Supporting Information

La Rosa et al. 10.1073/pnas.1409497111

SI Materials and Methods

Human Subjects. The protocol under which this study was approved by the Washington University School of Medicine in St. Louis Human Research Protection Office was entitled "The Neonatal Microbiome and Prematurity" (protocol 201104267). All data were analyzed without personal identifiers. Research was conducted according to the principles in the Declaration of Helsinki. Subjects' families provided written informed consent before enrollment.

Sample Collection. All infants weighing ≤ 1,500 g at birth, who were expected to live beyond the first week of life and who were hospitalized in the Neonatal Intensive Care Unit (NICU) at St. Louis Children's Hospital (SLCH) were eligible for enrollment. All stools produced were collected and stored temporarily at 4 °C before being frozen at −80 °C until analyzed. In total, 922 stools collected on multiple days from 58 premature babies were included in this analysis.

The SLCH NICU operates under a single joint physician–nurse model. All attending physicians are neonatology board-eligible or board-certified members of the Division of Newborn Medicine of the Department of Pediatrics of the Washington University School of Medicine in St. Louis. The nursing staff consists of registered and advanced practice nurses, who function under a single administrative structure at SLCH. Faculty, fellows, and nurses have no non-NICU responsibilities. The house staff and fellows who staff the NICU are within the same single Accreditation Council for Graduate Medical Education program at Washington University School of Medicine in St. Louis. Written protocols for clinical care, including feeding, thermal support, and infection control policies are implemented unitwide after deliberation by a unit-based joint practice team composed of physicians and nursing staff. This circumscribed clinical facility therefore strives to deliver uniform care, thereby reducing provider-associated variables.

Samples were selected from subjects whose gestational ages matched those of infants in our NICU who had necrotizing enterocolitis, based on current and historic (1) data. We sought to systematically sequence all samples up to day of life (DOL) 30, and from every third sample collected after DOL 30. Adjustments were made to this sampling strategy if quantity was not sufficient (<500 mg for specimens prepared before sequencing before March 2012, and <250 mg thereafter, $n = 55$ samples), or if sample did not meet internal quality control standards of immediate freeze down upon arrival in laboratory (16 samples from six subjects whose specimens were obtained before October 27, 2009), or were not identified in the database at the time subject's samples were sent for sequencing ($n = 147$, median 1/subject, IQR 0–3). Of the 1,016 samples sequenced, 57 were removed for producing <1,000 reads and 36 failed in the sequencing process because of poor amplification. A final number of 922 samples were included in the analysis, with a median of 15 (IQR 10–20) samples per subject. Each sample is associated with a DOL computed as the difference between the collection time of the sample and the day of the birth of the subject from which the specimen was collected. In practice this computation is performed by calculating the difference in days between these two dates plus the difference in hours divided by 24, i.e., the total number of hours in a day.

DNA was extracted from stool, 16S rRNA genes were amplified by targeted PCR, and sequences generated via 454-Roche pyrosequencing were analyzed to characterize the bacterial content of the gut. Taxonomic identities and relative abundances of the resident bacteria were determined by comparing 16S sequences to those in the Ribosomal Database Project (http://rdp.cme.msu.edu).

Clinical Data. Clinical data from each subject were entered into RedCap (2) including demographic characteristics such as weight and gestational age at birth, characteristics of labor and delivery, and postnatal clinical data such as medication and dietary history during the interval of stool collection. Relevant subject data are deposited in database of Genotypes and Phenotypes (dbGaP) (www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000247.v2.p2). Infants were grouped into subcohorts based on three different gestational age ranges: ≤25, 26–28, and >28 wk (3–5).

DNA Sequencing, Quality Control, and Taxonomic Classification. Consistent with the Human Microbiome Project (HMP) protocol for the 16S rRNA gene survey of stool samples, the V3 to V5 hypervariable regions (V3–V5) of 16S rRNA genes was amplified from the metagenomic DNA using primers 357F and 926R (357F: 5′-CCTACGGGAGGCAGCAG-3′ and 926R: 5′-CCGTCAATTCMTTTRAGT-3′). The oligonucleotides containing the 16S primer sequences also contain an adaptor sequence as well as one of 96 tag sequences, which are unique to each sample. PCR was performed with the following conditions: initial denaturation: 95 °C for 2 min; annealing: 50 °C for 0.5 min; elongation: 72 °C for 5 min; and final elongation: 72 °C for 5 min; 30 cycles. After removing the excess primers and nucleotides, the tagged amplicons were pooled and sequenced on the Roche 454 Titanium pyrosequencing platform, one-half picotiter plate per 94 samples with two controls. This produced >500,000 reads of 400–450 bases/run.

Roche's GL Titanium software initially processes the raw sequencing data. One mismatch in primers (excluding degeneracies) and zero mismatches in barcodes are allowed while extracting the FASTAS and quality scores from the standard flowgram files. Barcodes and primers are removed from reads, which are then further processed by removing reads of low quality (average quality <35), <200 bp in length, and ambiguous codons (N's). Chimeric reads are removed using Chimera Slayer software (microbiomeutil.sourceforge.net). The high-quality, chimera-free reads are classified from phylum to genus level at the 0.5 threshold using the Ribosomal Database Project (RDP) Naive Bayesian Classifier version 2.5, training set 9. RDP data were organized in matrix form with taxa in rows and subjects in columns. The entries in the table were the number of reads for each taxon for each subject. Subsampling was performed when sequencing depth is a covariate in the analysis. The table was the input for statistical and other analyses.

Data Structure. RDP data were organized in a matrix form with each row representing a unique sample and each column representing taxa at the class level for P samples and K distinct taxa at class level (table below).

		Taxa				
Subject	Sample	1	2	. . .	K	Total
1	1	X_{111}	X_{112}	. . .	X_{11K}	X_{11*}
1	2	X_{121}	X_{122}	. . .	X_{12K}	X_{12*}
1	3	X_{131}	X_{132}	. . .	X_{13K}	X_{13*}
2	1	X_{211}	X_{212}	. . .	X_{21K}	X_{21*}
2	2	X_{221}	X_{222}	. . .	X_{22K}	X_{22*}
3	1	X_{311}	X_{312}	. . .	X_{31K}	X_{31*}
j	p	X_{jp1}	X_{jp2}	. . .	X_{jpK}	X_{jp*}
Total		X_{**1}	X_{**2}	. . .	X_{**K}	X_{***}

The entries in the table under taxa columns are the number of reads, or counts, of taxon by subject and sample. Entries designated X_{jpk} represent the number of sequences from subject j, sample p, and taxon k. For example, for subject 1, sample 1, the number of reads RDP matched to taxon 1 is X_{111}, to taxon 2 is X_{112}, etc. The far right column indicates the total number of reads for each subject by sample row denoted as X_{jp*}, where the * indicates the counts were summed over the taxa. Similarly, the total number of reads for each taxon is indicated in the bottom row where the summation of the counts is down the column. These are denoted by X_{**k}.

Exploratory Multivariate Analysis: Biplot Analysis. To visually explore the existing relationships between the samples and the bacterial taxa forming these samples, we constructed biplots (6) where samples are displayed as points and taxa are displayed as vectors. Points near the end of a taxon vector represent samples that have high prevalence of that taxon and points in the opposite part of the plot from the end of a taxon vector represent samples that have low prevalence of that taxon.

In this analysis the biplots were constructed using nonmetric multidimensional scaling (7) applied to a sample pairwise distance matrix. To portray the temporal progression of samples in the biplot, samples were color coded by either chronologic (DOL) or postconception (weeks) age of the infant at time of sample production.

We used the Bray–Curtis distance to quantify the difference in taxa composition between all pairs of samples i and j according to the following equation:

$$d(i,j) = \frac{\sum_{k=1}^{K} |X_{ik} - X_{jk}|}{\sum_{k=1}^{K} X_{ik} + X_{jk}},$$

where X_{ik} and X_{jk} are the abundances of class k in samples i and j, respectively, and K is total number of distinct taxa present in both samples. The Bray–Curtis dissimilarity index ranges between 0 and 1, where 0 means two samples share all of the taxa in identical abundances, and 1 means two samples share no taxa at all. To illustrate the Bray–Curtis measure we analyze three mock samples and calculate the corresponding pairwise Bray–Curtis distances (Fig. S3). The taxa abundance composition of each sample is displayed by bar graphs where sizes are proportional to the abundances of each taxa, and where colors identify unique taxa. Samples 1 and 2 have different taxa compositions and a Bray–Curtis distance = 0.92, and samples 1 and 3 are less dissimilar resulting in a Bray–Curtis distance = 0.16.

Ordination techniques, such as biplots, extract information on distances between objects (metagenomic samples in this case), and project or display them in a low (i.e., 2) dimensional plot. The goal of these techniques is to allow the investigator to identify patterns in their data that might not be apparent in the original portrayal. Ordination methods typically operate on the dissimilarity matrix D formed by all sample pairwise distances:

$$D = \begin{bmatrix} d(1,1) & d(1,2) & \cdots & d(1,P) \\ d(2,1) & d(2,2) & \cdots & d(2,P) \\ \vdots & \vdots & \ddots & \vdots \\ d(P,1) & d(P,2) & \cdots & d(P,P) \end{bmatrix},$$

where $d(i,j)$ is the distance (Bray–Curtis in this analysis) between sample i and j. Fig. S4 illustrates the nonmetric multidimensional scaling (NMDS) data flow, which begins with the data composition table, and which includes the taxa abundances of samples and finishes with the NMDS plot based on the dissimilarity matrix D of the samples. The resulting 2D NMDS plot portrays sample 2 further apart from samples 1 and 3 as the corresponding Bray–Curtis indices are large with $d(1,2) = 0.92$ and $d(2,3) = 0.8$. Similarly, samples 1 and 3 are separated by small distances [$d(1,3)$] compared with others [e.g., $d(2,3)$ or $d(P,3)$] because the Bray–Curtis index between these samples is small [$d(1,3) = 0.16$].

Time-Series Analysis of Taxon Abundances Using Mixed Models. To formally model trends in the data, we performed a post hoc analysis of repeated samples for subjects using mixed models (Table 1). Using this model we identified linear relationships (trends) between taxa abundance (Clostridia, Bacilli, Gammaproteobacteria separately) and day of life, antibiotic effect (total antibiotic use), route of delivery, total breast milk consumption, and timing of the sample (before or after study midpoint). Breast milk was coded as a categorical variable (0, 1, 2, 3 of breast milk corresponding to 0%, >0 but < 10%, 10–50%, and > 50% of the total for the entire length of stay). The effect of sex and open versus single room was insignificant and excluded from final models.

Repeated measurements for each subject were modeled assuming an autoregressive within subject covariance structure. Separate models were fit for the three gestational age groups. All analyses were performed using SAS (version 9.3) and R (version 3.0.1).

Metabolic Capacity of the Microbial Communities. We used a recently described computational approach known as "phylogenetic investigation of communities by reconstruction of unobserved states (PICRUSt)" to characterize the metabolic capacities of the microbial communities (8). PICRUSt identifies the pathways inferred by the particular taxa in a population of bacteria, weighted by their relative abundance. We made three orthogonal comparisons of the identified Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways in the three different gestational age subcohorts (<26, 26–28, and >28 wk). First, in each subcohort, KEGG pathways inferred from the group of first-in-life samples were compared with the last sample within the interval equivalent to 33–36 wk postconceptional age (one subject whose first sample was produced in the 33–36 wk interval was not included in this analysis). Second, KEGG pathways were compared between subcohorts, with analysis confined to the first sample obtained during the first 15 d of life (11 subjects who did not produce a specimen in these 15 d were not included). Third, KEGG pathways were compared between subcohorts, with analysis confined to the last sample obtained in the interval equivalent to 33–36 wk postconceptional age.

Permutational multivariate analysis of variance was used in these intra- and intersubcohort comparisons. Metastats, a statistical methods based on Fisher's exact test developed for the HMP study, was used to identify differentially abundant KEGG pathways and the associated q values (9). KEGG pathways were considered noteworthy if (i) their abundances differed by more than a factor of 2 between two compared groups, and (ii) the false discovery rate for that pathway was <0.10 (q value). This cut point for the q value was chosen after visual inspection of the distribution of the obtained values.

1. Gonzalez-Rivera R, Culverhouse RC, Hamvas A, Tarr PI, Warner BB (2011) The age of necrotizing enterocolitis onset: An application of Sartwell's incubation period model. *J Perinatol* 31(8):519–523.
2. Harris PA, et al. (2009) Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 42(2):377–381.
3. Adams-Chapman I, et al.; Eunice Kennedy Shriver National Institute of Child Health and Human Development Neonatal Research Network (2013) Ten-year review of major birth defects in VLBW infants. *Pediatrics* 132(1):49–61.
4. Boghossian NS, et al. (2013) Late-onset sepsis in very low birth weight infants from singleton and multiple-gestation births. *J Pediatr* 162(6):1120–1124, 1124.e1.

5. Finer NN, et al.; SUPPORT Study Group of the Eunice Kennedy Shriver NICHD Neonatal Research Network (2010) Early CPAP versus surfactant in extremely preterm infants. *N Engl J Med* 362(21):1970–1979.
6. Gower JC, Hand DJ (1996) *Biplots* (Chapman & Hall, London), p 277.
7. Cox TF, Cox AA (2000) *Multidimensional Scaling* (Chapman and Hall/CRC, Boca Raton, FL), 2nd Ed, p 328.
8. Langille MG, et al. (2013) Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* 31(9):814–821.
9. White JR, Nagarajan N, Pop M (2009) Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLOS Comput Biol* 5(4): e1000352.

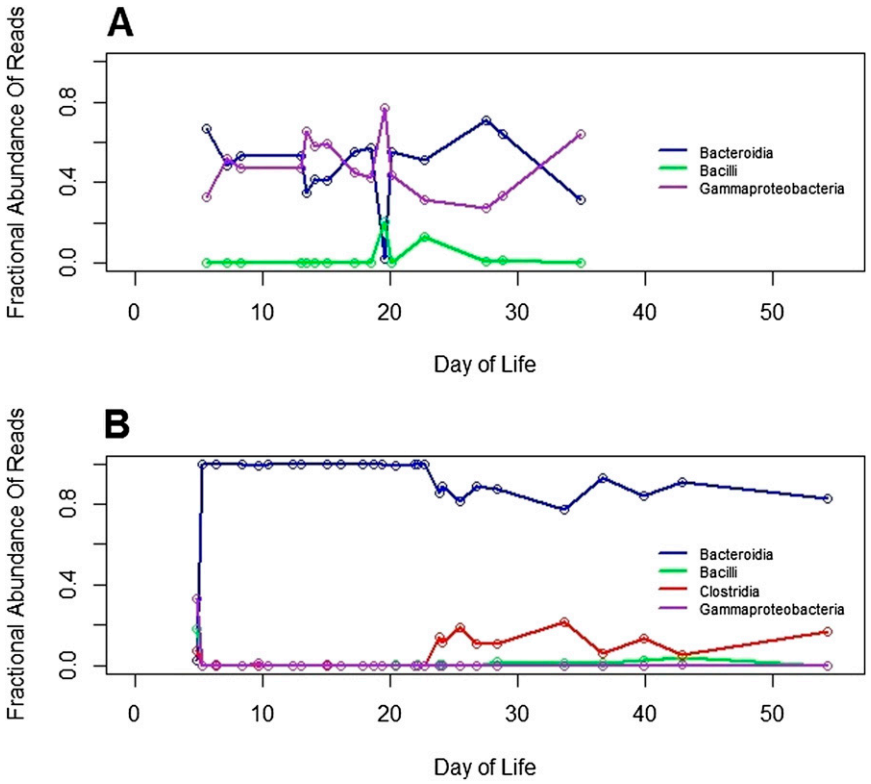


Fig. S1. Two subjects whose specimens largely consisted of bacteria from classes other than Bacilli, Gammaproteobacteria, and Clostridia. (*A*) Subject 35 is a girl born by vaginal delivery after 27 wk gestation (birth weight was 866 g) because of preterm labor. She received i.v. ampicillin for the first 10 d of life, and i.v. vancomycin and gentamicin for 8 d, starting on the third day of life. In the first 6 wk of life this subject received a combination of maternal or pasteurized donor human milk, or exclusive formula. (*B*) Subject 55 is a girl born by vaginal delivery after 29 wk gestation (birth weight was 1,390 g) because of preterm labor. She received i.v. ampicillin and gentamicin for 48 h after birth, and i.v. vancomycin and gentamicin for 48 h at 6 wk of life. Nutrition consisted of breast milk exclusively (first two weeks of life), followed by a combination of maternal breast milk and formula.

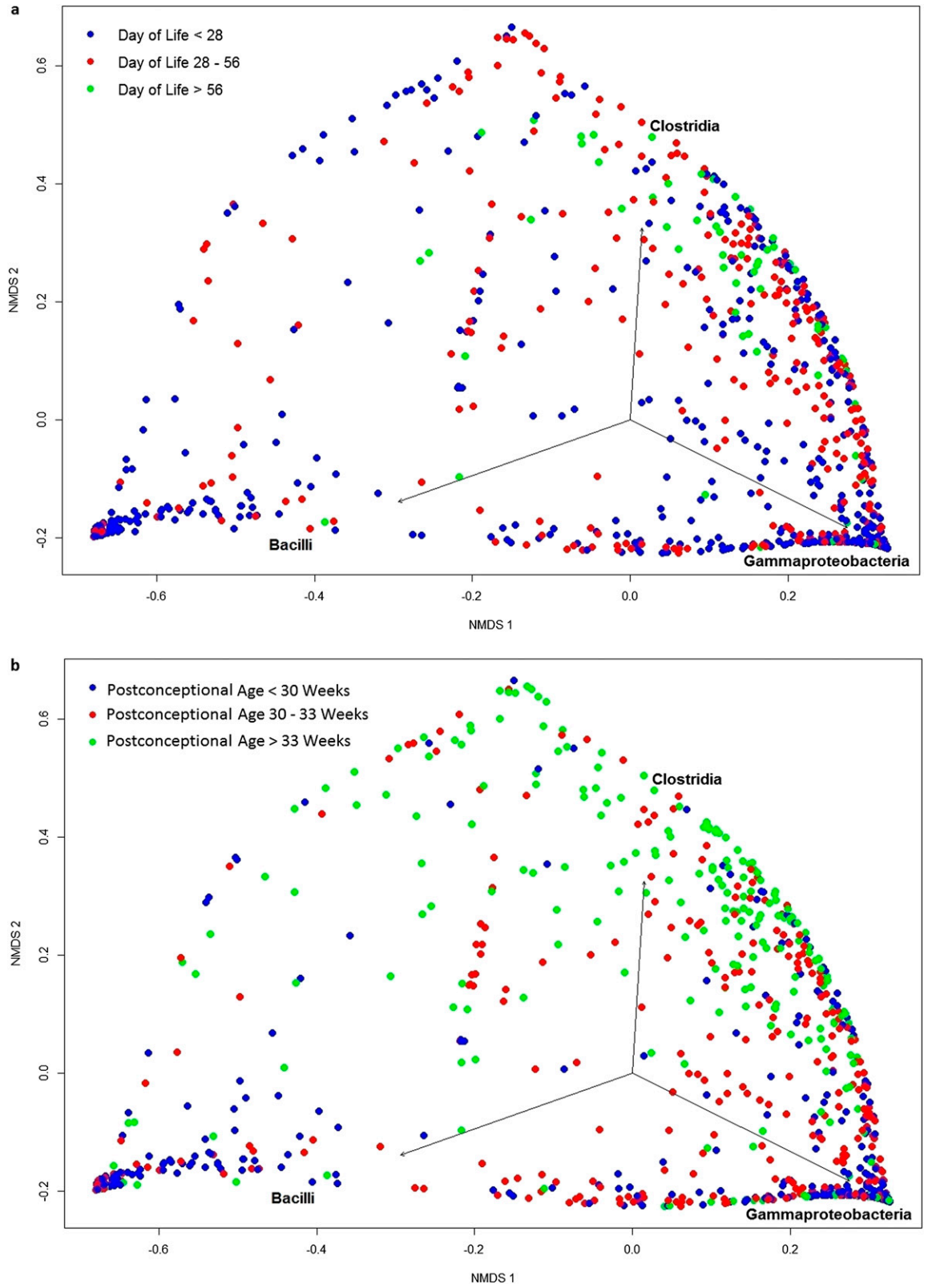


Fig. S2. Biplots representing bacterial class abundances. Nonmetric multidimensional scaling using the Bray–Curtis distance between the taxa composition for each specimen was used to construct a 2D portrayed of all samples. Arrows portray the directions in which the abundances of the three most predominant taxa tend to increase. (*A*) Each dot represents a single sample, color-coded by day of life of subject when sample was produced. (*B*) Each dot represents a single sample, color-coded by age postconception when sample was produced. The dots in the vicinity of the Clostridia arrowhead are more polychromatic in *A* than *B*, indicating that when samples are portrayed as originating from subjects grouped by postconceptional age, the convergence toward Clostridial abundance is more apparent.

Abundance	Taxon 1	Taxon 2	Taxon 3	Taxon 4	Taxon 5	Taxon 6
Sample 1	120	240	80	560	0	0
Sample 2	0	0	400	0	360	240
Sample 3	160	200	200	440	0	0

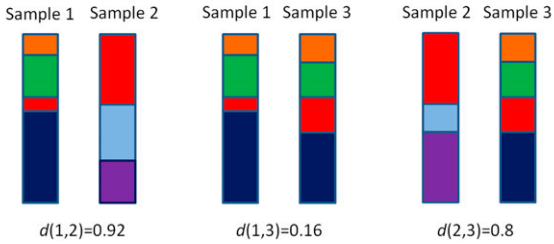


Fig. S3. Hypothetical taxa abundance for samples 1, 2, and 3. Bar graphs: Proportions of taxa abundance in each sample and their corresponding Bray–Curtis indices to quantify dissimilarities between sample pairs 1 and 2, 1 and 3, and 2 and 3.

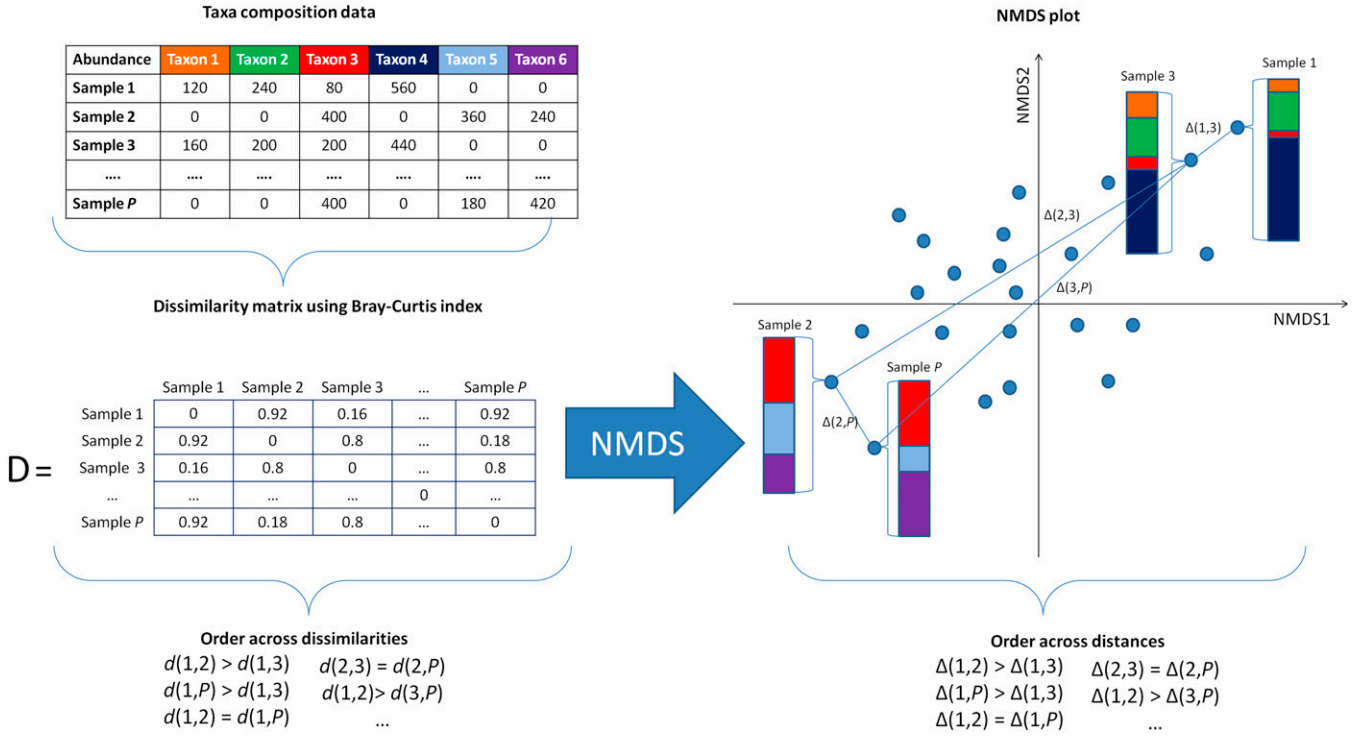
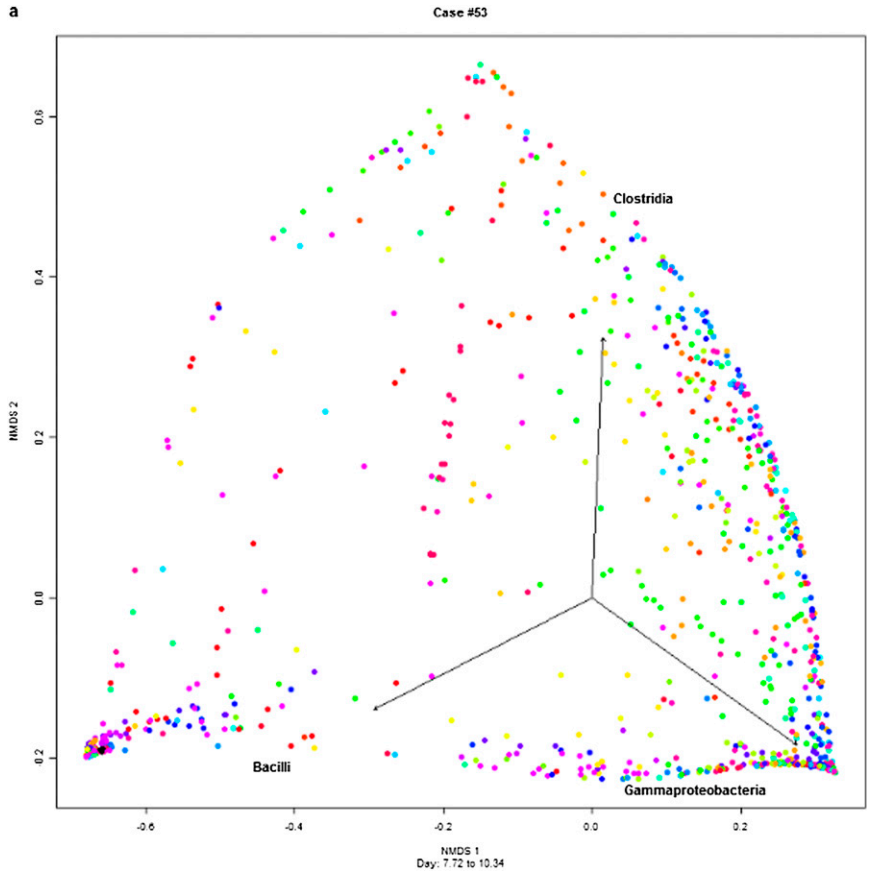


Fig. S4. Schematic of data flow processing for NMDS. The flow begins on the right with the computation of the dissimilarity matrix, using the taxa composition data and the Bray–Curtis index of dissimilarity. The flow ends on the left with the NMDS plot. Samples with the largest Bray–Curtis dissimilarity values are located on the NMDS plot at opposite extremes (e.g., samples 1 and 2, and samples 2 and 3), and samples with the smallest Bray–Curtis dissimilarity values are located much closer to each other.

Table S1. KEGG pathways identified in each of the three gestational age subcohorts that significantly differ in abundance between the first and last sampling intervals

Metabolic pathways	Median*		*P* value	*q* value
	Early	Late		
Subcohort born after <26 wk gestation				
KEGG pathways that increase between early and late sampling				
Lipopolysaccharide biosynthesis proteins	0.340	0.734	0.010	0.079
Lipopolysaccharide biosynthesis	0.194	0.444	0.008	0.068
Pertussis	0.068	0.248	0.008	0.068
Biosynthesis and biodegradation of secondary metabolites	0.098	0.197	0.003	0.037
Biosynthesis of siderophore group nonribosomal peptides	0.088	0.182	0.003	0.037
Nucleotide metabolism	0.088	0.181	0.002	0.030
Biosynthesis of ansamycins	0.050	0.134	0.001	0.020
Glycan biosynthesis and metabolism	0.046	0.123	0.015	0.099
Phenylpropanoid biosynthesis	0.041	0.114	0.012	0.088
Isoquinoline alkaloid biosynthesis	0.028	0.068	0.001	0.020
Proximal tubule bicarbonate reclamation	0.019	0.042	0.001	0.020
KEGG pathways that decrease between early and late sampling				
Secondary bile acid biosynthesis	0.014	0.003	0.001	0.020
Primary bile acid biosynthesis	0.015	0.003	0.001	0.020
Mineral absorption	0.018	0.000	0.001	0.020
Butirosin and neomycin biosynthesis	0.038	0.017	0.011	0.083
Bacterial invasion of epithelial cells	0.027	0.005	0.011	0.083
Carotenoid biosynthesis	0.039	0.001	0.001	0.020
Bisphenol degradation	0.066	0.017	0.003	0.037
Synthesis and degradation of ketone bodies	0.115	0.052	0.005	0.052
Bacterial toxins	0.144	0.047	0.004	0.044
Tetracycline biosynthesis	0.200	0.091	0.001	0.020
Naphthalene degradation	0.251	0.118	0.001	0.020
Staphylococcus aureus infection	0.228	0.005	0.001	0.020
Subcohort born after 26–28 wk gestation				
KEGG pathways that decrease between early and late sampling				
Bacterial toxins	0.093	0.034	0.008	0.082
Bisphenol degradation	0.056	0.012	0.004	0.049
Staphylococcus aureus infection	0.064	0.008	0.002	0.029
Subcohort born after >28 wk gestation				
KEGG pathways that decrease between early and late sampling				
Bisphenol degradation	0.066	0.018	0.004	0.049
Carotenoid biosynthesis	0.023	0.002	0.001	0.017
Staphylococcus aureus infection	0.103	0.007	0.002	0.029



Movie S1. Day-by-day progression of microbial content for three different subjects. These three cases exemplify the general trend in the progression from Bacilli to Gammaproteobacteria, to Clostridia, in an animation format used in prior studies of microbial populations over time (1, 2). Other general characteristics portrayed by these subjects are frequent abruptions in microbial content, and the slower rate of accumulation of Clostridia in the most premature subject, and the most rapid rate of accumulation of this bacterial class in the least premature subject. Arrows change at one second intervals, but their pace of appearance is not proportional to the days between samples. Numbers at bottom of each graph provide the days of life of each sample. Dots are color-coded, with unique shading for each subject. Examples A, B, and C represent subjects 53, 47, and 39, born after 25, 28, and 31 wk gestation, respectively.

Movie S1

1. Caporaso JG, et al. (2011) Moving pictures of the human microbiome. *Genome Biol* 12(5):R50.
2. Gerber GK, Onderdonk AB, Bry L (2012) Inferring dynamic signatures of microbes in complex host ecosystems. *PLOS Comput Biol* 8(8):e1002624.

Dataset S1. Details of each specimen and subject

Dataset S1

Each row represents one specimen. Columns designate, from left, subject number in this study, submitted subject-identifying number (dbGaP reference for subject), submitted sample-identifying number (dbGaP reference number for specimen and sequence), day of life specimen obtained (includes fraction of days expressed as decimals), gestational age at birth (weeks represented by number to the left of decimal point; days represented as 0.10 = 1 d, 0.2 = 2 d, 0.3 = 3 d, 0.4 = 4 d, 0.5 = 5 d, 0.6 = 6 d), postconceptional age (synonymous with postmenstrual age, which is equal to the sum of gestational age at birth plus day of life) on day sample was obtained, gender (0 = female, 1 = male), mode of birth (0 = vaginal delivery, 1 = Caesarian section), period of study during which sample was obtained (1 = before midpoint, 2 = after midpoint), room category (single = single room, open = multipatient room), milk (breast milk volume; 0 = 0%, 1 = <10%, 2 = 10–50%, 3 = >50% of enteral volume), days of antibiotics (total days prior to sample being obtained on which antibiotics were administered), and number of reads assigned to each class designated at top of column (including, in the rightmost column, reads that could not be classified). Episodes of bloodstream infections that occurred in members of this cohort are provided and described at bottom of column A.