# Hospitalized premature infants are colonized by related bacterial strains with distinct proteomic profiles

Christopher. T. Brown[1], Weili Xiong[2], Matthew R. Olm[1], Brian C. Thomas[3], Robyn Baker[4], Brian Firek[5], Michael J. Morowitz[5,6], Robert L. Hettich[7], and Jillian F. Banfield[3,8,9*]

[1]Department of Plant and Microbial Biology, University of California, Berkeley, CA, USA,
[2]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA,
[3]Department of Earth and Planetary Science, University of California, Berkeley, CA, USA,
[4]Division of Newborn Medicine, Children's Hospital of Pittsburgh and Magee-Womens Hospital of UPMC, Pittsburgh, United States
[5]Department of Surgery, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA,
[6]Division of Pediatric General and Thoracic Surgery, Children's Hospital of Pittsburgh of UPMC, Pittsburgh, PA, USA
[7]Chemical Science Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
[8]Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA, USA
[9]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

*Corresponding author
Email: jbanfield@berkeley.edu
Telephone: 510-643-2155
Address: McCone Hall, Berkeley, CA 94720

## Running Title

Colonization of the infant gut microbiome

## Keywords

Metagenomics, Metaproteomics, Microbial Genomics, iRep, Human Microbiome, Necrotizing Enterocolitis, Neonates, Microbial Colonization

## Abstract

During the first weeks of life, microbial colonization of the gut impacts human immune system maturation and other developmental processes. In premature infants, aberrant colonization has been implicated in the onset of necrotizing enterocolitis (NEC), a life-threatening intestinal disease. To study variability in the premature infant gut colonization process, genome-resolved metagenomics was conducted on 343 fecal samples collected during the first three months of life from 35 premature infants, 14 of which developed NEC, and metaproteomic measurements were made on 87 samples. Microbial proteomic profiles and community composition remained relatively stable on the time scale of a week, but the proteome was more variable than community composition. Although genetically similar organisms colonized many infants, most infants were colonized by distinct strains with metabolic profiles that could be distinguished using metaproteomics. Microbiome composition correlated with infant, antibiotics administration, and NEC diagnosis. Communities were found to cluster into seven primary types, and community type switched within infants, sometimes multiple times. Interestingly, some communities sampled from the same infant at subsequent time points clustered with those of other infants. In some cases, switches preceded onset of NEC; however, no species or community type could account for NEC across the majority of infants. In addition to a correlation of protein abundances with organism replication rates, we find that organism proteomes correlate with overall community composition. Thus, this genome-resolved proteomics study demonstrates that the contributions of individual organisms to microbiome development depend on microbial community context.

## Introduction

Infants have been characterized as having high levels of between-individual variation in microbiome composition compared with adult humans (Costello et al. 2009; Palmer et al. 2007). Variation in the infant microbiome exists at both the species and strain level (Raveh-Sadka et al. 2015; 2016). During the first one to two years of life the gut microbiomes of infants begin to converge upon an adult-like state (Palmer et al. 2007; Bokulich et al. 2016). However, aberrations in this process may contribute to diseases such as type 1 and 2 diabetes, irritable bowel disease, and necrotizing enterocolitis (NEC) in premature infants (Xavier and Podolsky 2007; Brown et al. 2011; Qin et al. 2012; Mshvildadze et al. 2010; Mai et al. 2011; Morrow et al. 2013). Because establishment of the microbiome is a key driver of immune system development, changes in the process of colonization may have life-long implications, even if they do not result in a drastically different microbiome composition later in life (Maslowski et al. 2009; Lathrop et al. 2011).

Infants born prematurely have low-diversity microbial communities compared with full term infants, and are susceptible to life-threatening diseases such as NEC (Neu and Walker 2011; Sharon et al. 2012; Brown et al. 2013; Raveh-Sadka et al. 2016; Pammi et al. 2017). While it has long been thought that bacterial infection may contribute to NEC pathogenesis, strain-resolved microbial community analysis has not identified a single pathogen that is responsible for the disease (Raveh-Sadka et al. 2015). However, it is still

likely that microbial communities play an important role, with the context-dependent metabolism of specific strains potentially critical to infant health and disease. Recent studies have applied proteomics and metabolomics to premature infant gut microbiomes to measure functional profiles in healthy premature infants and those that went on to develop NEC (Xiong et al. 2017; Stewart et al. 2016). These studies reported temporal variation in the infant proteome and identified metabolites associated with NEC. However, further study is required to better understand the range of functional and developmental patterns during the microbial colonization process.

To investigate microbial community assembly, and how microbes modulate their metabolisms and replication rates during the colonization process, we conducted a combined metagenomics and metaproteomics study of the microbiome of both healthy premature infants and infants that went on to develop NEC. Microbiome samples were collected during the first three months of life with the goal of measuring the physiological changes of dominant and ubiquitous bacterial species. Genomes assembled from metagenomes enabled analysis of microbial community membership and tracking of community composition and replication rates over time. The availability of genome sequences made it possible to map protein abundance measurements to bacterial species and strains. Microbial communities were clustered into distinct types in order to provide context for proteomics analyses. Statistical analyses showed that, while overall community composition was correlated with species and strain-specific proteomic profiles, the proteomes of members of the same species and strain were largely infant-specific. These analyses also show that bacterial proteome features are correlated with infant development, health status, and antibiotics administration.

## Results

### *Metagenome sequencing and genome binning*

In order to study the developing gut microbiome, stool samples were collected during the first three months of life for 35 infants born prematurely. Two of the infants in the study cohort developed sepsis (N1_017 and N1_019) and 14 infants developed necrotizing enterocolitis (NEC; **Table 1**). To study the gut microbiome, we analyzed 1,149 Gbp of DNA sequences generated by our laboratory (Raveh-Sadka et al. 2015; 2016; Brooks et al.). These sequences were from 343 metagenomes (average of 3.3 Gbp sequencing per sample; **Supplemental Figure 1 and Supplemental Table 1**). Metagenomes were assembled into 6.79 Gbp of scaffolds ≥1 Kbp that represented 92% of all sequenced DNA.

**Table 1 | Infant medical information.**

| infant | campaign | sex | delivery | mult. gest. | gestational age (weeks) | birth weight (g) | feeding | condition | NEC diagnosis (DOL) | infection (DOL) |
|---|---|---|---|---|---|---|---|---|---|---|
| N1_003 | NIH1 | F | C-section | Single | 26 | 822 | Breast | control | n/a | n/a |
| N1_004 | NIH1 | F | C-section | N1_005 | 32 | 1450 | Formula | control | n/a | n/a |
| N1_008 | NIH1 | F | Vaginal | Single | 32 | 1230 | Formula | NEC | 9 | n/a |
| N1_009 | NIH1 | M | C-section | Single | 29 | 1820 | Combination | control | n/a | n/a |
| N1_011 | NIH1 | M | C-section | N1_012 | 26 | 523 | Combination | NEC | 34, 62 | n/a |
| N1_014 | NIH1 | M | Vaginal | Single | 32 | 2035 | Combination | control | n/a | n/a |
| N1_017 | NIH1 | F | Vaginal | Single | 26 | 748 | Combination | NEC | 11 | 11 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| N1_018 | NIH1 | M | C-section | Single | 29 | 1133 | Combination | control | n/a | n/a |
| N1_019 | NIH1 | F | C-section | N1_020, N1_021 | 24 | 731 | Combination | control | n/a | 23, 35 |
| N1_021 | NIH1 | F | C-section | N1_019, N1_020 | 24 | 697 | Breast | NEC | 32 | 24 |
| N1_023 | NIH1 | F | Vaginal | Single | 27 | 875 | Breast | control | n/a | n/a |
| N2_031 | NIH2 | M | C-section | Single | 26 | 773 | Formula | control | n/a | 39, 57 |
| N2_035 | NIH2 | M | Vaginal | Single | 25 | 795 | Breast | control | n/a | n/a |
| N2_038 | NIH2 | F | C-section | N2_039 | 30 | 1381 | Combination | control | n/a | n/a |
| N2_039 | NIH2 | F | C-section | N2_038 | 30 | 1470 | Combination | NEC | 24 | n/a |
| N2_060 | NIH2 | M | C-section | Single | 30 | 1878 | Combination | control | n/a | n/a |
| N2_061 | NIH2 | M | Vaginal | Single | 28 | 1184 | Combination | NEC | 9, 34 | 4 |
| N2_064 | NIH2 | M | Vaginal | Single | 28 | 1100 | Combination | control | n/a | n/a |
| N2_065 | NIH2 | F | Vaginal | Single | 25 | 841 | Combination | control | n/a | n/a |
| N2_066 | NIH2 | F | Vaginal | Single | 28 | 1028 | Breast | control | n/a | n/a |
| N2_069 | NIH2 | M | C-section | N2_070 | 26 | 637 | Breast | NEC | 32 | n/a |
| N2_070 | NIH2 | F | C-section | N2_069 | 26 | 633 | Combination | control | n/a | 7 |
| N2_071 | NIH2 | M | C-section | Single | 25 | 754 | Combination | NEC | 31 | n/a |
| N2_088 | NIH2 | F | C-section | N2_089 | 28 | 1057 | Formula | control | n/a | n/a |
| N2_093 | NIH2 | M | C-section | Single | 26 | 924 | Breast | NEC | 12 | n/a |
| N3_172 | NIH3 | M | C-section | Single | 28 | 1250 | Breast | NEC | 37, 54 | n/a |
| N3_173 | NIH3 | M | C-section | Single | 29 | 1530 | Breast | NEC | 25 | n/a |
| N3_174 | NIH3 | F | C-section | Single | 30 | 980 | Breast | control | n/a | n/a |
| N3_175 | NIH3 | M | Vaginal | Single | 29 | 1480 | Combination | control | n/a | n/a |
| N3_176 | NIH3 | M | C-section | Single | 28 | 990 | Combination | control | n/a | n/a |
| N3_177 | NIH3 | F | Vaginal | Single | 28 | 900 | Combination | control | n/a | n/a |
| N3_178 | NIH3 | M | Vaginal | Single | 32 | 2050 | Combination | NEC | 16 | n/a |
| N3_182 | NIH3 | M | C-section | Single | 39 | 3010 | Combination | NEC | 6 | n/a |
| N3_183 | NIH3 | M | Vaginal | Single | 32 | 2410 | Combination | NEC | 11 | n/a |
| S2_010 | NIH3 | M | C-section | Single | 32 | 1810 | Combination | control | n/a | n/a |

Scaffolds assembled from metagenomes were grouped into 3,643 bins, 1,457 of which represented draft-quality genomes (≥50% complete with ≤5% contamination; **Supplemental Figure 2**, **Supplemental File 1**). These genomes were assigned to 270 groups approximating different bacterial sub-species based on sharing ≥98% average nucleotide identity (ANI) (**Supplemental Table 2**). These genomes account for 91% of the total sequencing. High-quality draft genomes suitable for iRep replication rate analysis (≥75% complete with ≤175 fragments/Mbp and ≤5% contamination) were available for 193 genome clusters (Brown et al. 2016).

***Protein quantification by metaproteomics***

Across all metagenomes, 5,233,047 proteins were predicted, 897,520 of which were from a non-redundant set of representative genomes clustered at 98% ANI. Proteins clustered into 121,746 putative families (**Supplemental File 2**). Metaproteomics measurements were conducted on 87 metagenome-matched samples that spanned 16 infants, six of which developed NEC and one of which was diagnosed with sepsis (N1_019; **Supplemental Figure 1**). Each sample was analyzed in technical duplicate. Conducting metagenomics and metaproteomics on the same samples was critical for obtaining an appropriate database for matching peptides to proteins. On average 71,676 unique bacterial spectral counts were detected per sample, and an average of 33% of predicted bacterial proteins were identified (**Supplemental Figure 1**, **Supplemental Table 2, and Supplemental File 3**).

***Premature infants are colonized by genetically similar organisms***

The majority of infants were colonized by *Enterococcus faecalis*, *Klebsiella pneumoniae*, and *Staphylococcus epidermidis* (**Figure 1a,b**). However, most genotypes were seen in only one infant. Interestingly, species presence and absence patterns correlated with NEC (unweighted UniFrac distance Permutational Multivariate Analysis of Variance, PERMANOVA, p-value = 6 x $10^{-3}$; **Figure 1c**), but no individual species was strongly associated with NEC (**Supplemental Figure 3**). We deduce that this correlation is primarily due to low species diversity in NEC cases. The principal coordinate analysis (PCoA; **Figure 1c**) shows that some infants that developed NEC had distinct genome inventories. However, hierarchical clustering indicates that infants that developed NEC are often colonized by organisms that are genetically similar to those colonizing other infants. Genome inventories were also correlated with gestational age and birth weight (Mantel test p-value = 2 x $10^{-3}$ and 3 x $10^{-3}$, respectively).

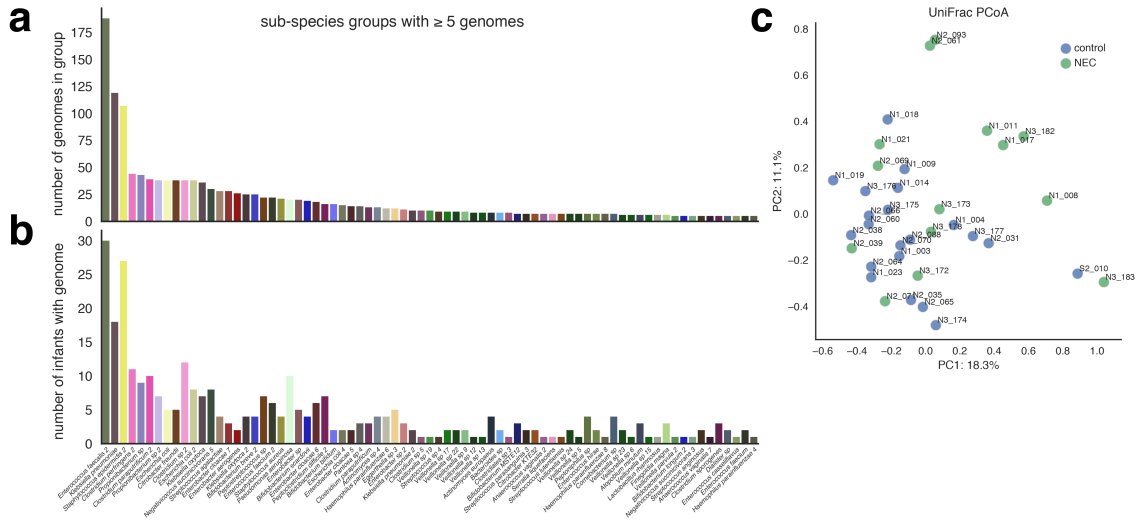


**Figure 1 | Premature infants are colonized by genetically similar bacteria.** Genomes reconstructed from metagenomes were clustered into sub-species groups based on sharing 98% average nucleotide identity (ANI). **a,** The number of genomes assigned to each group and **b,** the number of infants in the study with a reconstructed genome from the group. Shown are groups comprised of five or more genomes. **c,** Principal coordinate analysis (PCoA) clustering of infants based on unweighted UniFrac distances determined based on the ANI of assembled microbial genomes. Genome inventories from infants that developed NEC were significantly different from those that did not (PERMANOVA p-value = 3 x $10^{-3}$).

***Microbial communities cluster into seven primary types***

The premature infant microbiome was found to be highly variable. In some cases, samples collected from an infant at subsequent time points were as different from earlier samples as those collected from other infants (**Figure 2**). Communities were clustered based on species membership and abundance in order to identify microbial consortia common during the colonization process. In order to account for both genomic differences and organism abundance, clustering was conducted based on weighted UniFrac distances, where the tree used for calculating UniFrac was constructed using

genome ANI. Nine distinct community types were identified, seven of which were comprised of samples collected from multiple infants and were thus considered primary types (**Figure 2, Supplemental Figure 4, and Supplemental Figure 5**). Each community type is characterized by the dominance of different community members (**Supplemental Figure 6**). Microbiomes from different infants clustered into the same community type, and the microbiome of individual infants was found to switch types, sometimes multiple times, during the colonization process (**Figure 3**). Although infants shared community types, overall colonization patterns were not replicated across infants. Microbiomes associated with infants that did and did not go on to develop NEC were often classified in the same community type. In some cases, switches preceded onset of NEC, but no type or switch could explain all cases of NEC.

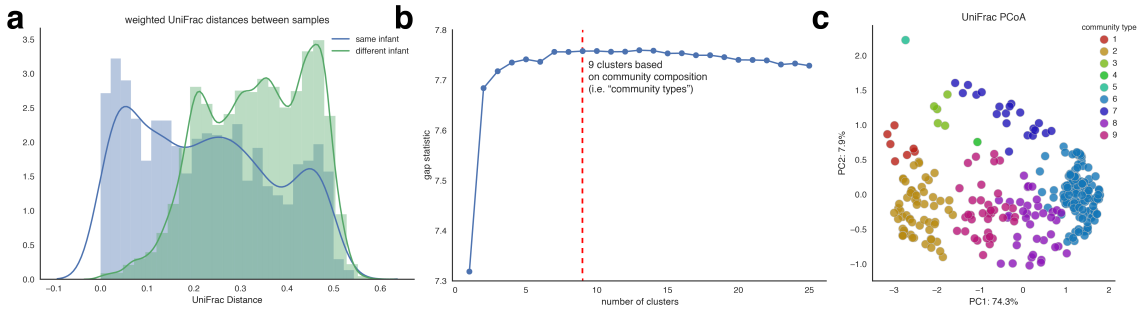


**Figure 2 | Premature infant gut microbial communities associate into seven primary types. a,** Pairwise weighted UniFrac distances calculated between all microbiome samples based on genome sequence ANI and abundance. **b,** The gap statistic was used to determine that nine distinct community types exist. Hierarchical clustering based on UniFrac distances was used to delineate samples into community types (**Supplemental Figure 4, Supplemental Figure 5, and Supplemental Figure 6**). Seven primary community types were identified based on being represented by samples collected from multiple infants. **c,** PCoA clustering of samples based on weighted UniFrac distances. Samples are colored based on community type assignment.

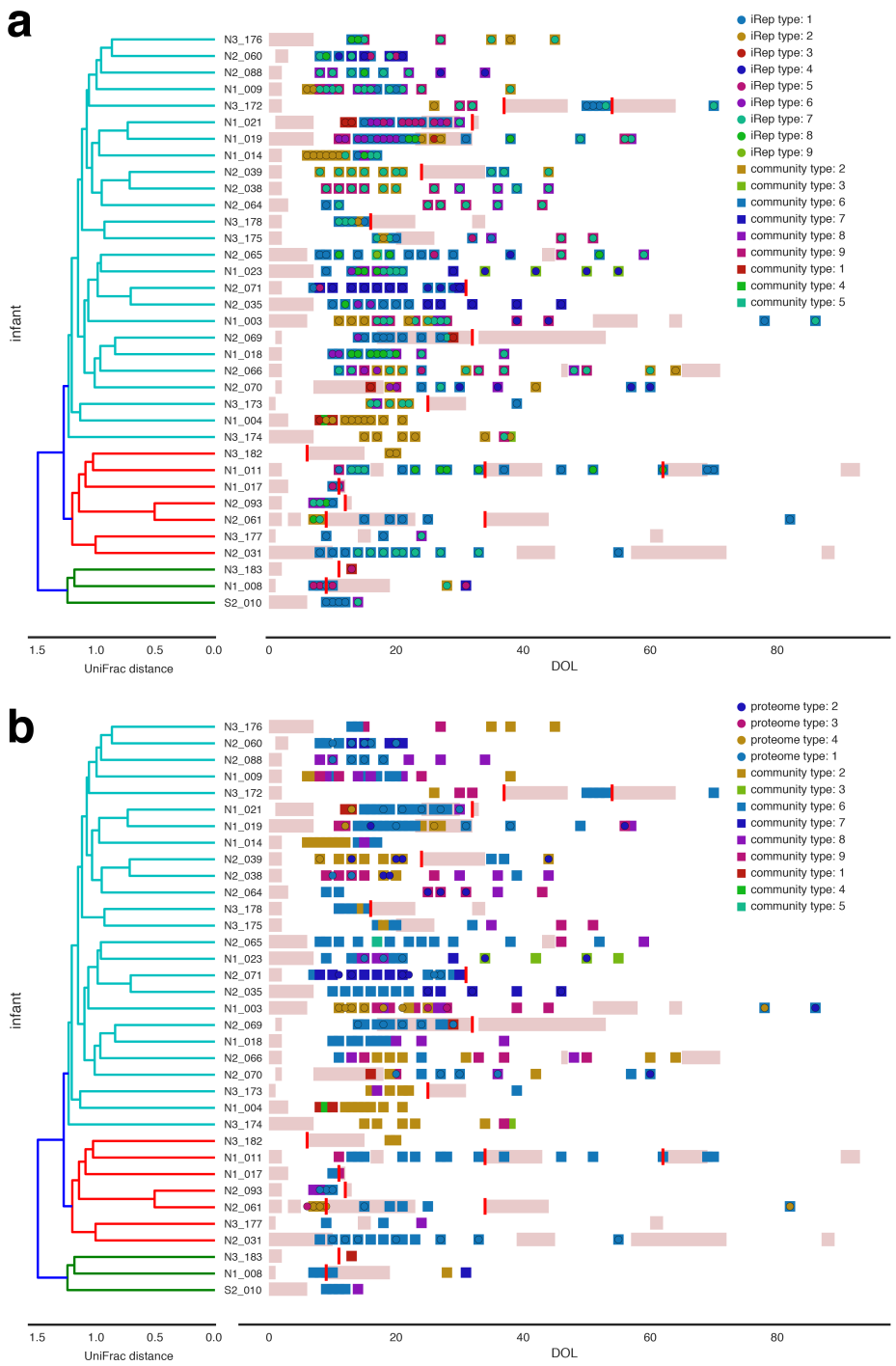


**Figure 3 | Microbial colonization patterns for preterm infants.** Samples were clustered into types based on microbial community composition ("community type"), bacterial iRep profiles ("iRep type"), and overall bacterial proteome composition ("proteome type"). Microbial community type is shown along with iRep **(a)** and proteome **(b)** types. Infants are arranged based on hierarchical clustering of unweighted UniFrac distances calculated based on the set of genomes recovered from each infant (**Supplemental Figure 3**). Antibiotics administration is indicated with pink bars and NEC diagnoses with red bars. DOL stands for day of life.

***Microbial community composition is correlated with infant health***

Premature infants that developed NEC had statically different microbial community abundance profiles (PERMANOVA p-value = 3 x $10^{-3}$; **Supplemental Figure 4g**). Interestingly, there was a variety of species that were detected in healthy infants, but never detected in those that developed NEC; however, the opposite was not true. It should be noted that species not detected in NEC infants were not consistently detected in healthy infants. No species identified five days prior to NEC diagnosis showed a significant difference in abundance, or was unique to NEC infants. Overall community composition was also correlated with each infant, antibiotics administration, birth weight, gestational age, and gestational age corrected day of life (GA + DOL; PERMANOVA or Mantel test p-value ≤0.01; **Supplemental Figure 4**). Several species were more abundant members of communities associated with infants that developed NEC: *Pantoea sp.*, *Pseudomonas aeruginosa*, *Enterobacter sp.*, *Propionibacterium sp.*, *Peptostreptococcus sp.*, and *Klebsiella oxytoca* (edgeR q-value ≤0.01; **Supplemental Table 3**), raising the question of whether or not these organisms are also replicating more quickly.

***Microbial replication rates and proteome composition***

iRep is a newly-developed method that enables measurement of bacterial replication rates based on metagenome sequencing data when high-quality draft genome sequences are available (Brown et al. 2016). We applied the iRep method using genomes recovered from metagenomes sequenced for each infant in the study, and quantified 1,328 iRep replication rates from 330 samples. Sample clustering was conducted based on community iRep profiles, identifying nine distinct iRep types that were correlated with community type (Mantel test p-value = 1 x $10^{-3}$, **Figure 3a**). Likewise, analysis of protein family abundance clustered samples into four distinct proteome types, which also correlated with community type (Mantel test p-value = 1 x $10^{-3}$, **Figure 3b**). Interestingly, there are several cases in which iRep and/or proteome type switched when community type was constant, or when community type switched but iRep and/or proteome type remained constant.

***Members of the same bacterial species replicate at different rates during colonization***

Across all infants, *Streptococcus agalactiae*, *Pseudomonas aeruginosa*, *Klebsiella pneumoniae*, and members of the genera *Veillonella* and *Clostridium* exhibited some of the highest replication rates (**Supplemental Figure 7**). Combined iRep values collected from infants that did and did not go on to develop NEC were not statistically different, even when considering only samples collected within the five days prior to NEC diagnosis (**Figure 4a**). However, iRep results show that populations of several individual species of bacteria were replicating more quickly in either infants that developed NEC or healthy controls (**Supplemental Table 4**). *Veillonella sp.* and *Pantoea sp.* were replicating faster in NEC infants, and *K. oxytoca* faster in control infants (Mann-Whitney U test, MW, p-value ≤0.01). Interestingly, *Pantoea sp.* were more abundant in infants that developed NEC. Several different species were active in control infants, but were not detected in infants that went on to develop NEC.

The iRep values for organisms sampled in this cohort during or immediately after antibiotics administration were not significantly different from those at other time points

(**Figure 4b**). This indicates that populations present after antibiotics administration are both resistant to antibiotics and are continuing to replicate. Members of several species were replicating quickly during or immediately following antibiotic treatment (*Veillonella sp.*, *Streptococcus agalactiae*, *Finegoldia magna*, and others; **Supplemental Table 4**). However, we did not detect overall higher iRep values following antibiotics administration, although this was reported previously (Brown et al. 2016). Most species were found only to be replicating in the absence of antibiotics, consistent with their susceptibility to the treatment.

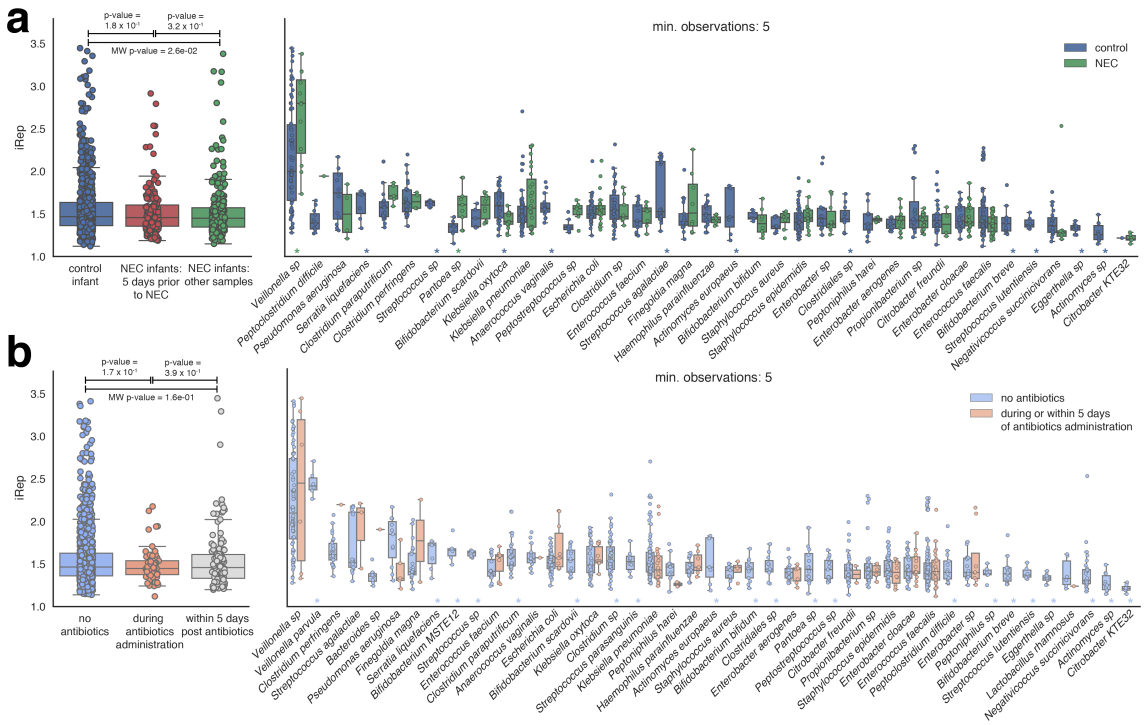


**Figure 4 | Replication rates for bacteria colonizing premature infants. a,** Replication rates for bacteria associated with infants that did and did not go on to develop NEC and **b,** sampled during periods with or without antibiotics administration. **a,** Overall replication rates were not statistically different between NEC and control samples or **b,** between samples collected during periods with or without antibiotics administration. Statistically significant differences between replication rates observed for individual species under different conditions are indicated with an asterisk (Mann-Whitney U Test p-value ≤0.01). Shown are all species with at least five observations.

***Low microbiome diversity is associated with NEC and antibiotics administration***

Microbial communities sampled from infants that developed NEC had significantly lower Shannon diversity compared with other infants (MW p-value = 4 x $10^{-4}$; **Figure 5a**). There was also a difference in diversity during and following antibiotics administration (MW p-value = 1.4 x $10^{-3}$). Analysis of Shannon diversity measurements indicate that lower diversity in NEC infant communities persists through the colonization period (**Figure 5b**).

Interestingly, microbiomes from both groups of infants increase in diversity during the first 20 days of life, and then decrease in diversity. This trend was also observed after correcting for gestational age (**Figure 5c**), and was also evident in healthy infants across samples where antibiotics were not administered (**Figure 5d**). Overall, differences in microbial community diversity are likely a driving factor in correlations between microbial community metrics and infant health.

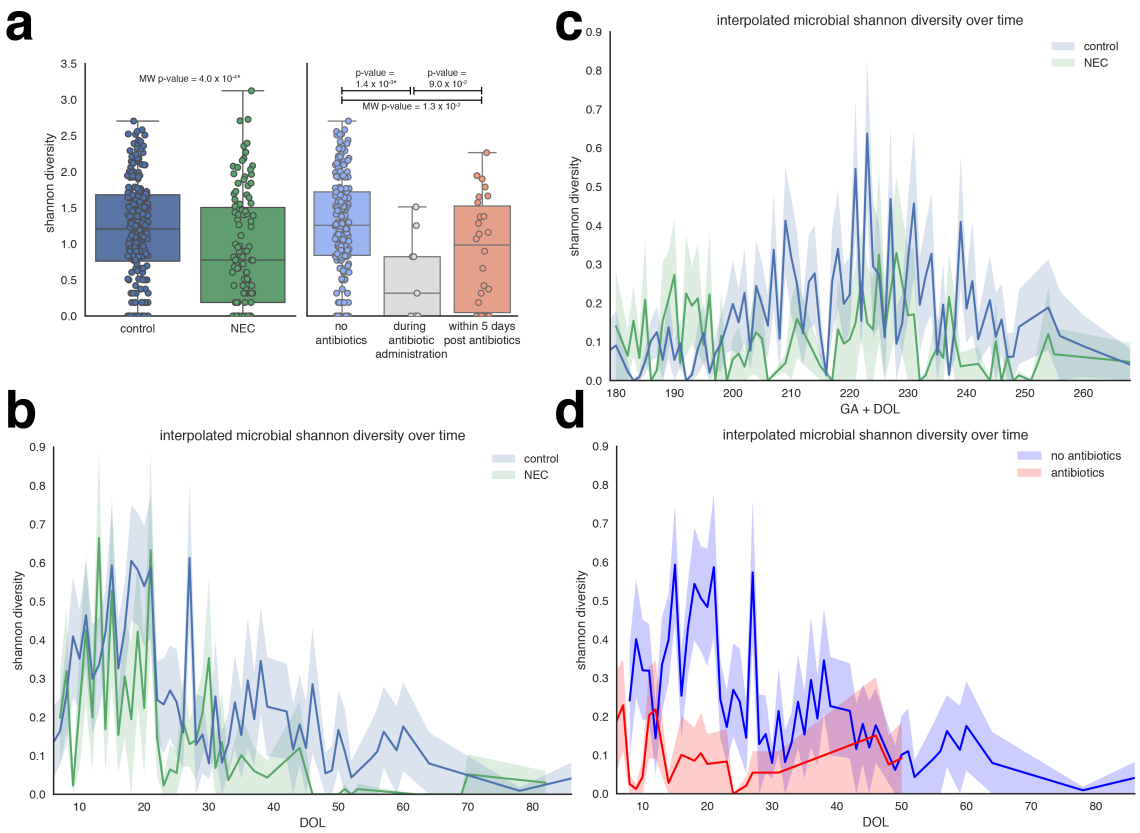


**Figure 5 | Microbial community diversity. a,** Shannon diversity measurements for microbial communities associated with infants that did and did not go on to develop NEC, and sampled during periods with or without antibiotics administration. Significant differences are indicated with an asterisk (Mann-Whitney U Test p-value ≤0.01). Interpolated Shannon diversity for infants that did and did not go on to develop NEC (**b, c**), and for samples collected from control infants during periods with or without antibiotics treatment (**d**). Infants experienced an increase in diversity during the first two weeks of life, followed by a decrease (**b**); this pattern was apparent even after correcting for gestational age (**c**). Plots are shown with a 95% confidence interval (**b-d**).

***Different species express varying amount of their proteome in the infant gut***

Microbes present in the gut environment are not expected to express their complete complement of proteins at all times. In order to investigate the extent of proteome expression for different bacteria, we compared depth of proteome sampling for each organism to the percent of the predicted proteome that could be detected (**Figure 6**). The median proteome detection across all samples was 11%, but this was largely due to low sampling depth. Higher depth of proteome sampling corresponded with detection of a

larger fraction of the predicted proteins. The median percent of the proteome detected for organisms with the best detection in each sample was 31% (max. 48%). For several frequently detected colonists, including *Klebsiella pneumoniae*, *Klebsiella oxytoca*, and members of the genus *Enterobacter,* maximum proteome expression was ~50%. However, *Propionibacterium sp.*, *Anaerococcus vaginalis*, and members of the genus *Bifidobacterium* expressed a greater proportion of their encoded genes than other organisms. We infer that these bacteria may be specifically adapted to environments and resource availability within the infant gut, whereas other bacteria may maintain capacities that enable adaption to other environments.

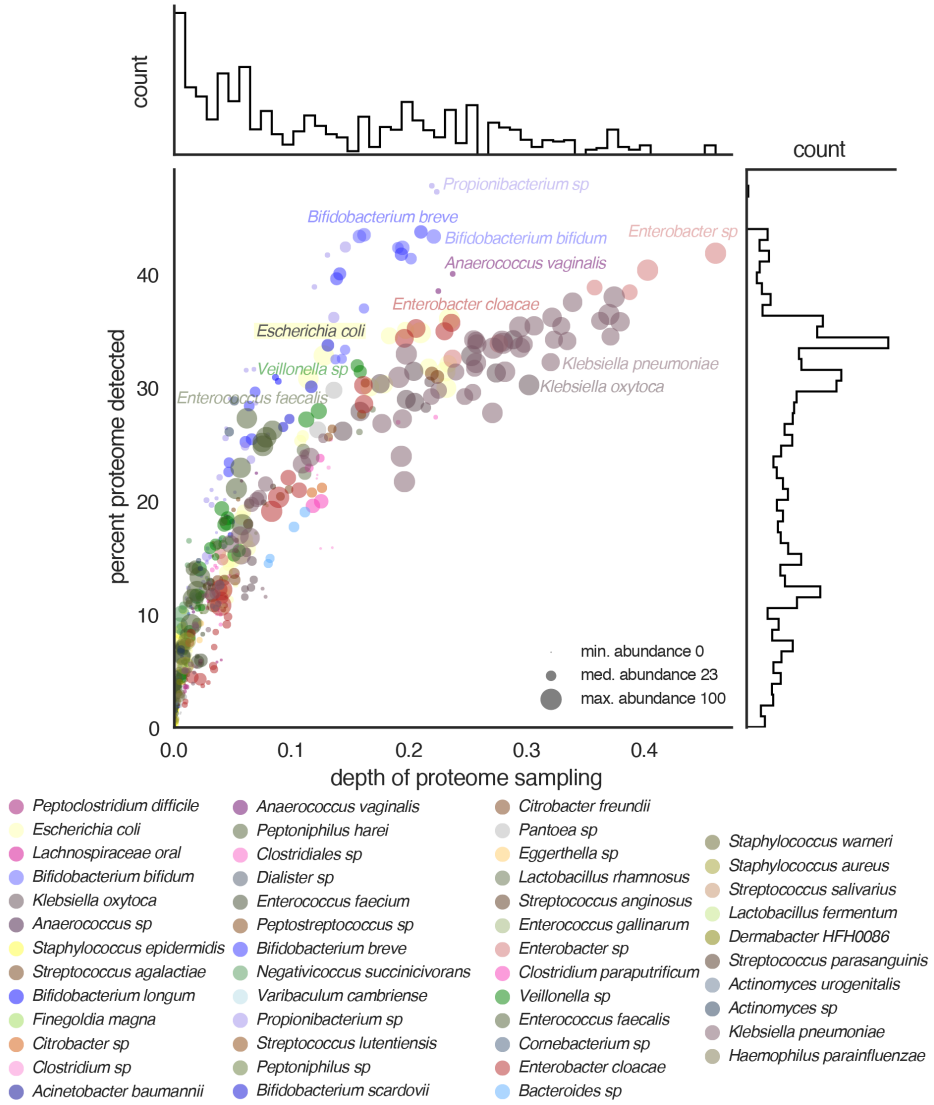


**Figure 6 | Proteome detection for species colonizing premature infants.** Depth of proteome sampling for organisms in each sample is compared against the percent of predicted proteins that could be detected. Data point sizes and histograms are scaled based on organism abundance as determined by metagenome sequencing.

***Microbiome development***

Peptide spectral counts were matched to infant-specific databases containing both human and microbial proteins. This allowed for the relative proportions of human and microbial proteins to be determined for each time point. Samples are dominated by human proteins during the first 10 days, and then microbial proteins become dominant around DOL 18. Ratios of human versus bacterial protein abundances show that the premature infant gut microbiome is established over a period of approximately two weeks (**Figure 7a**).

The presence of multiple data types (microbial community abundance and iRep, microbial community proteome composition, and human proteome composition) enabled tracking of various aspects of human and microbiome development during the first months of life (**Figure 7b,c**). All measurements from an infant were stable within the time scale of a week, but diverged over time. Interestingly, communities from different infants neither converged nor diverged over time in terms of similarity based on three of these five metrics. However, we observed that the human proteome measurements and microbial protein family abundances from different infants became increasingly different when samples with time separations of greater than three weeks were compared. Overall, the microbial proteome was more variable (higher variance) than community composition (**Figure 7d,e**). After approximately two weeks, both microbial community abundance and proteome measurements collected from the same infant became as different from each other as samples collected from other infants.

The majority of human and microbiome features recorded in our analyses were correlated with one another (**Figure 7f**). However, an exception is that microbial community abundance and iRep were not correlated with human proteome composition (Mantel test p-value $>0.01$). This is interesting in that it shows that there is no strong connection between the overall human proteome and either the composition or replication activity of the microbiome.

As shown in **Figure 7g**, microbial features were also correlated with a variety of infant factors, including infant health and development (gestational age and weight), as well as antibiotics administration (Mantel or PERMANOVA p-value $\leq 0.01$). Notably, whether or not an infant developed NEC ("condition") correlated with several microbiome factors (infant genome inventory, and both community composition and iRep), but not with proteome measurements.

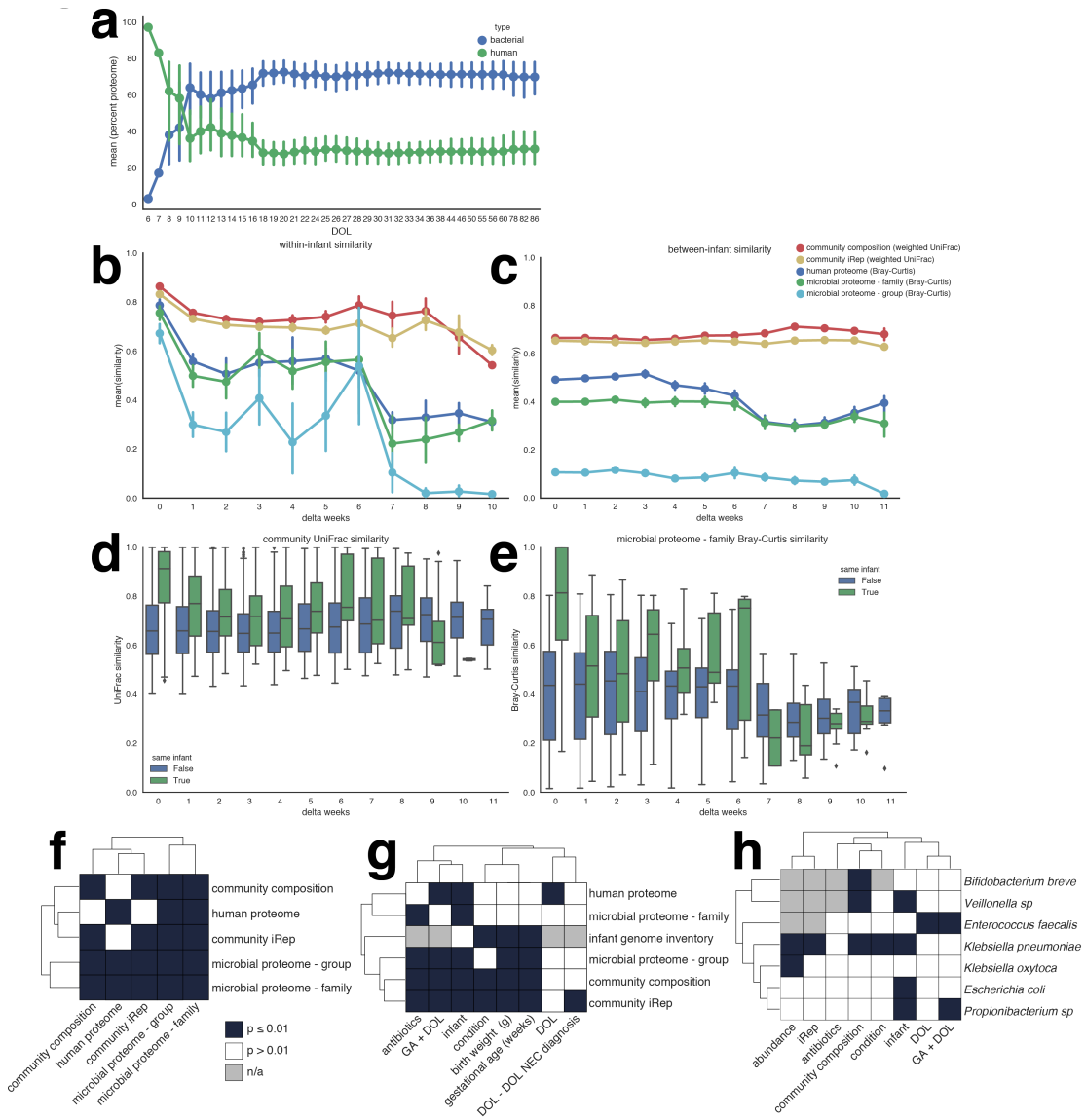


**Figure 7 | Microbiome stability and correlations.**
**a,** The relative contribution of human and bacterial proteins to overall proteome composition during development of the premature infant gut. **b,** Similarity measurements for microbiomes sampled either from the same infant or **c,** from different infants. Comparison of similarity measurements calculated between samples collected either form the same or different infants based either on weighted microbial community UniFrac (**d**), or weighted microbial proteome BrayCurtis (**e**) measurements. Human proteome and microbial community correlations calculated between one another (**f**), with infant metadata (**g**), and determined based on microbial species (**h**). Shown are PERMANOVA or Mantel test p-values (**f-h**).

***Microbial proteins associated with proteome type, NEC, and antibiotics administration***
As described above, we used protein abundance patterns to cluster microbial community proteomes into functionally distinct proteome types. Statistical analysis identified 3,085 differentially abundant proteins distinguish proteome types (edgeR q-value ≤ 0.01;

**Supplemental Table 5**). Of these, 461 were found to distinguish only one proteome type from all others. Notable amongst all of these proteins were those involved in central carbohydrate metabolism and energy metabolism (**Supplemental Figure 8**). Proteome types differ in terms of the amount and type of carbon degradation enzymes, as well as the propensity for aerobic versus anaerobic respiration (based on the abundance of oxidases and reductases).

Although overall community proteome abundance profiles were not correlated with NEC, microbial proteins from 160 different protein families, many with no known function, were more abundant in samples from infants that went on to develop NEC (identified in more than one NEC infant, edgeR q-value ≤0.01; **Supplemental Table 5**). The proteins with known functions were dominantly involved in transport of ions, metals, and other substrates, iron acquisition, and both motility and chemotaxis. Among proteins responsible for iron scavenging was subunit E of enterobactin synthase, a high-affinity siderophore involved in iron acquisition, which is often used by pathogenic organisms. Also more abundant was outer membrane receptor FepA, which is involved in transporting iron bound by extracellular enterobactin. Subunit F of enterobactin synthase was also identified in NEC infants, as were an iron-enterobactin ABC transporter substrate-binding protein, and an enterobactin esterase. The abundance of this protein suggests a possible role for iron acquisition by organisms that may contribute to disease onset.

Overall community proteome composition was correlated with antibiotics administration. Samples collected during treatment were enriched in 56 different proteins (identified in more than one treated infant, edgeR q-value ≤0.01; **Supplemental Table 5**). Amongst these proteins were those involved in secretion, transcription, and DNA degradation. Along with iRep results, the findings indicate that a subset of organisms remain active in the presence of antibiotics.

***Species-specific metabolic profiles are associated with specific infants, infant development, microbial community composition, and replication rate***

Normalization of proteome data for each genome indicates relative protein abundance levels, which can then be tracked across samples. Using this approach, we identified population-specific proteome profiles and tested whether or not they correlate with various human and microbial properties (**Figure 7h**, **Supplemental Figure 1**, **and Supplemental Table 6**). *Veillonella spp., Klebsiella pneumoniae, Escherichia coli,* and *Propionibacterium sp.* were all correlated with infant (PERMANOVA p-value ≤0.01), indicating that although similar organisms are colonizing different infants, each population is expressing a different complement of proteins. *K. pneumoniae* and *Veillonella spp.* proteomes also correlated with community type, as did the *Bifidobacterium breve* proteome (Mantel test p-value ≤0.01). This shows that populations are responding to their overall microbial community context. Interestingly, both *Enterococcus faecalis* and *Propionibacterium sp*. exhibited proteomes that were also correlated with infant development. Although overall microbial proteome correlated with antibiotics administration, species-specific proteome profiles did not; however, this may

be due to a lack of available data for the same species in multiple samples with and without antibiotics.

Because of the existence of 35 samples in which ≥10% of the *K. pneumoniae* proteome could be detected (max. = 38%, median = 25%), correlations between individual protein abundances and iRep could be determined. The *K. pneumoniae* proteome correlated with iRep and infant health. Amongst proteins positively correlated with iRep were a transcriptional regulator (LysR), proteins involved in cell wall biogenesis, and ribosomal proteins (Pearson ≥0.5, q-value ≤0.01, observed in ≥15 samples; **Supplemental Table 7**).

***Species-specific proteins correlated with NEC***

*Propionibacterium sp.* was found to be more abundant in microbiomes from infants that developed NEC; however, its proteome composition did not correlate with disease, and only three proteins were strongly correlated with NEC: glyceraldehyde-3-phosphate dehydrogenase (family 123), and two hypothetical proteins (family 9917 and 12000).

Interestingly, 21 *K. pneumoniae* proteins were correlated with NEC, including a ferrous iron transporter (family 2834) that was 3.9-fold more abundant in two infants that developed NEC. The abundance of this protein was also correlated with infant, proteome type, community type, and antibiotics administration.

*Pantoea sp.* was more abundant and replicating more quickly in infants that developed NEC. However, not enough samples were available to draw conclusions about proteins associated with NEC.

***Different infants are colonized by different strains with distinct proteomes***

The finding that *K. pneumoniae, E. coli, Propionibacterium sp.*, and *Veillonella spp.* have infant-specific proteomes raised the question of whether or not each infant was being colonized by different strains. All draft-quality genome sequences assembled for each species from each infant were compared with one another, and hierarchical clustering conducted on pairwise ANI values was used to delineate strains (**Supplemental Figure 9**). Clustering showed that in most cases each infant was indeed colonized by distinct strains, which proteomics analysis showed are functionally distinct. However, there were a few notable exceptions. Twin infants N2_069 and N2_070, as well as infant N1_003 were all colonized by the same strain of *K. pneumoniae*. The proteomic profiles for the strains colonizing N2_069 and N2_070 were more similar to one another than they were to profiles recovered from other strains; however, they were still distinguishable (**Figure 8**). Likewise, the same strain of *Propionibacterium sp.* colonized twin infants N2_038 and N2_039. As with shared strains of *K. pneumoniae*, their functional profiles clustered together but were still distinguishable from one another (**Supplemental Figure 10**).

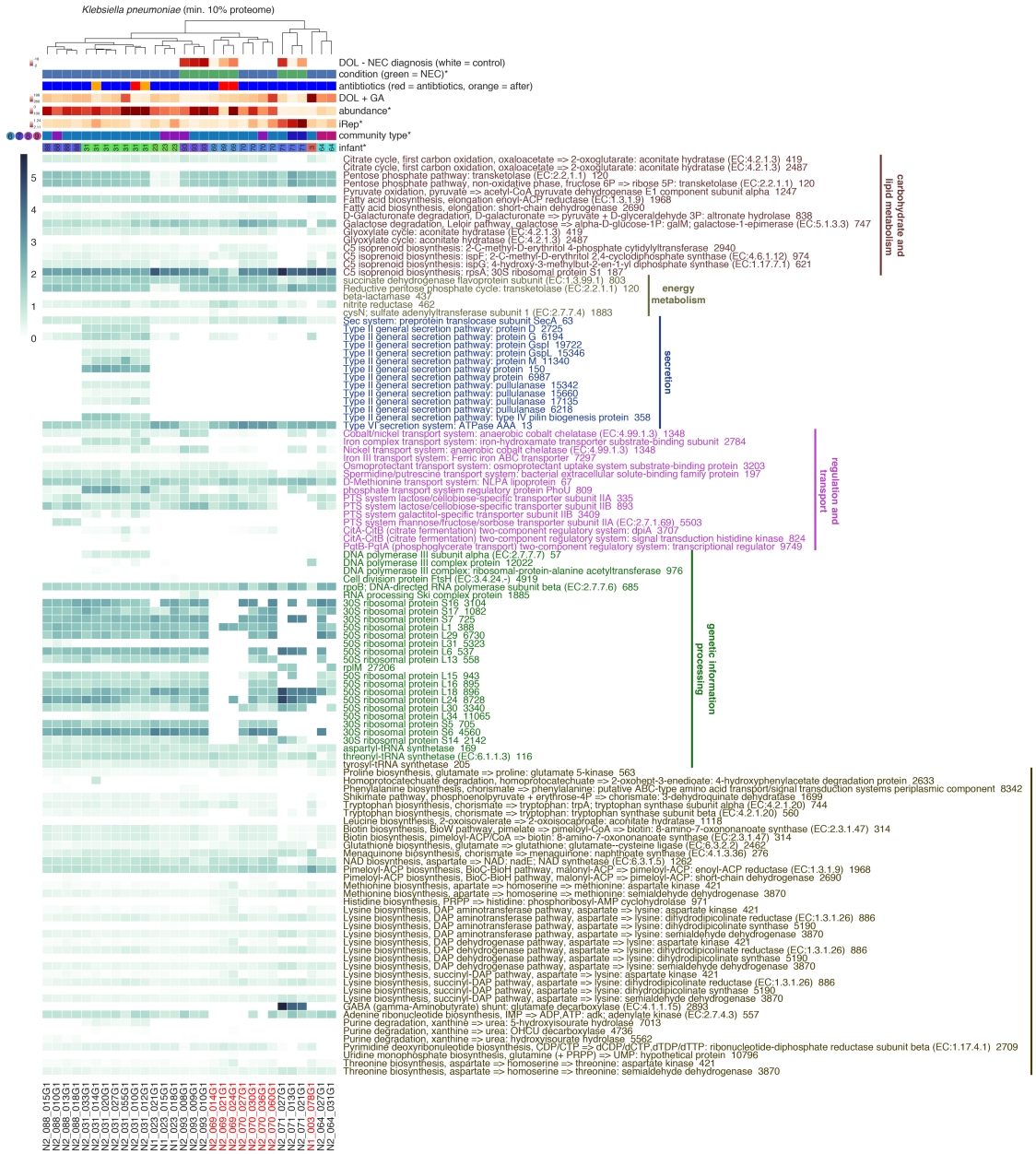


**Figure 8 | *Klebsiella pneumonia* proteins with infant-specific abundance profiles.** Hierarchical clustering was conducted on all *K. pneumonia* protein families, showing that strains colonizing different infants have distinct proteomic profiles. Infant and species metadata are shown for each sample. Metadata significantly correlated with the *K. pneumonia* proteome are indicated with an asterisk (PERMANOVA or Mantel test p-value ≤0.01). Protein families that correlated with at least one infant are shown in the heatmap (edgeR q-value ≤0.01). Samples colonized by the same *K. pneumonia* strain are shown with red text.

Analysis showed that few proteins were responsible for distinguishing proteomes of the same bacterial types in different infants (**Figure 8, Supplemental Figure 10, and Supplemental Table 6**). Common amongst these were proteins involved in nucleotide,

amino acid, carbohydrate and lipid metabolism. Also notable were several proteins produced by *K. pneumoniae* involved in central carbohydrate metabolism and both galactose degradation and D-galacturonate degradation, indicating different carbon preferences for strains colonizing different infants (**Figure 8**). Several proteins involved in bacterial secretion were different between *K. pneumonia* colonizing different infants, indicating differences in secretion potential that could affect human-microbe interactions. Relatedly, the abundance of proteins involved in transport of metals, ions, citrate, and several sugars also differed between infants.

## Discussion

Most studies to date have focused on the composition of the gut microbiome, typically at the low-resolution afforded by 16S rRNA gene amplicon methods. We used genome-resolved time-series metagenomics in conjunction with iRep replication rate and metaproteomics measurements to obtain a more comprehensive view of the colonization process. The dataset included information about the gut colonization trajectories of both healthy infants and infants that went on to develop NEC, enabling exploration of microbiome variability, at both the community composition and organism functional levels.

Microbial communities were classified into types based on the mixture of organisms present. Interestingly, most types occurred in multiple infants, a result that indicates the tendency of gut colonizing bacteria to form networks of interaction, possibly based on metabolic complementarity. An important factor determining the community type present may be the specific organisms that are introduced, and the extent to which they are able to colonize. Other factors that may dictate the community type include human genetic selection, diet, and antibiotics administration. Within a single infant, community types often switched several times over the observation period. Given the lack of evidence for consistent transitions from one type to another across multiple infants, the high degree of variation in iRep replication rates observed for members of the same species, and a lack of convergence of communities in different infants, we conclude that colonization is a chaotic process.

Overall microbial physiology, as measured by whole proteome abundance patterns, was more dynamic than community composition. Thus, metagenomics-enabled proteomic analyses indicate functional flexibility that does not depend on addition or loss of organisms. Shifts in the importance of specific pathways or metabolisms with environmental conditions would not be apparent in studies that only use organism identification or metabolic potential predictions.

An unexpected finding was that the microbiomes from all infants increase in diversity during the first 20 days of life, but then decrease in diversity. To some extent this could be an artifact of the study design, because infants that were hospitalized for longer may have been sicker and thus contributed substantially to the later time points. Alternatively, spatial heterogeneity in the gut may have reduced niche diversity, or decreasing diversity could reflect stabilization of the community. Investigation of infants over longer time

periods is required to clarify the extent of fluctuation along the path to establishment of diverse infant gut communities that are characteristic of older infants.

It is possible that onset of NEC is due to fast growth rates of potential pathogens within communities that are imbalanced due to low species richness, ultimately resulting in overgrowth by a pathogen. For this reason, we compared microbial community diversity and composition, growth rates, and metabolic features in infants that did and did not develop NEC. A clear finding of this study, and evident from prior research (Pammi et al. 2017), is that microbial communities associated with infants that develop NEC are of lower diversity compared with control infants at the same developmental age. Microbial community composition was correlated with NEC, but this may be largely due to lower diversity. Several different species have higher relative abundance in infants that developed NEC, but none of these species were consistently associated with the disease. The correlation could be the consequence of the loss of other organisms from the community rather than their higher absolute abundance. Thus, low community diversity may be a confounding factor in studies that have associated specific organisms with NEC.

The difference in diversity between healthy infants and those that develop NEC raises the question of how specific organisms modulate their behavior over the colonization period. We noted that specific bacteria had iRep values that, prior to onset of NEC, were anomalously high relative to values for the same organism in control infants (*Veillonella sp.* and *Pantoea* sp.). This may be medically important, but additional examples are needed to establish a link between rapid growth and NEC.

Interestingly, whether or not an infant developed NEC was not correlated with overall proteome composition. However, there were specific proteins that were associated with NEC, notably several involved in iron scavenging. Given that this is an important process often associated with pathogenesis, it is possible that increased activity of iron scavenging pathways could contribute to organism proliferation and onset of NEC. In addition, the *Klebsiella pneumoniae* proteome was correlated with NEC, including a protein involved in transport of iron. This is intriguing considering the prior finding that supplementation of lactoferrin, an abundant breast milk protein involved in modulating iron levels in the gut, decreases risk of developing necrotizing enterocolitis (Manzoni 2016; Raghuveer et al. 2002). Overall, these findings indicate that fine-scale, species-specific proteins are important for understanding disease onset. Although the microbial community, and specific microbial proteins were correlated with NEC, no individual organism or protein was significantly more abundant in all cases. This finding supports the hypothesis that NEC is a multifaceted disease with multiple routes that lead to onset.

Although species-specific proteome profiles were correlated with community composition, they were largely infant specific. This is an interesting observation because it implies a feedback between human physiological conditions in the gut, which likely vary substantially from infant to infant and over time, and microbiome function.

## Methods

### *Sample collection and metagenome sequencing*

Samples were collected, processed for metagenome sequencing, and sequenced as part of three prior studies (**accession numbers in Supplemental Table 1**) (Raveh-Sadka et al. 2015; 2016; Brooks et al.). Stool samples were collected from infants and stored at −80°C. DNA was extracted from frozen fecal samples using the MO BIO PowerSoil DNA Isolation Kit, with modifications (Raveh-Sadka et al. 2016). DNA libraries were sequenced on an Illumina HiSeq for 100 or 150 cycles (Illumina, San Diego, CA). All samples were collected with parental consent.

### *Metagenome assembly and genome binning*

We re-assembled and analyzed metagenomes generated as part of a prior study, referred to as NIH1 (Raveh-Sadka et al. 2016). The data were processed in a manner consistent with the two other prior studies analyzed, referred to as NIH2 (Brooks et al.) and NIH3 (Raveh-Sadka et al. 2015). All raw sequencing reads were trimmed using Sickle (https://github.com/najoshi/sickle). Each metagenome was assembled separately using IDBA_UD (Peng et al. 2012). Open reading frames (ORFs) were predicted using Prodigal (Hyatt et al. 2010) with the option to run in metagenome mode. Predicted protein sequences were annotated based on USEARCH (–ublast) (Edgar 2010) searches against UniProt (The UniProt Consortium 2015), UniRef100 (Suzek et al. 2007), and KEGG (Kanehisa et al. 2012; Minoru Kanehisa 2000). Scaffold coverage was calculated by mapping reads to the assembly using Bowtie2 (Langmead and Salzberg 2012) with default parameters for paired reads.

Scaffolds from NIH1 infants were binned to genomes using Emergent Self-Organizing Maps (ESOMs) generated based on time-series abundance profiles (Dick et al. 2009; Sharon et al. 2012). Reads from every sample were mapped independently to every assembly using SNAP (Zaharia et al. 2011), and the resulting coverage data were combined. Coverage was calculated over non-overlapping 3 Kbp windows. Coverage values were normalized first by sample, and then the values for each scaffold fragment were normalized from 0-1. Combining coverage data from scaffolds assembled from different samples prior to normalization made it possible to generate a single ESOM map for binning genomes assembled independently from each sample. ESOMs were trained for 10 epochs using the Somoclu algorithm (Wittek et al. 2013) with the option to initialize the codebook using Principal Component Analysis (PCA). Genomes were binned by manually selecting data points on the ESOM map using Databionics ESOM Tools (Ultsch 2005). Binning was aided by coloring scaffold fragments on the map based on BLAST (Altschul et al. 1990) hits to the genomes assembled in the prior study.

As part of the NIH2 and NIH3 studies, scaffolds were binned based on their GC content, DNA sequence coverage, and taxonomic affiliation using ggKbase tools (ggkbase.berkeley.edu). Genome bins from all three datasets were classified based on the consensus of taxonomic assignments for predicted protein sequences. Genome completeness and contamination were estimated for all genomes using CheckM with the taxonomy_wf option (Parks et al. 2015). Genomes with extra single copy genes, but with

≤175 fragments/Mbp (normalized for contamination) that were estimated to be ≥75% complete were manually curated based on scaffold GC content and coverage.

***Clustering genomes into sub-species groups***

Genomes were clustered into sub-species groups based on sharing ≥98% average nucleotide identity (ANI), as estimated by MASH (Ondov et al. 2016). Representative genomes were selected for each cluster as the largest genome with the highest expected completeness and smallest amount of contamination. Genomes were classified based on the lowest possible consensus of taxonomic assignments for predicted protein sequences. Taxonomic assignments for representative genomes were checked manually based on hits to ribosomal protein S3, or visual inspection of protein taxonomic assignments. In order to identify cases in which the same bacterial strain was present in multiple samples, sub-species groups were further analyzed with the ANIm algorithm (Richter and Rossello-Mora 2009) implemented in dRep (Olm et al. 2017).

***Measuring microbial community abundance and replication rates***

In order to achieve accurate abundance and replication rate measurements from read mapping, databases of representative genomes were created for each sample. Each database was constructed in order to include a representative genome from important sub-species groups. Priority was given to high-quality draft genome sequences reconstructed from the same sample. Genomes were classified as high-quality draft based on the requirements for iRep replication rate analysis (https://github.com/christophertbrown/iRep): ≥75% complete, ≤2.5% contamination, and ≤175 scaffolds per Mbp of sequence (Brown et al. 2016). Genomes were selected to represent sub-species groups using the following priority scheme: 1) high-quality draft genome assembled from the same sample, 2) high-quality draft genome from the same infant, 3) high-quality draft genome representative of sub-species group from any infant (if group had ≥5 representatives), 4) best genome from infant (if a genome was available). iRep was conducted using reads that mapped to genome sequences with ≤1 mismatch per read sequence. In cases where iRep values were ≥3, coverage plots were inspected and values were removed if there was evidence of strain variation.

We considered bacterial sub-species to be present in a sample if ≥97% of the genome was covered by an average of ≥2 reads. Abundance and iRep measurements were compared across samples by linking sample-specific representative genomes to sub-species groups. Relative abundance measurements for each sub-species group were calculated by converting DNA sequencing coverage values to a percentage. UniFrac (Lozupone and Knight 2005) analysis was conducted based on rarefied abundance data and a tree constructed based on pairwise genome ANI values measured using MASH (-ms 5000000).

***Metaproteomics analysis***

Metaproteomics sequencing was conducted on 0.3 g of stool as previously described (Xiong et al. 2017). Each sample was suspended in 10 mL cold phosphate buffered saline. Samples were filtered through a 20 μm size filter to enrich for microbial cells and proteins. Microbial cells were collected by centrifugation, boiled in 4% sodium dodecyl

sulfate for 5 minutes, and sonicated to lyse cells. The resulting protein extract was precipitated with 20% trichloroacetic acid at -80°C overnight. The protein pellet was washed with ice-cold acetone, solubilized in 8 M urea, reduced with 5 mM dithiothreitol, and cysteines were blocked with 20 mM iodoacetamide. Then sequencing grade trypsin was used to digest the proteins into peptides. Proteolyzed peptides were then salted and acidified by adjusting the sample to 200 mM NaCl, 0.1% formic acid, followed by filtering through a 10 kDa cutoff spin column filter to collect tryptic peptides.

Peptides were quantified by BCA assay and 50 µg peptides of each sample were analyzed via two-dimensional nanospray LC-MS/MS system on an LTQ-Orbitrap Elite mass spectrometer (Thermo Scientific). Each peptide mixture was loaded onto a biphasic back column containing both strong-cation exchange and reverse phase resins (C18). As previously described, loaded peptides were separated and analyzed using a 11-salt-pusle MudPIT protocol over a 22-h period (Xiong et al. 2015). Mass spectra were acquired in a data-dependent mode with following parameters: full scans were acquired at 30 k resolution (1 microscan) in the Orbitrap, followed by CID fragmentation of the 20 most abundant ions (1 microscan). Charge state screening and monoisotopic precursor selection were enabled. Unassigned charge and charge state +1 were rejected. Dynamic exclusion was enabled with a mass exclusion width of 10 ppm and exclusion duration of 30 seconds. Two technical replicates were conducted for each sample.

Protein databases were generated for each infant from protein sequences predicted from assembled metagenomes. The database also included human protein sequences (NCBI Refseq_2011), common contaminants, and reverse protein sequences, which were used to control the false discovery rate (FDR). Collected MS/MS spectra were matched to peptides using MyriMatch v2.1 (Tabb et al. 2007), filtered, and assembled into proteins using IDPicker v3.0 (Ma et al. 2009). All searches included the following peptide modifications: a static cysteine modification (+57.02 Da), an N-terminal dynamic carbamylation modification (+43.00 Da), and a dynamic oxidation modification (+15.99). A maximum 2% peptide spectrum match level FDR and a minimum of two distinct peptides per protein were applied to achieve confident peptide identifications (FDR <1%). To alleviate the ambiguity associated with shared peptides, proteins were clustered into protein groups by 100% identity for microbial proteins and 90% amino acid sequence identity for human proteins using USEARCH (Edgar 2010). Spectral counts were balanced between shared proteins.

***Identification of putative protein families***

Putative protein families were identified in order to track the presence and abundance of different protein types across samples. ORFs were first pre-clustered at 95% identity using USEARCH (-cluster_smallmem -target_cov 0.50 -query_cov 0.95 -id 0.95), and then all-versus-all protein searches were conducted (–ublast -evalue 10e-10 -strand both). Protein families were delineated from within the all-versus-all network graph using the MCL clustering algorithm (-I 2 -te 10) (Enright et al. 2002). The most common annotation observed across all protein sequences in the group was selected as the annotation for the putative protein family.

***Tracking human and bacterial protein abundances***
Human and bacterial protein abundances were normalized using the weighted trimmed mean method from EdgeR (Robinson et al. 2009). Species-specific proteomic profiles were normalized as the percent of total balanced spectral counts.

***Sample clustering and statistical analyses***
Sample clustering was conducted based on microbial community abundance and iRep profiles, and bacterial protein family abundance profiles. In each case, the number of clusters was determined using the gap statistic (Tibshirani et al. 2001), and then samples were grouped into the appropriate number of clusters using hierarchical clustering (average linkage method). Microbial community data was clustered based on weighted UniFrac distances, and protein data using Bray-Curtis distance. EdgeR was used to calculate statistically significant differences between conditions using quasi-likelihood linear modeling (glmQLFTest).

## Acknowledgements

Sample collection was approved by the University of Pittsburgh Institutional Review Board (PRO10090089). Funding was provided by National Institutes of Health grants R01-AI-092531 and R01-GM-103600.

## Author Contributions

MJM oversaw sample collection, RB collected all samples and managed metadata, and BF coordinated sample processing for DNA sequencing and proteomics analysis. CTB and MRO assembled and annotated the metagenome data. CTB and JFB carried out the genome binning and curation. CTB conducted the microbial community time series abundance and iRep analyses. WX and RLH generated the proteomics data, which was analyzed by CTB. CTB, MRO, and BCT provided bioinformatics support. CTB and JFB wrote the paper, and all authors provided input to the final text.

## Disclosure Declaration

The authors declare no competing financial interests.

## Supplemental Materials

### *Supplemental Figures*

**Supplemental Figure 1 | Metagenome sequencing and metaproteomics conducted on microbiome samples collected from premature infants.** Frequency of sample collection for metagenomics (**a**) and metaproteomics (**b**) based on infant day of life (DOL). **c,** Metagenome sequencing, and **d**, the percentage of each metagenome represented by assembled draft-quality genome sequences. **e**, The number of proteomics spectral counts that could be uniquely assigned to human or bacteria. **f,** The percent of predicted proteins that could be detected in each sample. **g,** The percent of species-specific proteomes that could be detected for species where ≥10% of the proteome could

be detected in at least one sample. **h,** Histogram showing the distribution of the maximum percent of the proteome detected for all species present in each sample.

**Supplemental Figure 2 | ESOM genome binning.** Genome binning was conducted based on Emergent Self-Organizing Map (ESOM) clustering of scaffolds assembled from individual metagenomes. Data points represent three Kbp fragments of assembled scaffolds. Coloring is based on the species-level assignment of reconstructed draft-quality genomes. The map is periodic, and red boxes indicate a single period.

**Supplemental Figure 3 | Infants that developed NEC and healthy controls are colonized by genetically similar bacteria.** Presence (dark boxes) and absence (white boxes) of members of bacterial sub-species in microbial communities from different infants. Sub-species were identified based on sharing ≥98% genome average nucleotide identity (ANI), and were determined to be present if ≥97% of the genome was covered by an average of ≥2 reads. Hierarchical clustering was conducted based on unweighted UniFrac distances calculated between infant genome inventories.

**Supplemental Figure 4 | Studied infant gut microbial communities associate into seven primary community types. a**, Hierarchical clustering was conducted based on the abundance of bacterial sub-species using weighted UniFrac distances. Microbial community types are identified by colored boxes. Metadata are shown for each sample, and indicated with an asterisk if significantly correlated with microbial community abundance data (PERMANOVA or Mantel test p-value ≤0.01). **b-i**, PCoA clustering of microbial communities with associated metadata: antibiotics administration (**b**), infant (**c**), developmental age (**d**; number of days since conception: gestational age + day of life, GA + DOL), proteome type (**e**), iRep type (**f**), infant health (**g**), days prior to NEC diagnosis (**h;** DOL – NEC diagnosis), and human proteome type (**i**).

**Supplemental Figure 5 | Microbial community abundance and replication rate profiles.** Relative abundance (bars) and iRep replication rate (scatter plot) values for bacterial sub-species colonizing studied premature infants. The five days following antibiotics administration are indicated with a color gradient.

**Supplemental Figure 6 | Microbial community types are distinguished by their abundant members.** Rank abundance curves showing the average and range (95% confidence interval) of relative abundance values for sub-species groups associated with each community type.

**Supplemental Figure 7 | iRep replication rates for members of bacterial species colonizing premature infants.** Shown are all species with at least five observations.

**Supplemental Figure 8 | Proteome types are distinguished by the abundance of proteins from different KEGG modules.** Hierarchical clustering of proteome types was conducted based on the abundance of proteins associated with KEGG modules. The relative abundance of proteins associated with each module was summed for each sample, and then the average was taken across all samples associated with each proteome type.

**Supplemental Figure 9 | Hierarchical clustering of genomes for members of the same sub-species group.** dRep results show ANI clustering of assembled genomes. Genome names indicate the metagenome that each genome was assembled from (**see Supplemental Table 2**). Clustering dendrograms show that most infants are colonized by different strains.

**Supplemental Figure 10 | Multiple species have infant-specific proteome profiles. a**, Analysis of *Veillonella spp.* genomes shows the presence of four different species. **b-e**, Proteome profiles for different species colonizing premature infants. Hierarchical clustering was conducted based on all detected protein families, and shows that strains colonizing different infants typically have distinct proteomic profiles. Infant and species metadata are shown for each sample. Metadata significantly correlated with the species proteome are indicated with an asterisk (PERMANOVA or Mantel test p-value ≤0.01). Protein families that correlated with at least one infant are shown in the heatmap (edgeR q-value ≤0.01). Samples colonized by the same strain are shown with colored text.

***Supplemental Tables***

**Supplemental Table 1 | DNA sequencing and metaproteomics statistics.**

**Supplemental Table 2 | Genomes reconstructed from metagenomes.**

**Supplemental Table 3 | Species relative abundance and statistical analysis.**

**Supplemental Table 4 | Species iRep replication rates and statistical analysis.**

**Supplemental Table 5 | Microbial protein family abundance and statistical analysis.**

**Supplemental Table 6 | Species-specific microbial protein family abundance and statistical analysis.**

**Supplemental Table 7 | Correlation of species-specific protein family abundances with iRep replication rates and gestational age corrected day of life (GA + DOL).**

***Supplemental Files***

**Supplemental File 1 | Scaffolds binned to reconstructed genomes.**

**Supplemental File 2 | Proteins assigned to putative families.**

**Supplemental File 3 | Metaproteomics spectral counts.**

## References

Altschul SF, Gish W, Miller W, Meyers EW, Lipman DJ. 1990. Basic Local Alignment Search Tool. *J Mol Biol* **215**: 403–410.

Bokulich NA, Chung J, Battaglia T, Henderson N, Jay M, Li H, Lieber AD, Wu F, Perez-Perez GI, Chen Y, et al. 2016. Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Sci Transl Med* **8**.

Brooks B, Olm MR, Firek BA, Baker R, Thomas BC, Morowitz MJ, Banfield JF. Strain-resolved analysis of the hospital room and hospitalized infants reveals overlap between the human and room microbiome. *Under Revision*.

Brown CT, Davis-Richardson AG, Giongo A, Gano KA, Crabb DB, Mukherjee N, Casella G, Drew JC, Ilonen J, Knip M, et al. 2011. Gut microbiome metagenomics analysis suggests a functional model for the development of autoimmunity for type 1 diabetes ed. R.M. Roop. *PLoS ONE* **6**: e25792.

Brown CT, Olm MR, Thomas BC, Banfield JF. 2016. Measurement of bacterial replication rates in microbial communities. *Nat Biotechnol* **34**: 1256–1263.

Brown CT, Sharon I, Thomas BC, Castelle CJ, Morowitz MJ, Banfield JF. 2013. Genome resolved analysis of a premature infant gut microbial community reveals a *Varibaculum cambriense* genome and a shift towards fermentation-based metabolism during the third week of life. *Microbiome* **1**: 30.

Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. 2009. Bacterial community variation in human body habitats across space and time. *Science* **326**: 1694–1697.

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

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

Edgar RC. UBLAST. *drive5com*. http://drive5.com/usearch/ (Accessed July 8, 2014).

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**: 1575–1584.

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.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. 2012. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* **40**: D109–14.

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

Lathrop SK, Bloom SM, Rao SM, Nutsch K, Lio C-W, Santacruz N, Peterson DA,

Stappenbeck TS, Hsieh C-S. 2011. Peripheral education of the immune system by colonic commensal microbiota. *Nature* **478**: 250–254.

Lozupone C, Knight R. 2005. UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *Appl Environ Microbiol* **71**: 8228–8235.

Ma Z-Q, Dasari S, Chambers MC, Litton MD, Sobecki SM, Zimmerman LJ, Halvey PJ, Schilling B, Drake PM, Gibson BW, et al. 2009. IDPicker 2.0: Improved Protein Assembly with High Discrimination Peptide Identification Filtering. *J Proteome Res* **8**: 3872–3881.

Mai V, Young CM, Ukhanova M, Wang X, Sun Y, Casella G, Theriaque D, Li N, Sharma R, Hudak M, et al. 2011. Fecal microbiota in premature infants prior to necrotizing enterocolitis. *PLoS ONE* **6**: e20647–e20647.

Manzoni P. 2016. Clinical Benefits of Lactoferrin for Infants and Children. *J Pediatr* **173**: S43–S52.

Maslowski KM, Vieira AT, Ng A, Kranich J, Sierro F, Yu D, Schilter HC, Rolph MS, Mackay F, Artis D, et al. 2009. Regulation of inflammatory responses by gut microbiota and chemoattractant receptor GPR43. *Nature* **461**: 1282–1286.

Minoru Kanehisa SG. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28**: 27.

Morrow AL, Lagomarcino AJ, Schibler KR, Taft DH. 2013. Early microbial and metabolomic signatures predict later onset of necrotizing enterocolitis in preterm infants. *Microbiome* **1**: 13.

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**: 20–25.

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

Olm MR, Brown CT, Brooks B, Banfield JF. 2017. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J*.

Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**: 1.

Palmer C, Bik EM, DiGiulio DB, Relman DA, Brown PO. 2007. Development of the human infant intestinal microbiota. *PLoS Biol* **5**: e177.

Pammi M, Cope J, Tarr PI, Warner BB, Morrow AL, Mai V, Gregory KE, Kroll JS, McMurtry V, Ferris MJ, et al. 2017. Intestinal dysbiosis in preterm infants preceding

necrotizing enterocolitis: a systematic review and meta-analysis. *Microbiome* **5**: 31.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**: gr.186072.114–1055.

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

Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D, et al. 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**: 55–60.

Raghuveer TS, McGuire EM, Martin SM, Wagner BA, Rebouché CJ, Buettner GR, Widness JA. 2002. Lactoferrin in the Preterm Infants' Diet Attenuates Iron-Induced Oxidation Products. *Pediatric Research 2002 52:6* **52**: pr2002280–972.

Raveh-Sadka T, Firek B, Sharon I, Baker R, Brown CT, Thomas BC, Morowitz MJ, Banfield JF. 2016. Evidence for persistent and shared bacterial strains against a background of largely unique gut colonization in hospitalized premature infants. *ISME J*.

Raveh-Sadka T, Thomas BC, Singh A, Firek B, Brooks B, Castelle CJ, Sharon I, Baker R, Good M, Morowitz MJ, et al. 2015. Gut bacteria are rarely shared by co-hospitalized premature infants, regardless of necrotizing enterocolitis development. *Elife* **4**: –.

Richter M, Rossello-Mora R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. *PNAS* **106**: 19126–19131.

Robinson MD, McCarthy DJ, Smyth GK. 2009. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139–140.

Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, Banfield JF. 2012. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res* **23**: 111–120.

Stewart CJ, Embleton ND, Marrs ECL, Smith DP, Nelson A, Abdulkadir B, Skeath T, Petrosino JF, Perry JD, Berrington JE, et al. 2016. Temporal bacterial and metabolic development of the preterm gut reveals specific signatures in health and disease. *Microbiome* **4**: 67.

Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**: 1282–1288.

Tabb DL, Fernando CG, Chambers MC. 2007. MyriMatch: Highly Accurate Tandem

Mass Spectral Peptide Identification by Multivariate Hypergeometric Analysis. *J Proteome Res* **6**: 654–661.

The UniProt Consortium. 2015. UniProt: a hub for protein information. *Nucleic Acids Res* **43**: D204–D212.

Tibshirani R, Walther G, Hastie T. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**: 411–423.

Ultsch A. 2005. *ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM*.

Wittek P, Gao SC, Lim IS, Zhao L. 2013. Somoclu: An Efficient Parallel Library for Self-Organizing Maps.

Xavier RJ, Podolsky DK. 2007. Unravelling the pathogenesis of inflammatory bowel disease. *Nature* **448**: 427–434.

Xiong W, Abraham PE, Li Z, Pan C, Hettich RL. 2015. Microbial metaproteomics for characterizing the range of metabolic functions and activities of human gut microbiota eds. P. Jagtap, T. Griffin, and J. Armengaud. *Proteomics* **15**: 3424–3438.

Xiong W, Brown CT, Morowitz MJ, Banfield JF, Hettich RL. 2017. Genome-resolved metaproteomic characterization of preterm infant gut microbiota development reveals species-specific metabolic shifts and variabilities during early life. *Microbiome* **5**: 72.

Zaharia M, Bolosky WJ, Curtis K, Fox A, Patterson D, Shenker S, Stoica I, Karp RM, Sittler T. 2011. Faster and More Accurate Sequence Alignment with SNAP.