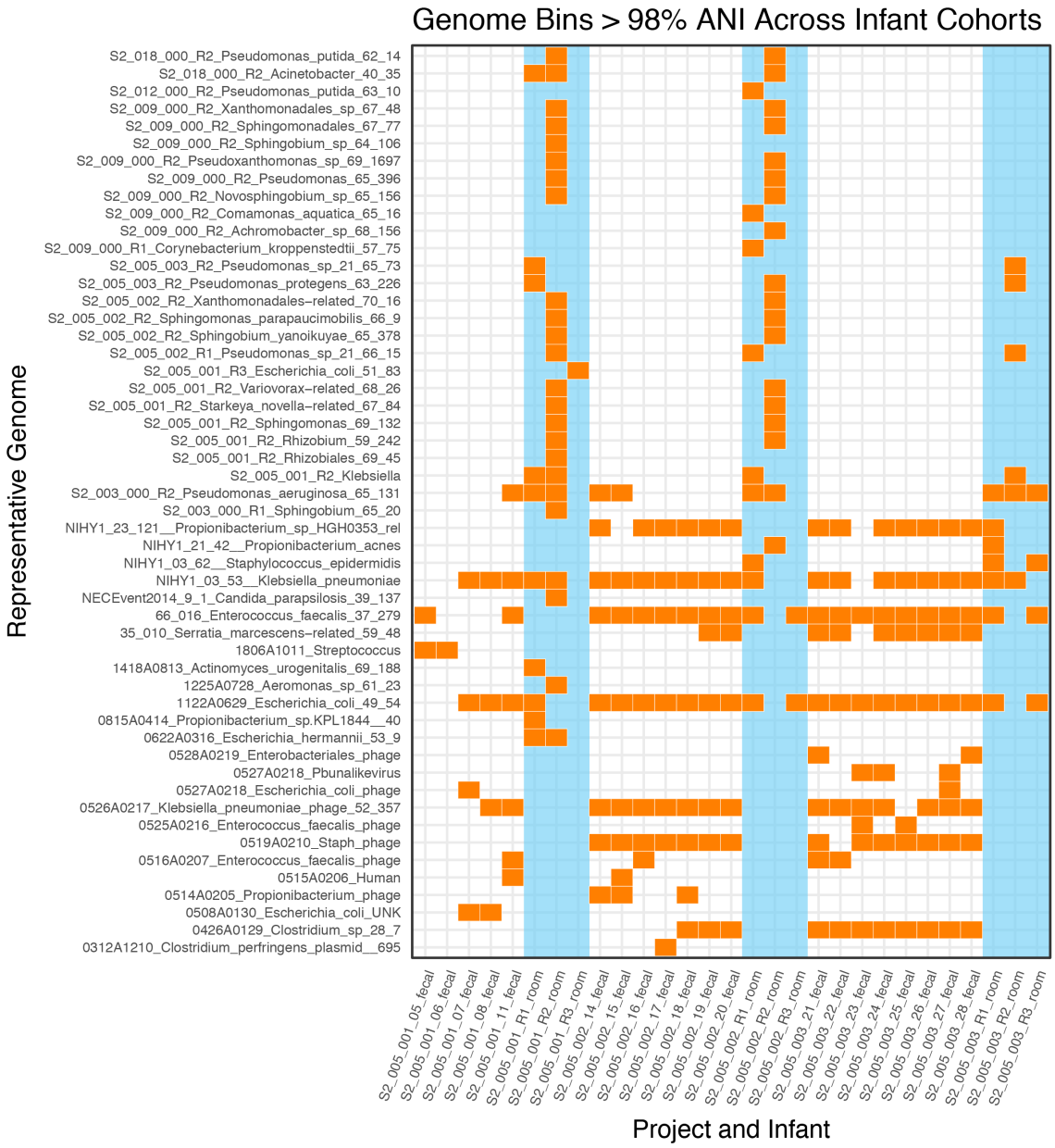


Table 4-1: Strains isolated from hospital sources have varying degrees of similarity to publicly available reference genomes

Room reads were mapped to a database of publicly available genomes and genomes previously isolated from this hospital by our lab to observe the global ubiquity of room microbes. The top 10 hits are provided.

SNPs	breadth	consensus_ANI	genome	sample	location	date deposited	source	
Pseudomonas aeruginosa								
68	0.999405028	0.999989394	GCF_900095805.1_PA14Or_genomic.fna	Infant 18 sink samples	Orsay, France	2016	NA	
86	0.999007125	0.999987096	NECEvent2014_3_6_Pseudomonas_aeruginosa_65_49.contigs.fna	Infant 18 sink samples	Pittsburgh, PA	2014	infant sample	fecal
78	0.998582212	0.999987809	GCF_000014625.1_ASM1462v1_genomic.fna	Infant 18 sink samples	Boston, MA	2006	clinical isolate; burn patient	
388	0.995511975	0.999935215	Infant_2_PA.contigs.fna	Infant 12 sink samples	Pittsburgh, PA	2011	infant sample	fecal
1426	0.987371486	0.999683952	NECEvent2014_3_6_Pseudomonas_aeruginosa_65_49.contigs.fna	Infant 12 sink samples	Pittsburgh, PA	2014	infant sample	fecal
464	0.987338114	0.999922458	NIHY1_18_45__Pseudomonas_aeruginosa.contigs.fna	Infant 12 sink samples	Pittsburgh, PA	2012	infant sample	fecal
105	0.986111251	0.999974379	GCF_900095805.1_PA14Or_genomic.fna	Infant 12 sink samples	Orsay, France	2016	NA	
85	0.983749355	0.999979234	GCF_000014625.1_ASM1462v1_genomic.fna	Infant 12 sink samples	Boston, MA	2006	clinical isolate; burn patient	
439	0.982505499	0.999931941	65_007_Pseudomonas_aeruginosa_66_63.contigs.fna	Infant 18 sink samples	Pittsburgh, PA	2013	infant sample	fecal
272	0.980995951	0.999957789	NIHY1_14_298__Pseudomonas_aeruginosa.contigs.fna	Infant 18 sink samples	Pittsburgh, PA	2012	infant sample	fecal
Klebsiella oxytoca								
123	0.991351015	0.999978546	NIHY1_19_62__Klebsiella_oxytoca.contigs.fna	Infant 5 early swab samples	Pittsburgh, PA	2012	infant sample	fecal
84	0.991057012	0.999985337	NIHY1_19_62__Klebsiella_oxytoca.contigs.fna	Infant 3 sink samples	Pittsburgh, PA	2012	infant sample	fecal
95	0.990613175	0.999983394	NIHY1_19_62__Klebsiella_oxytoca.contigs.fna	Infant 5 early sink samples	Pittsburgh, PA	2012	infant sample	fecal
121	0.989889236	0.999978808	NIHY1_19_62__Klebsiella_oxytoca.contigs.fna	Infant 6 sink samples	Pittsburgh, PA	2012	infant sample	fecal
341	0.989550781	0.99994017	NIHY1_19_62__Klebsiella_oxytoca.contigs.fna	Infant 12 sink samples	Pittsburgh, PA	2012	infant sample	fecal
164	0.988948185	0.99997711	NIHY1_19_62__Klebsiella_oxytoca.contigs.fna	Infant 5 late sink samples	Pittsburgh, PA	2012	infant sample	fecal
187	0.987765478	0.999966816	NIHY1_19_62__Klebsiella_oxytoca.contigs.fna	Infant 18 sink samples	Pittsburgh, PA	2012	infant sample	fecal

45	0.974341384	0.999989695	NIHY1_19_62__Klebsiella_oxytoca.contigs.fna	Infant 3 swab samples	Pittsburgh, PA	2012	infant fecal sample	
48	0.961381845	0.999985613	NIHY1_21_47__Klebsiella_oxytoca_rel.contigs.fna	Infant 5 early swab samples	Pittsburgh, PA	2012	infant fecal sample	
42	0.959261836	0.999987355	NIHY1_21_47__Klebsiella_oxytoca_rel.contigs.fna	Infant 3 sink samples	Pittsburgh, PA	2012	infant fecal sample	

5 Concluding remarks and future perspectives

The data and analyses presented here suggest the place an infant is born has major implications for its health and development. Through a series of experiments we show several ESKAPE related organisms are readily recovered from NICU surface samples before and after their detection in infant gut samples (Figures 1-5, 3-10, 3-11, 4-1, 4-2). Using multiple molecular techniques we show strong evidence that room microbes colonize infants, but also that infants are major contributors to shaping the microbiome of each NICU room (Figure 3-7). The ability of a patient to change the microbiome of the built environment has been previously reported, but most of these studies focus on nosocomial infection. For example, if the previous patient has a vancomycin resistant enterococcus (VRE), methicillin-resistant *Staphylococcus aureus*, *Clostridium difficile*, or *Acinetobacter baumannii* infection, subsequent patients in the same room have a 73% increase chance of acquiring this infection (Carling and Bartley, 2010). The approaches presented here offer an unbiased perspective of what microbes persist on NICU surfaces, what strains successfully colonize occupants, and what adaptations may contribute to this cycle of room to occupant exchange. However, assuming the ultimate application of the knowledge generated from hospital microbiome studies is curbing nosocomial infection and facilitating better occupant health, many questions remain unanswered.

Recent studies have highlighted that very few strains are shared amongst cohoused preterm infants in a NICU (Raveh-Sadka *et al.*, 2015). Most of the strains colonizing preterm infants appear to be sourced from a vast reservoir containing, at the very least hundreds, likely thousands of strain types (Raveh-Sadka *et al.*, 2016). Since very few strains are shared, two possible sources of colonization seem likely. One, the room may have little influence on infant colonization and microbes detected in the gut are sourced from the birthing process. This seems unlikely since recent studies using ddPCR to quantify biomass in infant fecal samples show the antibiotic treatments administered cause biomass to become undetectable (Raveh-Sadka *et al.*, 2015). Two, most microbes are sourced from the NICU environment but the large reservoir of strain types makes the probability of observing shared strains less likely. If this were true, with more infant fecal samples collected, more shared strains across infants would be observed.

While not explicitly addressed in Chapter 4, there is a trend emerging in the number of "persister" strains seen in this NICU over time as more samples are collected. For example, *K. oxytoca* was not identified as a persistent strain in a 2016 study (Raveh-Sadka *et al.*, 2016), but with the data presented here, it is very clearly a "persister." The same strain of *K. oxytoca* is recovered from many of the room metagenomes and several infant fecal samples. To fully characterize the level of strain diversity in the NICU, more room metagenomic samples are needed. A simple approach in designing a follow-up study would be to plot a collectors curve based on the data generated (*i.e.* number of samples on the *x*-axis and number of genomes recovered on the *y*-axis). Using such a curve as a guide, the campaign size could then be extrapolated from this preliminary data. As the cost of sequencing continues to decrease, exhaustively sampling a NICU to recover the entirety of strain-diversity may not be too cost prohibitive.

As the database of NICU strain diversity grows, so will understanding of the population structure of these NICU-adapted, ESKAPE related organisms. Recent studies detailing the expansion of the *K. pneumoniae* population provide an apt case study in what trends may emerge as researchers begin applying isolate-independent, metagenomic techniques in hospitals. The population structure of *K. pneumoniae* has been elucidated from a variety of molecular methods

(Bialek-Davenet *et al.*, 2014; Brisse *et al.*, 2009; Diancourt *et al.*, 2005). Popular over the past decade has been the use of multilocus sequence typing (MLST) in which seven chromosomally encoded housekeeping genes are targeted for sequencing (Brisse *et al.*, 2009; Diancourt *et al.*, 2005). These methods originally produced three phylogenetically distinct *K. pneumoniae* groupings (types I, II, and III), which were later classified as *K. pneumoniae*, *K. quasipneumoniae*, and *K. variicola* (Rosenblueth *et al.*, 2004; Brisse *et al.*, 2014; Holt *et al.*, 2015). Increased efforts in whole genome sequencing in recent years, however, has revealed a vast complexity within the *K. pneumoniae* population. Using 289 *K. pneumoniae* genomes in a whole-genome based analysis, a recent study found 157 distinct lineages (Holt *et al.*, 2015), 155 of which have been documented using MLST methods (Bialek-Davenet *et al.*, 2014). However, this campaign was far from an exhaustive survey of the *K. pneumoniae* population. A recent review plotted the number of *K. pneumoniae* lineages discovered versus the number of sequenced isolates to create a collector's curve. This curve indicates the number of possible *K. pneumoniae* lineages to be in the thousands (Wyres and Holt, 2016). The persistence of this many lineages has yet to be explained but may be due to *K. pneumoniae's* ability to occupy a wide variety of ecological niches (Ullmann, 1998; Bagley, 1985; Holt *et al.*, 2015).

Many of the *K. pneumoniae* and ESKAPE related studies are implemented using a targeted approach focused on disease or are based on isolation via culturing. The approaches presented here offer several advantages over these culture-dependent techniques, but the most important are whole-genome information and scalability. While all ESKAPE organisms are relatively easy to culture, using an isolate based approach is low throughput. Using a sample pooling scheme across NICU environments, we were able to recover dozens of *K. pneumoniae* and *K. oxytoca* genomes, which are not represented in the above mentioned Klebsiella strain type databases. Since sink basin samples are dominated by Enterobacteriaceae, a follow up study focusing on sink samples could greatly expand the Klebsiella radiation previously documented by low throughput methods. Another added benefit of metagenomics approaches is whole genome information and community context of genomes in the environment.

Whole genome information and community context could be leveraged to provide interventions to prevent nosocomial infection. One possible approach could be to introduce a standard metagenomic surveillance protocol of room surfaces to be included in the regular surveillance activities common in most hospitals. Typically infectious disease specialists use antibiotic resistance panels and enrichment media to monitor relative contamination levels on hospital surfaces (Sydnor and Perl, 2011). In conducting metagenomic surveillance, recovering the entire genome of hospital surface associated organisms would reveal the pathogenic potential of surface microbes, their antibiotic resistance potential, and provide information as to what features allow persistent surface colonization.

Several examples using metabolic potential to predict what features enable hospital surface colonization were highlighted here. A good example is biofilm formation, discussed in Chapter 1. Biofilms on hospital surfaces often have increased antibiotic resistance potential and are highly resistant to removal via cleaning (Weiss-Muszkat *et al.*, 2010; Romanova *et al.*, 2007; Hu *et al.*, 2015). Utilizing the genomic information from these communities, it may be possible to genetically engineer communities unable to form biofilms or engineer biofilms that are more susceptible to cleaning. With recent advances in gene editing technology, *e.g.* CRISPR-Cas9 (Ledford, 2016), introducing a genetically engineered probiotic consortia of hospital surface associated microbes is becoming increasingly more attainable. Another possibility could be the implementation of cleverly designed building materials that act as prebiotics to selectively enrich

for innocuous microbes to dominate hospital surfaces. Creating prebiotic building materials would be challenging for numerous reasons but essential first steps is generating more metagenomic, as well as transcriptomic, data to better understand the metabolic processes contributing to surface colonization.

Hospital acquired infections remain a significant problem in the US, costing approximately $30 billion per year to manage (CDC, 2016). The data presented here suggests not only are pathogens sourced from the hospital environment to colonize patients, but so are commensal organisms in the case of preterm infants. Using metagenomics techniques we show many strains from infant fecal samples are recoverable year after year and these strains are identical to strains found on NICU surfaces. Some of these strains are endemic while others are more globally dispersed. Future studies should continue to generate genome focused data using high throughput techniques. Perhaps unattainable years ago due to sequencing costs, scientists should aim to sequence everything in the hospital (*i.e.* surfaces, air, water, and people). The data generated will be invaluable information in better understanding organisms that have significant implications for human and building health.

6 References

Abbad-Andaloussi S, Durr C, Dürr C, Raval G, Petitdemange H. (1996). Carbon and electron flow in *Clostridium butyricum* grown in chemostat culture on glycerol and on glucose. *Microbiology* **142**: 1149–58.

Adams RI, Miletto M, Taylor JW, Bruns TD. (2013). Dispersal in microbes: fungi in indoor air are dominated by outdoor air and show dispersal limitation at short distances. *ISME J* **7**: 1262–73.

Adler A, Gottesman G, Dolfin T, Arnon S, Regev R, Bauer S, *et al.* (2005). *Bacillus* species sepsis in the neonatal intensive care unit. *J Infect* **51**: 390–5.

Albenberg L, Esipova T V, Judge CP, Bittinger K, Chen J, Laughlin A, *et al.* (2014). Correlation between intraluminal oxygen gradient and radial partitioning of intestinal microbiota in humans and mice. *Gastroenterology* **147**: 1–9.

Ardissone AN, de la Cruz DM, Davis-Richardson AG, Rechcigl KT, Li N, Drew JC, *et al.* (2014). Meconium microbiome analysis identifies bacteria correlated with premature birth. *PLoS One* **9**: e90784.

Arias CA, Murray BE. (2012). The rise of the *Enterococcus*: beyond vancomycin resistance. *Nat Rev Microbiol* **10**: 266–78.

Arrieta M-C, Stiemsma LT, Dimitriu PA, Thorson L, Russell S, Yurist-Doutsch S, *et al.* (2015). Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Sci Transl Med* **7**: 307ra152.

Bagley ST. (1985). Habitat association of *Klebsiella* species. *Infect Control* **6**: 52–58.

Barberan A, Dunn RR, Reich BJ, Pacifici K, Laber EB, Menninger HL, *et al.* (2015). The ecology of microscopic life in household dust. *Proc R Soc B* **282**: 20151139.

Barnhart MM, Chapman MR. (2006). Curli biogenesis and function. *Annu Rev Microbiol* **60**: 131–47.

Basler M, Ho BT, Mekalanos JJ. (2013). Tit-for-tat: type VI secretion system counterattack during bacterial cell-cell interactions. *Cell* **152**: 884–94.

Van Den Berg RWA, Claahsen HL, Niessen M, Muytjens HL, Liem K, Voss A. (2000). *Enterobacter cloacae* outbreak in the NICU related to disinfected thermometers. *J Hosp Infect* **45**: 29–34.

Bhangar S, Brooks B, Firek B, Licina D, Tang X, Morowitz MJ, *et al.* (2016). Pilot study of sources and concentrations of size-resolved airborne particles in a neonatal intensive care unit. *Build Environ* **106**: 10–19.

Bialek-Davenet S, Criscuolo A, Ailloud F, Passet V, Jones L, Delannoy-Vieillard AS, *et al.* (2014). Genomic definition of hypervirulent and multidrug-resistant *Klebsiella pneumoniae* clonal groups. *Emerg Infect Dis* **20**: 1812–1820.

Bizzarro MJ, Ehrenkranz RA, Gallagher PG. (2014). Concurrent bloodstream infections in infants with necrotizing enterocolitis. *J Pediatr* **164**: 61–6.

Bokulich NA, Mills DA, Underwood M a. (2013). Surface microbes in the neonatal intensive care unit: Changes with routine cleaning and over time. *J Clin Microbiol* **51**: 2617–24.

Bonora M, Ligozzi M, Fatima M De. (2004). Vancomycin-resistant *Enterococcus faecium* isolates causing hospital outbreaks in northern Italy belong to the multilocus sequence typing C1 lineage. *Microb Drug* **10**: 114–123.

Bradley CR, Fraise AP. (1996). Heat and chemical resistance of enterococci. *J Hosp Infect* **34**: 191–6.

Brisse S, Fevre C, Passet V, Issenhuth-Jeanjean S, Tournebize RR, Diancourt L, *et al.* (2009). Virulent clones of *Klebsiella pneumoniae*: identification and evolutionary scenario based on genomic and phenotypic characterization. *PLoS One* **4**: e4982.

Brisse S, Passet V, Grimont PAD. (2014). Description of *Klebsiella quasipneumoniae* sp. nov., isolated from human infections, with two subspecies, *Klebsiella quasipneumoniae* subsp. *quasipneumoniae* subsp. nov. and *Klebsiella quasipneumoniae*. *Int J Syst Evol Microbiol* **64**: 3146–52.

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.

Brooks B, Mueller RS, Young JC, Morowitz MJ, Hettich RL, Banfield JF. (2015). Strain-resolved microbial community proteomics reveals simultaneous aerobic and anaerobic function during gastrointestinal tract colonization of a preterm infant. *Front Microbiol* **6**: 654.

Buffet-Bataillon S, Branger B, Cormier M, Bonnaure-Mallet M, Jolivet-Gougeon A. (2011). Effect of higher minimum inhibitory concentrations of quaternary ammonium compounds in clinical *E. coli* isolates on antibiotic susceptibilities and clinical outcomes. *J Hosp Infect* **79**: 141–6.

Buffet-Bataillon S, Rabier V, Bétrémieux P, Beuchée A, Bauer M, Pladys P, *et al.* (2009). Outbreak of *Serratia marcescens* in a neonatal intensive care unit: contaminated unmedicated liquid soap and risk factors. *J Hosp Infect* **72**: 17–22.

Buffet-Bataillon S, Tattevin P, Bonnaure-Mallet M, Jolivet-Gougeon A. (2012). Emergence of resistance to antibacterial agents: the role of quaternary ammonium compounds--a critical review. *Int J Antimicrob Agents* **39**: 381–9.

Cahenzli J, Köller Y, Wyss M, Geuking MB, McCoy KD. (2013). Intestinal microbial diversity during early-life colonization shapes long-term IgE levels. *Cell Host Microbe* **14**: 559–570.

Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R. (2010a). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26**: 266–7.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, *et al.* (2010b). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**: 335–6.

Carling PC, Bartley JM. (2010). Evaluating hygienic cleaning in health care settings: What you do not know can harm your patients. *Am J Infect Control* **38**: S41–S50.

CDC. (2016). Healthcare-associated Infections. *Heal Infect*. http://www.cdc.gov/winnablebattles/healthcareassociatedinfections/ (Accessed November 26, 2016).

Chase J, Fouquier J, Zare M, Sonderegger DL, Knight R, Kelley ST, *et al.* (2016). Geography and location are the primary drivers of office microbiome composition. *mSystems* **1**: e00022-16.

Checinska A, Probst AJ, Vaishampayan P, White JR, Kumar D, Stepanov VG, *et al.* (2015). Microbiomes of the dust particles collected from the International Space Station and spacecraft assembly facilities. *Microbiome* **3**: 50.

Condell O, Iversen C, Cooney S, Power K a, Walsh C, Burgess C, *et al.* (2012). Efficacy of biocides used in the modern food industry to control *Salmonella enterica*, and links

between biocide tolerance and resistance to clinically relevant antimicrobial compounds. *Appl Environ Microbiol* **78**: 3087–97.
Costello EK, Stagaman K, Dethlefsen L, Bohannan BJM, Relman DA. (2012). The application of ecological theory toward an understanding of the human microbiome. *Science* **336**: 1255–62.
David LA, Materna AC, Friedman J, Campos-Baptista MI, Blackburn MC, Perrotta A, *et al.* (2014). Host lifestyle affects human microbiota on daily timescales. *Genome Biol* **15**: R89.
DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, *et al.* (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**: 5069–72.
Diancourt L, Passet V, Verhoef J, Patrick a D, Grimont P a D, Brisse S. (2005). Multilocus sequence typing of *Klebsiella pneumoniae* nosocomial isolates. *J Clin Microbiol* **43**: 4178–82.
Diaz Heijtz R, Wang S, Anuar F, Qian Y, Björkholm B, Samuelsson A, *et al.* (2011). Normal gut microbiota modulates brain development and behavior. *Proc Natl Acad Sci U S A* **108**: 3047–52.
Dick GJ, Anantharaman K, Baker BJ, Li M, Reed DC, Sheik CS. (2013). The microbiology of deep-sea hydrothermal vent plumes: ecological and biogeographic linkages to seafloor and water column habitats. *Front Microbiol* **4**: 124.
Dominguez-Bello MG, Costello EK, Contreras M, Magris M, Hidalgo G, Fierer N, *et al.* (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc Natl Acad Sci U S A* **107**: 11971–5.
La Duc MT, Dekas A, Osman S, Moissl C, Newcombe D, Venkateswaran K. (2007). Isolation and characterization of bacteria capable of tolerating the extreme conditions of clean room environments. *Appl Environ Microbiol* **73**: 2600–11.
La Duc MT, Vaishampayan P, Nilsson HR, Torok T, Venkateswaran K. (2012). Pyrosequencing-derived bacterial, archaeal, and fungal diversity of spacecraft hardware destined for Mars. *Appl Environ Microbiol* **78**: 5912–22.
Edgar RC. (2004a). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 113.
Edgar RC. (2004b). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–7.
Edgar RC. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–1.
Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**: 2194–200.
Ellegaard K, Engel P. (2016). Beyond 16S rRNA community profiling: intra-species diversity in the gut microbiota. *Front Microbiol* **7**: 1475.
Eppley JM, Tyson GW, Getz WM, Banfield JF. (2007). Strainer: software for analysis of population variation in community genomic datasets. *BMC Bioinformatics* **8**: 398.
Erickson AR, Cantarel BL, Lamendella R, Darzi Y, Mongodin EF, Pan C, *et al.* (2012). Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PLoS One* **7**: e49138.
Fadrosh DW, Ma B, Gajer P, Sengamalay N, Ott S, Brotman RM, *et al.* (2014). An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina

MiSeq platform. *Microbiome* **2**: 6.
Fairchild CI, Tillery MI. (1982). Wind tunnel measurements of the resuspension of ideal particles. *Atmos Environ* **16**: 229–38.
Fernández L, Hancock REW. (2012). Adaptive and mutational resistance: role of porins and efflux pumps in drug resistance. *Clin Microbiol Rev* **25**: 661–81.
Fierer N, Lauber CL, Zhou N, Mcdonald D, Costello EK, Knight R. (2010). Forensic identification using skin bacterial communities. *Proc Natl Acad Sci U S A* **107**: 6477–81.
Franzosa EA, Morgan XC, Segata N, Waldron L, Reyes J, Earl AM, *et al.* (2014). Relating the metatranscriptome and metagenome of the human gut. *Proc Natl Acad Sci U S A* **111**: E2329-38.
Gasparrini AJ, Crofts TS, Gibson MK, Tarr PI, Warner BB, Dantas G. (2016). Antibiotic perturbation of the preterm infant gut microbiome and resistome. *Gut Microbes* **7**: 443–9.
Gastmeier P, Loui A, Stamm-Balderjahn S, Hansen S, Zuschneid I, Sohr D, *et al.* (2007). Outbreaks in neonatal intensive care units - they are not like others. *Am J Infect Control* **35**: 172–6.
Gibson MK, Wang B, Ahmadi S, Burnham C-AD, Tarr PI, Warner BB, *et al.* (2016). Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome. *Nat Microbiol* **1**: 16024.
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.
Guaraldi F, Salvatori G. (2012). Effect of breast and formula feeding on gut microbiota shaping in newborns. *Front Cell Infect Microbiol* **2**: 94.
Hamady M, Lozupone C, Knight R. (2010). Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J* **4**: 17–27.
Hamilton BE, Martin JA, Osterman MJKS. (2015). Births: Preliminary data for 2015 National Vital Statistics reports. *Natl Vital Stat Reports* **65**: 1–15.
Hewitt KM, Mannino FL, Gonzalez A, Chase JH, Caporaso JG, Knight R, *et al.* (2013). Bacterial diversity in two neonatal intensive care units (NICUs). *PLoS One* **8**: e54703.
Hildebrand F, Tadeo R, Voigt AY, Bork P, Raes J. (2014). LotuS: an efficient and user-friendly OTU processing pipeline. *Microbiome* **2**: 1–7.
Hoffman LR, D'Argenio DA, MacCoss MJ, Zhang Z, Jones RA, Miller SI. (2005). Aminoglycoside antibiotics induce bacterial biofilm formation. *Nature* **436**: 1171–5.
Hold GL. (2014). Role of the gut microbiota in inflammatory bowel disease pathogenesis: What have we learnt in the past 10 years? *World J Gastroenterol* **20**: 1192.
Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse C a., Dance D, *et al.* (2015). Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. *Proc Natl Acad Sci U S A* **112**: E3574-81.
Hsiao EY, Mcbride SW, Hsien S, Sharon G, Hyde ER, Mccue T, *et al.* (2013). Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. *Cell* **155**: 1451–63.
Hu H, Johani K, Gosbell IB, Jacombs ASW, Almatroudi A, Whiteley GS, *et al.* (2015). Intensive care unit environmental surfaces are contaminated by multidrug-resistant bacteria in biofilms: Combined results of conventional culture, pyrosequencing, scanning electron microscopy, and confocal laser microscopy. *J Hosp Infect* **91**: 35–44.

Huffnagle GB. (2010). The microbiota and allergies/asthma. *PLoS Pathog* **6**: e1000549.
Hug LA, Castelle CJ, Wrighton KC, Thomas BC, Sharon I, Frischkorn R, *et al.* (2013). Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. *Microbiome* **1**: 22.
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.
Jani AJ, Cotter PA. (2010). Type VI Secretion: Not just for pathogenesis anymore. *Cell Host Microbe* **8**: 2–6.
Jovel J, Patterson J, Wang W, Hotte N, O'Keefe S, Mitchel T, *et al.* (2016). Characterization of the gut microbiome using 16S or shotgun metagenomics. *Front Microbiol* **7**: 1–17.
Juge N. (2012). Microbial adhesins to gastrointestinal mucus. *Trends Microbiol* **20**: 30–9.
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**: 199–205.
Kelley ST, Gilbert JA. (2013). Studying the microbiology of the indoor environment. *Genome Biol* **14**: 202.
Kembel SW, Jones E, Kline J, Northcutt D, Stenson J, Womack AM, *et al.* (2012). Architectural design influences the diversity and structure of the built environment microbiome. *ISME J* **6**: 1469–79.
Kembel SW, Meadow JF, O'Connor TK, Mhuireach G, Northcutt D, Kline J, *et al.* (2014). Architectural design drives the biogeography of indoor bacterial communities. *PLoS One* **9**: e87093.
Kerfeld CA, Erbilgin O. (2014). Bacterial microcompartments and the modular construction of microbial metabolism. *Trends Microbiol* **23**: 22–34.
Kleanthous C. (2010). Swimming against the tide: progress and challenges in our understanding of colicin translocation. *Nat Rev Microbiol* **8**: 843–8.
Klepeis NE, Nelson WC, Ott WR, Robinson JP, Tsang AM, Switzer P, *et al.* (2001). The National Human Activity Pattern Survey (NHAPS): A resource for assessing exposure to environmental pollutants. *J Expo Anal Environ Epidemiol* **11**: 231–52.
Knights D, Kuczynski J, Charlson E, Zaneveld J, Mozer MC, Collman RG, *et al.* (2011). Bayesian community-wide culture-independent microbial source tracking. *Nat Methods* **8**: 761–3.
Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight R, *et al.* (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proc Natl Acad Sci U S A* **108**: 4578–85.
Kong HH, Oh J, Deming C, Conlan S, Grice EA, Beatson MA, *et al.* (2012). Temporal shifts in the skin microbiome associated with atopic dermatitis disease flares and treatment in children with atopic dermatitis. *Genome Res* **22**: 850–9.
Lagesen K, Hallin P, Rødland EA, Staerfeldt H-H, Rognes T, Ussery DW, *et al.* (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* **35**: 3100–8.
Langmead B, Salzberg SL. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–9.
Lax S, Smith DP, Hampton-Marcell J, Owens SM, Handley KM, Scott NM, *et al.* (2014). Longitudinal analysis of microbial interaction between humans and the indoor

environment. *Science* **345**: 1048–52.
Lazarevic V, Gaïa N, Girard M, Schrenzel J. (2016). Decontamination of 16S rRNA gene amplicon sequence datasets based on bacterial load assessment by qPCR. *BMC Microbiol* **16**: 73.
Ledford H. (2016). Riding the CRISPR wave. *Nature* **531**: 156–9.
Licina D, Bhangar S, Brooks B, Baker R, Firek B, Tang X, *et al.* (2016). Concentrations and sources of airborne particles in a neonatal intensive care unit. *PLoS One* **11**: e0154991.
Lindsley WG, Blachere FM, Thewlis RE, Vishnu A, Davis KA, Cao G, *et al.* (2010). Measurements of airborne influenza virus in aerosol particles from human coughs. *PLoS One* **5**: e15100.
Louis P, Duncan SHSSH, Mccrae SSI, Jackson MS, Flint HJ, Millar J. (2004). Restricted distribution of the butyrate kinase pathway among butyrate-producing bacteria from the human colon. *J Bacteriol* **186**: 2099–2106.
Lowe TM, Eddy SR. (1996). TRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**: 955–64.
Luangasanatip N, Hongsuwan M, Limmathurotsakul D, Lubell Y, Lee AS, Harbarth S, *et al.* (2015). Comparative efficacy of interventions to promote hand hygiene in hospital: systematic review and network meta-analysis. *BMJ* **351**: h3728.
Luo C, Walk ST, Gordon DM, Feldgarden M, Tiedje JM, Konstantinidis KT. (2011). Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *Proc Natl Acad Sci U S A* **108**: 7200–5.
Luoma M, Batterman SA. (2001). Characterization of particulate emissions from occupant activities in offices. *Indoor Air* **11**: 35–48.
Mackelprang R, Waldrop MP, DeAngelis KM, David MM, Chavarria KL, Blazewicz SJ, *et al.* (2011). Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature* **480**: 368–71.
Madan JC, Salari RC, Saxena D, Davidson L, Toole GAO, Moore JH, *et al.* (2012). Gut microbial colonisation in premature neonates predicts neonatal sepsis. *Arch Dis Child Fetal Neonatal Ed* **97**: F456-62.
Mahnert A, Vaishampayan P, Probst AJ, Auerbach A, Moissl-Eichinger C, Venkateswaran K, *et al.* (2015). Cleanroom maintenance significantly reduces abundance but not diversity of indoor microbiomes. *PLoS One* **10**: e0134848.
Manson JM, Hancock LE, Gilmore MS. (2010). Mechanism of chromosomal transfer of *Enterococcus faecalis* pathogenicity island, capsule, antimicrobial resistance, and other traits. *Proc Natl Acad Sci U S A* **107**: 12269–74.
Maurice CF, Haiser HJ, Turnbaugh PJ. (2013). Xenobiotics shape the physiology and gene expression of the active human gut microbiome. *Cell* **152**: 39–50.
McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, *et al.* (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* **6**: 610–8.
Meadow JF, Altrichter AE, Bateman AC, Stenson J, Brown G, Green JL, *et al.* (2015). Humans differ in their personal microbial cloud. *PeerJ* **3**: e1258.
Meadow JF, Altrichter AE, Kembel SW, Moriyama M, O'Connor TK, Womack AM, *et al.* (2014). Bacterial communities on classroom surfaces vary with human contact. *Microbiome* **2**: 7.
Miller CS, Baker BJ, Thomas BC, Singer SW, Banfield JF. (2011). EMIRGE: reconstruction of

full-length ribosomal genes from microbial community short read sequencing data. *Genome Biol* **12**: R44.
Miller CS, Handley KM, Wrighton KC, Frischkorn KR, Thomas BC, Banfield JF. (2013). Short-read assembly of full-length 16S amplicons reveals bacterial diversity in subsurface sediments. *PLoS One* **8**: e56018.
Morowitz MJ, Denef VJ, Costello EK, Thomas BC, Poroyko V, Relman DA, *et al.* (2010a). Strain-resolved community genomic analysis of gut microbial colonization in a premature infant. *Proc Natl Acad Sci U S A* **108**: 1128–33.
Morowitz MJ, Poroyko V, Caplan M, Alverdy J, Liu DC. (2010b). Redefining the role of intestinal microbes in the pathogenesis of necrotizing enterocolitis. *Pediatrics* **125**: 777–85.
Mshvildadze M, Neu J, Shuster J, Theriaque D, Li N, Mai V. (2010). Intestinal microbial ecology in premature infants assessed with non-culture-based techniques. *J Pediatr* **156**: 20–5.
Murray BE. (1990). The life and times of the Enterococcus. *Clin Microbiol Rev* **3**: 46–65.
Naesens R, Jeurissen A, Vandeputte C, Cossey V, Schuermans A. (2009). Washing toys in a neonatal intensive care unit decreases bacterial load of potential pathogens. *J Hosp Infect* **71**: 197–8.
Joshi NA, Fass JN. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at https://github.com/najoshi/sickle.
Nawrocki EP, Eddy SR. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**: 2933–5.
Nistal E, Caminero A, Herrán AR, Arias L, Vivas S, de Morales JMR, *et al.* (2012). Differences of small intestinal bacteria populations in adults and children with/without celiac disease: effect of age, gluten diet, and disease. *Inflamm Bowel Dis* **18**: 649–56.
Oberauner L, Zachow C, Lackner S, Högenauer C, Smolle K-H, Berg G. (2013). The ignored diversity: complex bacterial communities in intensive care units revealed by 16S pyrosequencing. *Sci Rep* **3**: 1413.
Olm MR, Brown CT, Brooks B, Firek BA, Baker R, Soenjoyo K, *et al.* (2016). Identical bacterial populations colonize premature infant gut, skin, and oral microbiomes and exhibit different in situ growth rates. *Genome Res* (Accepted).
Ondov BD, Treangen TJ, Mallonee AB, Bergman NH, Koren S, Phillippy AM. (2015). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* 29827.
Ong SH, Kukkillaya VU, Wilm A, Lay C, Ho EXP, Low L, *et al.* (2013). Species identification and profiling of complex microbial communities using shotgun Illumina sequencing of 16S rRNA amplicon sequences. *PLoS One* **8**: e60811.
Palmer C, Bik EM, DiGiulio DB, Relman DA, Brown PO. (2007). Development of the human infant intestinal microbiota. *PLoS Biol* **5**: 1556–73.
Park AJ, Collins J, Blennerhassett PA, Ghia JE, Verdu EF, Bercik P, *et al.* (2013). Altered colonic function and microbiota profile in a mouse model of chronic depression. *Neurogastroenterol Motil* **25**: 733-e575.
Penders J, Thijs C, Vink C, Stelma FF, Snijders B, Kummeling I, *et al.* (2006). Factors influencing the composition of the intestinal microbiota in early infancy. *Pediatrics* **118**: 511–21.
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**:

1420–8.
Perkins SD, Mayfield J, Fraser V, Angenent LT. (2009). Potentially pathogenic bacteria in shower water and air of a stem cell transplant unit. *Appl Environ Microbiol* **75**: 5363–72.
Poza M, Gayoso C, Gómez MJ, Rumbo-Feal S, Tomás M, Aranda J, *et al.* (2012). Exploring bacterial diversity in hospital environments by GS-FLX Titanium pyrosequencing. *PLoS One* **7**: e44105.
Price MN, Dehal PS, Arkin AP. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**: e9490.
Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, *et al.* (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**: 7188–96.
Qian J, Hospodsky D, Yamamoto N, Nazaroff WW, Peccia J. (2012). Size-resolved emission rates of airborne bacteria and fungi in an occupied classroom. *Indoor Air* **22**: 339–51.
Raveh-Sadka T, Firek B, Sharon I, Baker R, Brown CT, Thomas BC, *et al.* (2016). Evidence for persistent and shared bacterial strains against a background of largely unique gut colonization in hospitalized premature infants. *ISME J* **10**: 2817–30.
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* **2015**: 1–25.
Di Rienzi SC, Sharon I, Wrighton KC, Koren O, Hug LA, Thomas BC, *et al.* (2013). The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. *Elife* **2**: e01102.
Rintala H, Pitkäranta M, Toivola M, Paulin L, Nevalainen A. (2008). Diversity and seasonal dynamics of bacterial community in indoor environment. *BMC Microbiol* **8**: 56.
Romanova NA, Gawande P V, Brovko LY, Griffiths MW. (2007). Rapid methods to assess sanitizing efficacy of benzalkonium chloride to *Listeria monocytogenes* biofilms. *J Microbiol Methods* **71**: 231–7.
Rosenblueth M, Martínez L, Silva J, Martínez-Romero E. (2004). *Klebsiella variicola*, a novel species with clinical and plant-associated isolates. *Syst Appl Microbiol* **27**: 27–35.
Ruiz-Calderon JF, Cavallin H, Song SJ, Novoselac A, Pericchi LR, Hernandez JN, *et al.* (2016). Walls talk: Microbial biogeography of homes spanning urbanization. *Sci Adv* **2**: e1501061.
Sharon 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**: 111–20.
Shin H, Pei Z, Martinez KA, Rivera-Vinas JI, Mendez K, Cavallin H, *et al.* (2015). The first microbial environment of infants born by C-section: the operating room microbes. *Microbiome* **3**: 59.
Sim K, Powell E, Shaw AG, McClure Z, Bangham M, Kroll JS. (2013). The neonatal gastrointestinal microbiota: the foundation of future health? *Arch Dis Child - Fetal Neonatal Ed* **98**: F362–F364.
Singh N, Campbell J, Short B. (2005). Control of vancomycin-resistant enterococci in the neonatal intensive care unit. *Infect Control* **26**: 646–9.
Smith MI, Yatsunenko T, Manary MJ, Trehan I, Mkakosya R, Cheng J, *et al.* (2013). Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. *Science* **339**: 548–54.
Smith PA. (2014). Is your body mostly microbes? Actually, we have no idea. *Boston Globe* 1.

Song SJ, Dominguez-Bello MG, Knight R. (2013). How delivery mode and feeding can shape the bacterial community in the infant gut. *Can Med Assoc J* **185**: 373–4.
Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, Rubin EM. (2007). Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* **318**: 1449–52.
Stackebrandt E, Goodfellow M. (1991). 16S/23S rRNA sequencing. In: *Nucleic acid techniques in bacterial systematics*. John Wiley & Son Ltd: Chichester, United Kingdom, pp 115–175.
Stoll BJ, Hansen NI, Bell EF, Shankaran S, Laptook AR, Walsh MC, *et al.* (2010). Neonatal outcomes of extremely preterm infants from the NICHD Neonatal Research Network. *Pediatrics* **126**: 443–56.
Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. (2007). UniRef: Comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**: 1282–8.
Sydnor ERM, Perl TM. (2011). Hospital epidemiology and infection control in acute-care settings. *Clin Microbiol Rev* **24**: 141–73.
Tan J, McKenzie C, Potamitis M, Thorburn AN, Mackay CR, Macia L. (2014). The role of short-chain fatty acids in health and disease. *Adv Immunol* **121**: 91–119.
Tenaillon O, Skurnik D, Picard B, Denamur E. (2010). The population genetics of commensal *Escherichia coli*. *Nat Rev Microbiol* **8**: 207–17.
Touati A, Achour W, Cherif A, Hmida HH Ben, Afif FB, Jabnoun S, *et al.* (2009). Outbreak of *Acinetobacter baumannii* in a neonatal intensive care unit: Antimicrobial susceptibility and genotyping analysis. *Ann Epidemiol* **19**: 372–8.
Tringe SG, Zhang T, Liu X, Yu Y, Lee WH, Yap J, *et al.* (2008). The airborne metagenome in an indoor urban environment. *PLoS One* **3**: e1862.
Trosvik P, Stenseth NC, Rudi K. (2010). Convergent temporal dynamics of the human infant gut microbiota. *ISME J* **4**: 151–8.
Tu Q, He Z, Zhou J. (2014). Strain/species identification in metagenomes using genome-specific markers. *Nucleic Acids Res* **42**: 1–12.
Turnbaugh PJ, Ley RE, Mahowald M a, Magrini V, Mardis ER, Gordon JI. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**: 1027–31.
Ullmann U. (1998). *Klebsiella* spp. as nosocomial pathogens : Epidemiology, taxonomy, typing methods, and pathogenicity factors. *Clin Microbiol Rev* **11**: 589–603.
Ultsch A, Mörchen F. (2005). ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM. Marburg.
van Veen HW. (2010). Structural biology: Last of the multidrug transporters. *Nature* **467**: 926–7.
Wang Q, Garrity GM, Tiedje JM, Cole JR. (2007). Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**: 5261–7.
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**: 944–54.
Waters JM, Fraser CI, Hewitt GM. (2013). Founder takes all: Density-dependent processes structure biodiversity. *Trends Ecol Evol* **28**: 78–85.
Watkins DJ, Besner GE. (2013). The role of the intestinal microcirculation in necrotizing enterocolitis. *Semin Pediatr Surg* **22**: 83–7.
Weinmaier T, Probst AJ, La Duc MT, Ciobanu D, Cheng J-F, Ivanova N, *et al.* (2015). A

viability-linked metagenomic analysis of cleanroom environments: eukarya, prokaryotes, and viruses. *Microbiome* **3**: 62.
Weiss-Muszkat M, Shakh D, Zhou Y, Pinto R, Belausov E, Chapman MR, *et al.* (2010). Biofilm formation by and multicellular behavior of *Escherichia coli* O55:H7, an atypical enteropathogenic strain. *Appl Environ Microbiol* **76**: 1545–54.
Winter SE, Lopez CA, Bäumler AJ. (2013). The dynamics of gut-associated microbial communities during inflammation. *EMBO Rep* **14**: 319–27.
Wright ES, Yilmaz LS, Noguera DR. (2012). DECIPHER, a search-based approach to chimera identification for 16S rRNA sequences. *Appl Environ Microbiol* **78**: 717–25.
Wu M, Eisen JA. (2008). A simple, fast, and accurate method of phylogenomic inference. *Genome Biol* **9**: R151.
Wyres KL, Holt KE. (2016). *Klebsiella pneumoniae* population genomics and antimicrobial-resistant clones. *Trends Microbiol* **In Press**: 1–13.
Yamamoto N, Shendell DG, Peccia J. (2011). Assessing allergenic fungi in house dust by floor wipe sampling and quantitative PCR. *Indoor Air* **21**: 521–30.
Young JC, Pan C, Adams RM, Brooks B, Banfield JF, Morowitz MJ, *et al.* (2015). Metaproteomics reveals functional shifts in microbial and human proteins during a preterm infant gut colonization case. *Proteomics* **15**: 3463–73.
Yuan C, Lei J, Cole J, Sun Y. (2015). Reconstructing 16S rRNA genes in metagenomic data. *Bioinformatics* **31**: i35–i43.
Zerbino DR, Birney E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**: 821–9.