Supporting Information

Morowitz et al. 10.1073/pnas.1010992108

SI Materials and Methods

DNA Extraction, PCR Amplification, and Sequencing of Amplicons and Metagenomic Libraries. Microbial DNA was isolated from frozen fecal samples using the QIAamp DNA Stool mini-Kit (Qiagen) with modifications (1). Bacterial 16S rRNA genes were amplified using the forward primer (5′- CTA TGC GCC TTG CCA GCC CGC TCA GNN NNN NNN NNA GAG TTT GAT CCT GGC TCA G-3′), which contained the 454 Life Sciences primer B sequence, the broadly conserved bacterial primer 8–27F, a unique 10-nt multiplex identifier (MID) used to tag each amplicon (designated by NNNNNNNNNN), and the reverse primer (5′-CGT ATC GCC TCC CTC GCG CCA TCA GGG ACT ACC AGG GTA TCT AA-3′), which contained the 454 Life Sciences primer A sequence and the broad-range bacterial primer 788–806R. PCR products were purified using AMPure Kits (Agencourt Bioscience). Preparation of a shotgun metagenomic library and pyrosequencing of both the genomic library and the 16S rRNA amplicons were performed on the 454 Genome Sequencer FLX-Titanium system at the W. M. Keck Center for Comparative and Functional Genomics at the University of Illinois at Urbana–Champaign, according to manufacturer's instructions (454 Life Sciences) (2). Signal processing and base calling were performed using the bundled 454 Data Analysis Software version 2.0.00.

16S rRNA gene sequences were processed using the QIIME software package (3) and removed from the analysis if they were <350 or >550 nt in length, contained >2 ambiguous bases, had a mean quality score <25, contained a homopolymer run exceeding 6 nt, or did not contain a primer and barcode sequence (Table S1*A*). Similar sequences were clustered into operational taxonomic units (OTUs) using UCLUST software (4) and minimum identities of 100 and 97%. The most abundant sequence was chosen to represent each OTU. Taxonomy was assigned to each unique sequence (i.e., representatives of OTUs picked at 100% identity) using the Ribosomal Database Project (RDP) classifier (5) with a minimum support threshold of 80% and the RDP taxonomic nomenclature. To confirm that the removal of numerous short reads did not taxonomically bias our results, we also attempted to assign taxonomy to each quality-filtered, unique, short sequence (50–350 nt) (Fig. S1 and Table S1*B*). Representatives of OTUs picked at 97% identity were aligned against the Greengenes core set (6) using PyNAST software (7) with a minimum alignment length of 150 and a minimum identity of 75%. The PH Lane mask was used to screen out hypervariable regions after alignment. A phylogenetic tree was inferred using FastTree software (8) and Kimura's two-parameter model. Good's coverage of OTUs picked at 97% identity was calculated using the full high-quality dataset (Table S1*C*). To facilitate comparisons among samples, OTU-based and phylogenetic α-diversity metrics were calculated using rarefied datasets: the number of sequences in the smallest sample ($n = 87$) was randomly drawn 10 times from each sample and the averages are reported. Unweighted and weighted UniFrac distances (i.e., phylogenetic beta diversity metrics) (9) were calculated between all pairs of samples. UniFrac-based sample clustering was performed using principle coordinates analysis (PCoA) and jackknifed hierarchical clustering [unweighted-pair group method with arithmetic mean (UPGMA)]. The statistical significance of the UniFrac-based sample clustering was tested using PERMANOVA in the PRIMER software package (10). Means and SDs are reported.

Taxonomic assignments within the family Enterobacteriaceae (which includes *Serratia*) inferred based on metagenomic analysis (Fig. 2) are better resolved than those deduced from the PCR-based 16S rRNA gene sequence analysis (Fig. 1*A*) because the unassembled, and thus much shorter, 16S rRNA reads, which are also highly conserved within this taxon, generated subthreshold genus-level assignments: these assignments "fall back" to the lowest taxonomic rank confidently assigned, in this case the family Enterobacteriaceae. Differences may also be because of PCR or other biases.

Fecal 16S rRNA gene sequences from previous studies were obtained directly from GenBank or provided by the authors and pooled with sequences from the present study. To reduce sequence-length variation among studies, sequences were aligned as described above, and then truncated to ∼500 nt between *Escherichia coli* positions 27 to 515 (the region where the studies' sequences overlap) using a custom mask. The sequences were then unaligned and mapped to reference OTUs and a reference phylogeny (both derived from the Greengenes database), as described previously (3) using UCLUST. Pairwise unweighted UniFrac distances were calculated and subjected to PCoA. Sample clustering results were relatively robust to minimum sequence identities spanning 85 to 97% and differences in the number of sequences per sample.

Metagenomic Data Analyses. Sequencing reads from the four libraries were coassembled using Newbler (GSassembler v. 2.0.01; Roche) using default parameters except for a 95% nucleotide identity and 40-nt minimum overlap requirement. Replicated reads were identified using a previously described protocol based on CD-HIT clustering (11) (> 95% identity, > five identical bases at the start of the read, no equal length requirement). Within each cluster, reads that shared the same start position on the assembled contigs were removed, except for the longest read. Using identical parameters, a second assembly was performed using this filtered dataset.

We annotated contigs larger than 1,500 bp with an in-house annotation pipeline using Prodigal gene calls (12), BLAST-based similarity searches (against NR, KEGG, UniRef 90, COG), and HMM-based functional domain recognition searches [Interproscan (13)]. Sequence bin assignments were based on a combination of manual assembly curation, blastn, blastp, GC%, sequencing depth, SNP density, and emergent self-organizing maps (eSOM) based on tetranucleotide frequency in combination with a K-means clustering of the temporal profiles of the reads of each contig. We executed the eSOM training algorithm using the parameters optimized by (14) using tetranucleotide frequencies calculated over 3,000-bp intervals of the large contigs, in combination with relevant reference genomes (including their plasmids) (Fig. S2).

Library affiliations of all reads in each contig were extracted using custom Ruby scripts. Numbers were normalized for each contig based on each library's total number of reads. Temporal profiles were grouped by K-means clustering [cluster, 10 clusters, 100 iterations, uncentered correlation similarity metric (15)]. Clustering was performed separately for fragments >1,500 bp and fragments between 500 and 1,500 bp. Small contig clusters were named based on similarity to the large contigs clusters when appropriate.

Final sequence bin assignment for large contigs was performed manually by reconciling the different sources of information. In case of ambiguity, contigs were assigned to a higher phylogenetic category (e.g., Enterobacteriaceae, Firmicutes). Contigs of virus and plasmid origin were identified based on boom-and-bust dy-

namics deduced from read temporal profiles, colocalization with plasmid/phage reference genome fragments on the eSOM map, and their functional annotations.

Contigs between 500 and 1,500 bp were assigned to genomic bins based on an approach similar to that used for the large contigs, except for the use of projection onto the eSOM map trained using the large contigs and reference genomes (assignment of fragment to a location on the map of a large fragment that is most similar to the projected small fragment). Because of the vague boundaries between most Enterobacteriaceae on the trained map, the combined eSOM-temporal profile information was only used for assigning *Pseudomonas*, *Enterococcus*, and *Staphylococcus* fragments. Contigs smaller than 500 nt that were not incorporated during manual assembly curation were not further analyzed.

Assemblies for the dominant bacterial, viral and plasmid populations were manually curated in Consed (16). The taxonomic affiliation of almost all final *Serratia* and *Citrobacter* contigs was confirmed based on the rRNA sequences on one or both ends. Obvious homopolymer errors in the consensus contig sequences were corrected before functional annotation. We used a custom Ruby script to identify and correct frame shifts due to homopolymer errors postannotation. Contigs and reads that matched the human genome (blastn e-value cutoff of $1e^{-35}$) were tallied and then removed from the dataset.

Strain-Resolved Analysis of *Citrobacter*. After manual assembly curation, each contig was viewed in its entirety in Consed to identify and correct clear homopolymer errors and to the extent possible, select the UC1CIT-i sequence as the reference. A few gaps were closed using reads from a few thousand additional sequence reads derived from the same libraries.

Next, each contig was imported from the .ace file into Strainer using the recompute alignment option. Strain sequence types were identified based on SNP patterns and separated in Strainer (17). In cases where read abundances did not clearly resolve the sequence variant type, choices were always made such that two rather than three strains groups were generated. Where the minor strain (UC1CIT-ii) was not represented by multiple linked reads, a read with high quality SNPs was chosen to represent the minor strain so long as the read had two or more separated SNPs that were clearly not in homopolymer regions. Cases where two independent reads have a single shared SNP were also chosen to represent the minor strain where it was not otherwise sampled. Reads present in the assembly in multiple copies, because of joining of Newbler-generated contigs, were ignored. Once the blocks with UC1CIT-ii–identified SNPs were created, a second .xml file was generated in which all of the blocks taken to represent the UC1CIT-ii strain were linked. For each contig, the list of reads that comprised the dominant (UC1CIT-i) and minor (UC1CIT-ii) strains was generated from the linked strain blocks. To verify that the linking of strain blocks into a minor strain appropriately represented the day's distribution, the day's distribution was calculated for just each strain variant sequence block over ~300 kb.

The sequence representative of the major strain (UC1CIT-i) was exported from Strainer and annotated. The UC1CIT-ii sequence was also exported, and used to calculate strain sequence identity (BLASTN), both when gaps were filled by UC1CIT-i sequence and when they were not. At some loci, additional sequence types with very high SNP density were considered to derive from low abundance strains or species and were excluded from the analysis. A custom Ruby script was used to count the day's distribution of the reads in each strain after removing redundant reads.

Regions of length divergence in intergenic regions were identified primarily because they terminated the automated *Citrobacter* assembly. Intergenic and flanking sequences for the two strains were reconstructed, compared, and flanking-gene positions identified. Secondary structure predictions for the identified intergenic regions used CentroidFold (http://www.ncrna.org/centroidfold). Similarities between intergenic sequences and previously published sRNA sequences were evaluated with BLAST searches against the sRNAMap database (18).

Modeling of *Citrobacter* Strain Growth Dynamics. We made use of a simplified model of interstrain competition within the colon assuming chemostat dynamics as proposed by Freter (19), and modified by Ballyk et al. (20). A first approximation, aimed at examining differential growth rate as the controlling factor of strain population dynamics, we assumed a constant growth rate across each separate time interval and no cell attachment to the gut wall (Fig. S3*D*, Eq. **1**). Evaluated colon residence times (3, 6, and 12 h) were adjusted downward based on studies in children between 4 and 15 y of age, indicating times between 12 and 84 h (21).

In a second approach, we incorporated the possibility of wall attachment and changes in growth rate over time, as a function of the *Citrobacter* carrying capacity saturation, and evaluated the presence of the remainder of the population as well. Equations according to Ballyk et al. (20), except for Eqs. **6** and **7** (Fig. S3, Eqs. **2**–**7**). Parameters were adjusted from Ballyk et al. (20) to be reasonable for a preterm infant colon and to fit the empirical data (Fig. S3). Although differential die-off because of a phage bloom could be integrated in this model as well, we evaluated this hypothesis based on the genomic data at hand. Instead of substrate-dependent growth rates, we made these a function of the carrying capacity for all *Citrobacter* of the system so that growth rates decrease as carrying capacity gets more saturated. Colon dimensions were 50 cm in length (22), 1-cm radius; the number of cells in 1 g of cell weight = 1.8×10^{12}; total cell concentration = 1.8×10^{10} cells/mL = 1×10^{-2} g/mL; dilution rate D = 0.0833 h^{-1} (12-h colon transit time); wall affinity constant $\alpha_{major} = 1 \times 10^{-3} hr^{-1}$, $\alpha_{minor} = 0.1\ h^{-1}$ (six and eight orders of magnitude above the Freter model); sloughing rate $\beta_{major} = \beta_{minor} = 0.01\ h^{-1}$; conversion factor δ = surface area/volume = $2\pi rl/\pi r^2 l = 2\ cm^{-1}$; maximum concentration of cells on the intestinal wall $w_{max} = 4.71 \times 10^{-3}$ g cell weight/cm^2 (three orders of magnitude larger than in the Freter model); maximum growth rates in lumen were set at two times and one times the growth rate calculated for the major strain in the simplified chemostat model at the lowest (3 h) transit time for the major and minor strain, respectively: $\mu_{max,u_1} = 0.68\ h^{-1}$, $\mu_{max,u_2} = 0.34\ h^{-1}$; the maximum growth rates for wall growth were set an order of magnitude smaller and equal for both strains: $\mu_{max,w_1} = \mu_{max,w_2} = 0.034\ h^{-1}$. Carrying capacity for *Citrobacter* ($C_{Citrobacter}$) was set to the sum of major and minor strain at the beginning of the simulation: $u_{1,0} = 1.63 \times 10^{-3}$ g/mL, $u_{2,0} = 1.07 \times 10^{-3}$ g/mL, and wall-attached growth was set to zero at the beginning of the simulation: $w_1 = w_2 = 0$ g/cm^2. This result implies that a sudden increase of the available colonization loci occurred around day 16. Initial input values were based on the strained *Citrobacter* data (Fig. 3), and initial wall-attached populations were set at zero. We assumed a total cell concentration of ~1.8·10^10 cells/mL (numbers on graph are in grams cell dry weight, based on the units used in the Freter model equations).

1. Zoetendal EG, et al. (2006) Isolation of DNA from bacterial samples of the human gastrointestinal tract. *Nat Protoc* 1:870–873.
2. Margulies M, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380.
3. Caporaso JG, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7:335–336.
4. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26:2460–2461.

5. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73:5261–5267.
6. DeSantis TZ, et al. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 72: 5069–5072.
7. Caporaso JG, et al. (2010) PyNAST: A flexible tool for aligning sequences to a template alignment. *Bioinformatics* 26:266–267.
8. Price MN, Dehal PS, Arkin AP (2009) FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* 26:1641–1650.
9. Lozupone C, Knight R (2005) UniFrac: A new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 71:8228–8235.
10. Clarke KR, Gorley RN (2006) PRIMER v6: User Manual/Tutorial (PRIMER-E, Plymouth, United Kingdom).
11. Gomez-Alvarez V, Teal TK, Schmidt TM (2009) Systematic artifacts in metagenomes from complex microbial communities. *ISME J* 3:1314–1317.
12. Hyatt D, et al. (2010) Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119.
13. Quevillon E, et al. (2005) InterProScan: Protein domains identifier. *Nucleic Acids Res* 33(Web Server issue):W116–W120.
14. Dick GJ, et al. (2009) Community-wide analysis of microbial genome sequence signatures. *Genome Biol* 10:R85.
15. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 95:14863–14868.
16. Gordon D, Abajian C, Green P (1998) Consed: A graphical tool for sequence finishing. *Genome Res* 8:195–202.
17. Eppley JM, Tyson GW, Getz WM, Banfield JF (2007) Strainer: Software for analysis of population variation in community genomic datasets. *BMC Bioinformatics* 8: 398.
18. Huang HY, et al. (2009) sRNAMap: Genomic maps for small non-coding RNAs, their regulators and their targets in microbial genomes. *Nucleic Acids Res* 37(Database issue):D150–D154.
19. Freter R (1983) *Mechanisms That Control the Microflora in the Large Intestine* (Academic Press, New York).
20. Ballyk MM, Jones DA, Smith HL (2001) Microbial competition in reactors with wall attachment. *Microb Ecol* 41:210–221.
21. Wagener S, Shankar KR, Turnock RR, Lamont GL, Baillie CT (2004) Colonic transit time—What is normal? *J Pediatr Surg* 39:166–169, discussion 166–169.
22. Touloukian RJ, Smith GJ (1983) Normal intestinal length in preterm infants. *J Pediatr Surg* 18:720–723.

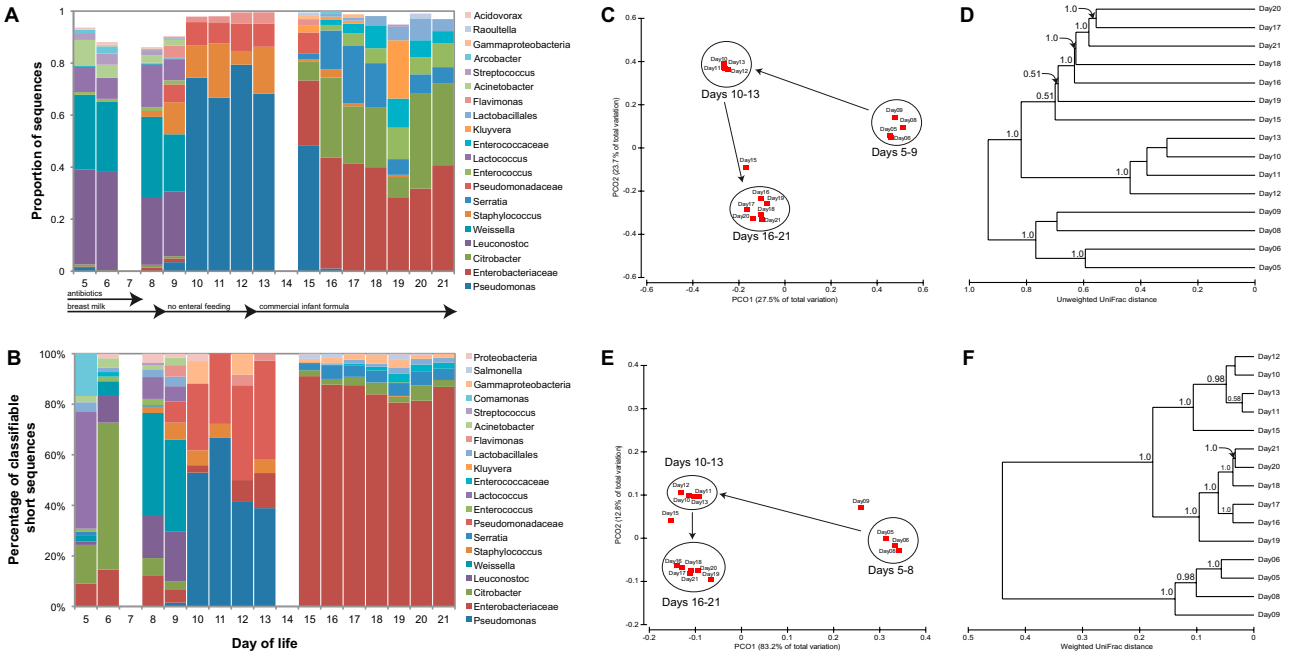

Fig. S1. Multiple stable taxonomic profiles exhibited by a premature infant's gastrointestinal microbiota. (*A*) Relative abundances of the 20 most abundant bacterial taxa found in 15 fecal microbiota samples collected approximately daily between 5 and 21 d of life. Only the high-quality sequences shown in Table S1*A* were used for this analysis. (*B*) Relative abundances of most of the taxa shown in *A* among the classifiable short (50–350 nt) sequences screened out of our analysis. Sequences were classified to the highest taxonomic level to which they could be confidently assigned using the RDP classification algorithm and taxonomic hierarchy. The results show large-scale shifts in the proportional abundances of the dominant taxa around days 9 and 15, and that our length-based sequence screen was not taxonomically biased. (*C* and *E*) Principle coordinates analysis and (*D* and *F*) hierarchical clustering of 15 fecal microbiota samples collected approximately daily between 5 and 21 d of life. Samples were compared using a phylogenetic measure of differences in overall bacterial community membership (unweighted UniFrac) (*C* and *D*) and a similar measure that also accounts for relative abundance (weighted UniFrac) (*E* and *F*). The percentage of the variation explained by the plotted principle coordinates (PCO1 and PCO2) is indicated on the axes (*C* and *E*). Robustness of UPGMA clusters was assessed using jackknifing and shown if >50% (*D* and *F*). Only the high quality sequences shown in Table S1*A* were used for this analysis. The results show large-scale adjustments in bacterial community membership and structure around days 9 and 15.

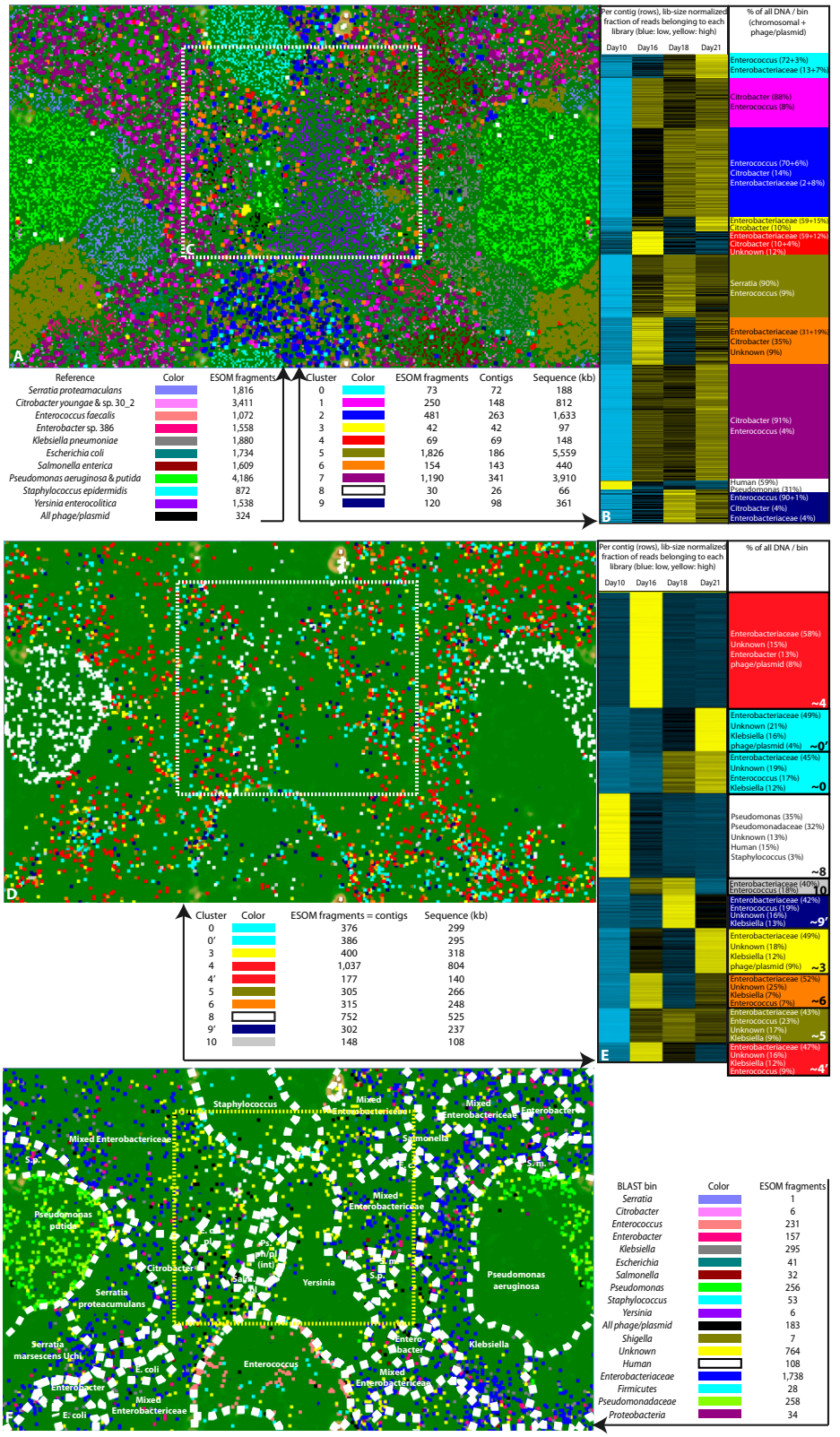

Fig. S2. Emergent self-organizing map based on tetranucleotidesignatures. (*A*) The eSOM map trained by using genomic data of both isolates (small squares) as well as the large contigs (> 1,500 bp) resulting from the uncurated Newbler assembly of the preterm infant metagenomic data (bold squares). The map is continuous from top to bottom and side-to-side and based on a 3-kb sequence window size (each square is a sequence fragment). Reference genomes were selected based on BLAST analysis of the large contigs and colored as indicated in the legend. *Citrobacter* sp. 30_2 (GG657366-83), *Citrobacter youngae* ATCC29220(GG730299-308),*Enterobacter* sp. 638 (CP000653-4), Enterobacteriophage ES18 (AY736146), Enterobacteriophage N15 (AF064539), *Enterococcus faecalis* V583 (AE016830-3), *Escherichia coli* UMN026 (CU928148-9,63), *Klebsiella pneumonia* 342 (CP000964-6), *Pseudomonas aeruginosa* LESB58 (FM209186), *Pseudomonas putida* F1 (CP000712), *Salmonella enterica* subsp. *enterica* serovarNewportstr. SL254 (CP000604,1112-3), *Serratia proteamaculans* 568

Legend continued on following page

(CP000826-7), *Staphylococcus epidermidis* RP62A (CP000028-9), *Yersinia enterocolitica* subsp. *enterocolitica* 8081 (AM286415-6). Squares located more closely to each other are more similar in their tetranucleotide content, unless the underlying topography (U-Matrix), which is colored from green over brown to white (low to high elevation), indicates more separation. Although different taxonomic groups typically cluster separately, significant admixing is observed for most Enterobacteriaceae genomes, although a subset of each reference genome's fragments also clusters in a specific region on the eSOM. Phages and plasmids of the reference genomes were all grouped together (black squares). The large contigs data points are colored by their temporal profiles (see *B*). (*B*) K-means clustering of the large contigs based on the library size-normalized frequency data of read origins (day 10, 16, 18, or 21 metagenomic library). Blocks next to the clustering representation are colored corresponding to the colors in the legend and include the bin names to which most contigs in the respective cluster belong as well as the fraction of all sequence in the cluster represented by these bins. (*C*) Region on the map enriched in phage/plasmid sequences, based on reference genomes and the enrichment of contigs with boom-and-bust dynamics in *B*. (*D*) Projection of contigs between 500 and 1,500 bp resulting from the uncurated Newbler assembly of the preterm infant metagenomic data onto the trained eSOM map of Fig S3. Data points are colored by their temporal profiles (see *E*). (*E*) K-means clustering of the 500 to 1,500 bpcontigs based on the library size-normalized frequency data of read origins (day 10, 16, 18, or 21 metagenomic library). Blocks next to the clustering representation are colored corresponding to the colors in the legend and include the bin names (based on BLAST) to which most contigs in the respective cluster belong as well as the fraction of all sequence in the cluster represented by these bins. Correspondence to the clusters in *A* is indicated as well. Denotation of 0′, 4′, and 9′ indicates the temporal distributions in these clusters are variants of those in 0, 4, and 9. (*F*) Same as *D*, except for color overlay, which is based on the BLAST-based binning information. For comparison purposes, we delineated and identified by their taxonomic group the different sectors of the map based on the analyses in *A* to *C*. In addition, the region identified as enriched in phage/plasmid sequences in *A* has been marked in both *D* and *F* as well.

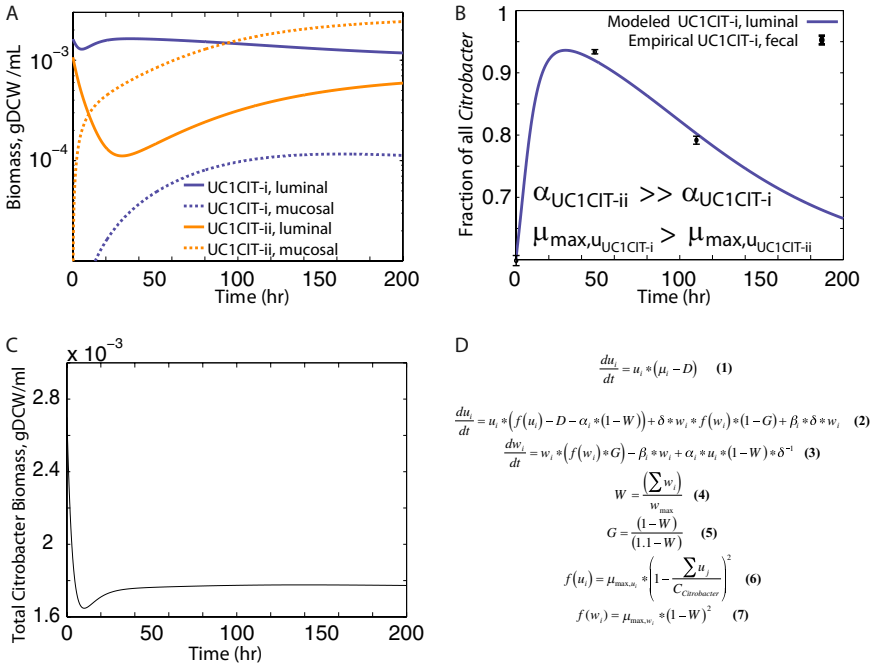

Fig. S3. Modeling of *Citrobacter* UC1CIT-i and -ii strain population dynamics. (*A*) Matlab simulation during the first 200 h (time 0 = day 16 sampling) of the luminal and wall-attached populations, the UC1CIT-ii population having a higher affinity for the wall, but the UC1CIT-i population has a higher growth rate. (*B*) UC1CIT-i dynamics as a fraction of all *Citrobacter* cells and the empirical data with SE across the evaluated contigs. (*C*) As seen in our data, an initial drop and then recovery of the total *Citrobacter* population occurs, although the current model does not restore the numbers as seen in the empirical data. These dynamics were achieved by setting a much higher affinity for wall colonization of the UC1CIT-ii compared with the UC1CIT-i strain (and assuming there has been a sudden increase of wall sites available; for example, by increased mucus production). In addition, to avoid too rapid washout of the UC1CIT-i strain, we set its maximum growth rate in the lumen/feces to be double that of UC1CIT-ii. In chemostat models with wall attachment, all else being equal, the more proficient wall colonizer will eventually outcompete the less proficient strain. In our model we also assumed equal maximum growth rates for wall attached cells for both strains, but significantly lower than their growth rates in the lumen. Sloughing rates are also set to be equal for both (nonspecific process).

A Intergenic variant case 4

Flanking genes in UC1 genome: arginine transporter permease subunit ArtM (+)
arginine-binding periplasmic protein 2 (+)

Best BLASTN hits against sRNAMap: major variant: E coli C0664, e = 0.08
minor variant: E coli C0362, e = $4e^{-07}$

Pairwise alignment of C0362 and intergenic sequence from MAJOR variant:

```
Score = 790
Length of alignment = 206
Sequence major :  1 - 228 (Sequence length = 228)
Sequence C0362 :  1 - 316 (Sequence length = 316)

major AATCGGTATTTGTGCCTTTGTAGGTCGGATAAGGTCTAACACCGCCATCCGAAAAATGTGCATAAG
      ||| | |    |||    |||||| | ||||||  |    | || |||| |  ||   ||| |  |
C0362 AATAGATTGCAGTGAACGTGTAGGCCTGATAAG--CGT--AGCG-CATCAGGCAATGTTGCGTTTG

major -CAAAAAATACAAA----GACGGACAA-----CAACCTAA-ATTGT---C--CGT---CTTTTTTT
       ||  |  | ||||    |  | | ||     ||   | | |||||   |  |||   |||| ||
C0362 TCATCAGTTTCAAATGGCGCTGTAAAAGGCGTCATTTTCATATTGTAGACAACGTAGGCTTTGTTC

major ATGCC--ATTAA-----AAATATTTAATC---ATTTTTATTGCAT-ATAAATTCATTAAATGGCA-
      |||||  ||        ||    |||  |   |||    |||  | || | |||  ||||||||
C0362 ATGCCGGATGCGGCGTGAACGCCTTATCCGGCATTGCGTTTG--TCATCAGTTC--TAAATGGCGC

major -TTGTTAA
       || | ||
C0362 TTTATAAA

Percentage ID = 52.43
```

Pairwise alignment of C0362 and intergenic sequence from MINOR variant:

```
Score = 1140
Length of alignment = 264
Sequence minor :  1 - 262 (Sequence length = 262)
Sequence C0362 :  1 - 316 (Sequence length = 316)

minor GCATATGCCTGATG--GCGCT-ACGC-TTATCAGGCCT----ACGGTTCA-----------TGCA
      |   ||||| ||||  |||   |||| ||||| ||| |    ||  ||||           ||||
C0362 GTTCATGCCGGATGCGGCGTGAACGCCTTATCCGGCATGAAAACCCTTCAAATCCAATAGATTGCA

minor CCTTT--TGTAGGCCGGATAAGGTGCTAGCACCACCATCCGGCAAATATGC-------AT-AAATT
            |||||||| ||||||      | | |  |||| |||||   |||       || |  ||
C0362 GTGAACGTGTAGGCCTGATAAG-----CGTAGC-GCATCAGGCAATGTTGCGTTTGTCATCAGTTT

minor AAAATAA-----TAAAGACGGACAACAACCTAAATTGT---C--CGT---CTTTTTTATGCC--A
       ||||       ||||    | ||    |  | |||||   |  |||   |||| || |||||  |
C0362 CAAATGGCGCTGTAAAAGGCGTCATTTTC--ATATTGTAGACAACGTAGGCTTTGTTCATGCCGGA

minor TTAA-----AAATATTTAATC---ATTTTTATTGCAT-ATAAATTCATTAAATGGCA--TTGTTAA
      |         ||    |||  |   |||    |||  | || | |||  ||||||||   || | ||
C0362 TGCGGCGTGAACGCCTTATCCGGCATTGCGTTTG--TCATCAGTTC--TAAATGGCGCTTTATAAA

Percentage ID = 51.89
```

B Intergenic variant case 5

Flanking genes in UC1 genome: NAD-dependent epimerase/dehydratase (+)
N-acetylmuramoyl-L-alanine amidase (-)

Best BLASTN hits against sRNAMap: major variant: E coli C0664, e = 0.055
minor variant: E coli C0664, e = $9e^{-09}$

Pairwise alignment of C0664 and intergenic sequence from MAJOR variant:

```
Score = 590
Length of alignment = 53
Sequence major :  1 - 48 (Sequence length = 48)
Sequence C0664 :  1 - 113 (Sequence length = 113)

major CCATTGCCGGATGGCGGCGCAAGCGCC--ATCAGGCATTGGTATTC---TGCG
        |  |||||||   ||||    ||||  ||| |||| |||| | |    |||
C0664 T-AG-GCCGGAT-AAGGCGTTTACGCCGCATCCGGCAATGGTGTCCAAATGCA

Percentage ID = 60.38
```

Pairwise alignment of C0664 and intergenic sequence from MINOR variant:

```
Score = 1460
Length of alignment = 114
Sequence minor :  1 - 105 (Sequence length = 105)
Sequence C0664 :  1 - 113 (Sequence length = 113)

minor CCATTGCCGGATGGCGGCGCAAGCGC--CTTATCCGGCCTACAAAATCCAGCCTAAATTAGCCGTA
        | || ||||| ||| |||  ||||  |||||||| |||||         |    |    |||
C0664 A-AATGTCGGAT-GCGACGCTGGCGCGTCTTATCCGACCTAC--------GGGGACGC-ATGTGTA

minor GGCCTGATAAG-CG---AAGCACCATCAGGCATTGGTATTC---TGCG
      |||| |||||| ||   | ||  |||| |||| |||| | |   |||
C0664 GGCCGGATAAGGCGTTTACGCCGCATCCGGCAATGGTGTCCAAATGCA

Percentage ID = 60.53
```

Fig. S4. (*A* and *B*) Alignment of known sRNA sequences with strain-resolved intergenic sequences in the UC1 genome. BLASTN analysis of all intergenic sequences in Table S10 in Dataset S2 was performed against the sRNAMap database (1). BLASTN hits were further investigated by aligning the published nucleotide sequences of sRNA candidates with the corresponding UC1 intergenic sequences using MUSCLE (2). Selected examples of these alignments are shown here to illustrate the effect of intergenic sequence variation on alignment with known intergenic sRNAs.

1. Huang HY, et al. (2009) sRNAMap: Genomic maps for small non-coding RNAs, their regulators and their targets in microbial genomes. *Nucleic Acids Res* 37(Database issue):D150–D154.
2. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.

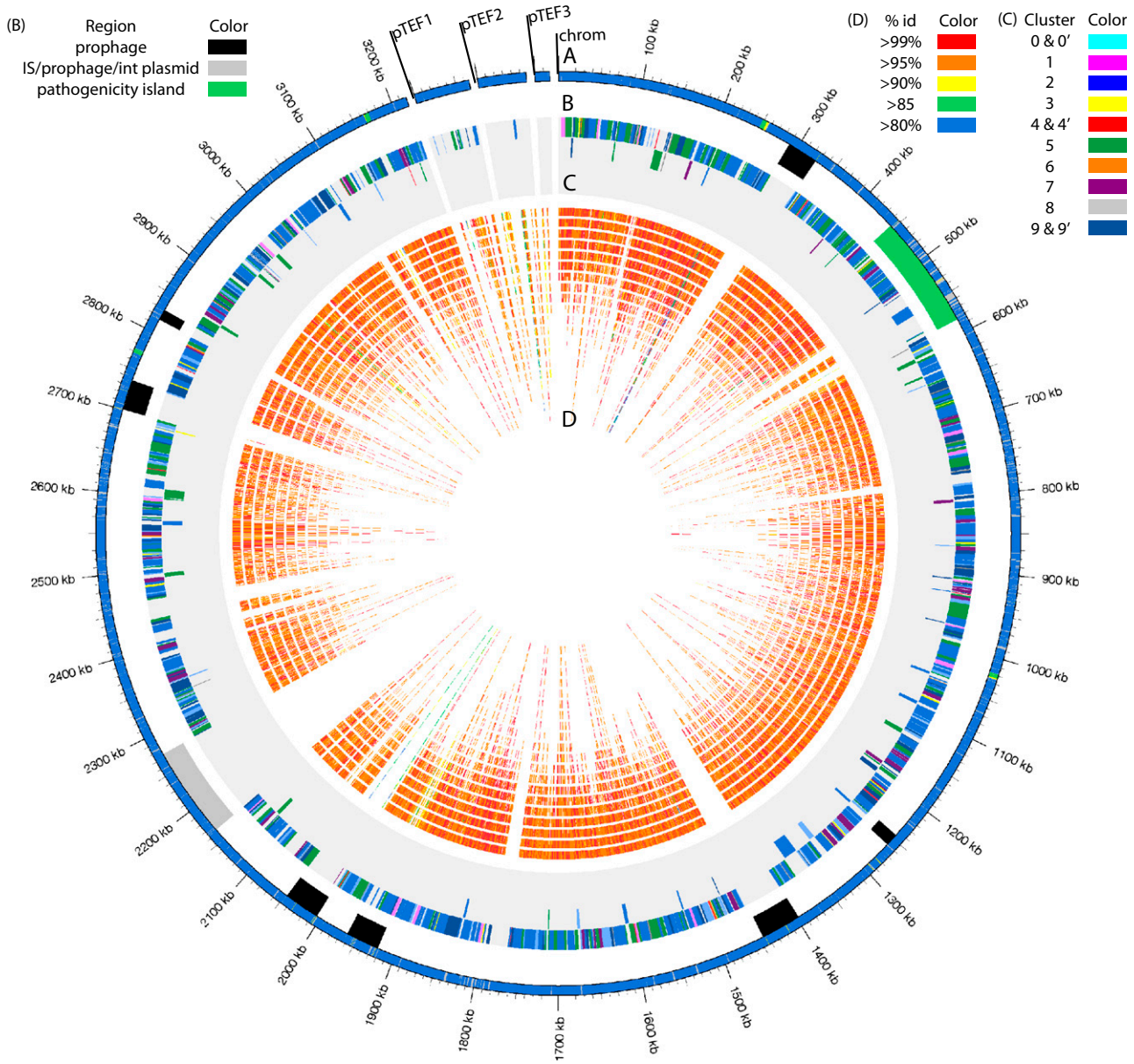

Fig. S5. Contig and read recruitment to the *Enterococcus faecalis* V583 reference genome. (*A*) Outer circle represents the V583 genome, including its three plasmids (pTEF1-3); blue, protein coding sequence; green, rRNA; yellow, tRNA. (*B*) Black bars indicate the prophage regions identified in the V583 genome, most of which are missing, but other prophage seem to be present in the preterm infant's *E. faecalis* genome. The gray bar marks a genomic island of mixed origin (EF2240-350), which is absent in the preterm infant *E. faecalis* population, that contains vancomycin resistance genes (EF2293-300), a bacteriocin (EF2314), and a cluster of sugar uptake and metabolism genes (EF2257-73). The green bar delineates the pathogenicity island, which is present except for EF0591-611 and EF0562-74, which are regions in the reference genome with a high concentration of pseudogenes, hypothetical proteins, and an operon encoding a potassium-transporting ATPase. (*C*) Tiles within the gray background area represent *Enterococcus* bin contigs (>500 bp) aligned to the V583 genome (MEGABLAST parameters –e1e^{-25} –N2 –t 18 –W 11 –A 50 –gF –v 1 –b 1). Partial overlap of these alignments forces the tile to the next line. Colors are based on the temporal clusters (see legend and Figs. S3 and S4). (*D*) Tiles represent reads that aligned to the V583 genome (BLASTN parameters –e 1e^{-35}) colored by percent-nucleotide identity. Partial overlap of these alignments forces the tile to the next line.

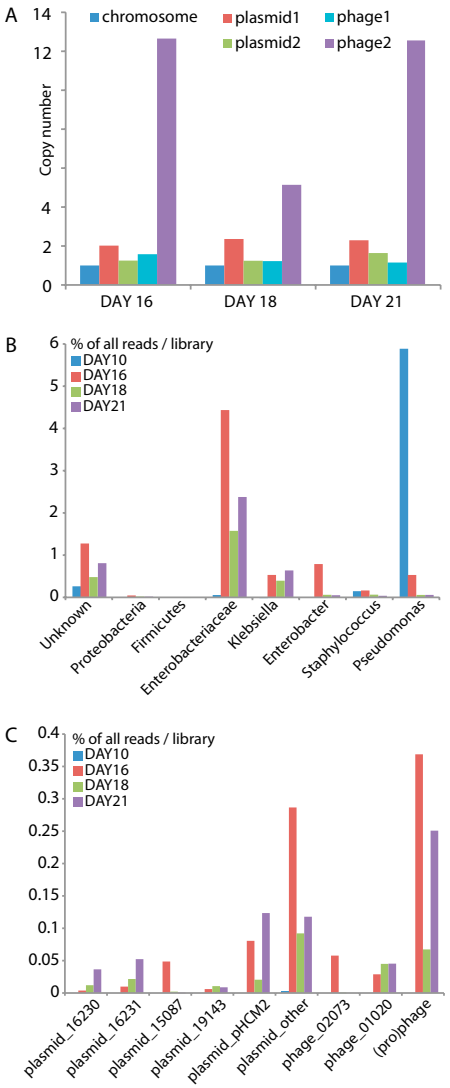

Fig. S6. Plasmid, phage, and potential host population dynamics. (*A*) *E. faecalis* plasmid and phage dynamics. The number of reads present in the contigs of each *Enterococcus* subbin (chromosome, plasmids, phages) was normalized by the subbin sequence length relative to the chromosome bin sequence length. This process was performed separately for the metagenomic libraries from days 16, 18, and 21. The ratio of this normalized value to the normalized number of chromosome reads for each considered day provides an estimate of the copy number for each replicon relative to the chromosome as it varies over time. (*B*) Distribution of reads across the minor population bins and the dynamics of these "populations" over time. Affiliation to a specific organism bin (rather than higher taxonomic groups) was based on stringent blast cutoffs (> 90% identity across > 90% of the contig or > 90% identity across > 90% of all proteins identified on the contig). (*C*) Similar analysis to that in *B* for the contigs identified to be of (pro)phage or plasmid origin. Phage and plasmids for which a complete sequence was available were given a number corresponding to their curated contig in the assembly except for pHCM2, which is a set of contigs similar to the pHCM2 plasmid of *Salmonella* (GenBank: AL513384). Particular correspondence can be observed between the dynamics of plasmid_15087, phage_02073, and *Enterobacter* in *B*. The pHCM2-likeplasmid is likely a *Klebsiella* plasmid based on similar temporal dynamics. Also notable is the distribution of plasmid_other, (pro)phage and the Enterobacteriaceae sequence bin, which all reflect the UC1CIT-ii strain dynamics (Fig. 2). Presumably, these bins contain some of the UC1CIT-ii strain contigs that were not anchored to the major strain path because of limited overall sequence coverage of the UC1CIT-ii strain.

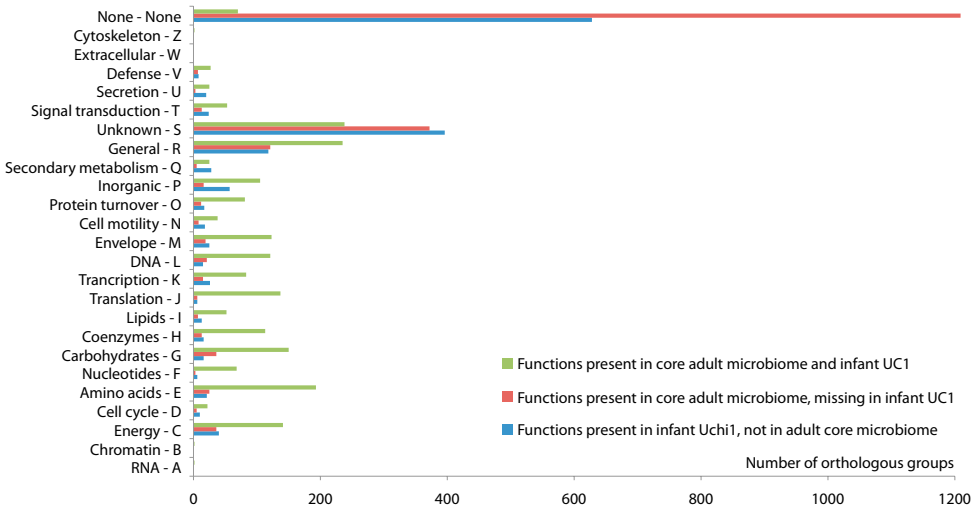

Fig. S7. Comparison of individual preterm infant microbiome functions to those identified in the adult core microbiome. The 17,487 annotated proteins from all contigs> 500 bp were compared with the orthologous groups database used by Qin et al. (1) (eggNOG version 1) using BLASTP (cutoff $1e^{-05}$). Out of all annotated proteins, 13,668 proteins were matched to 3,611 unique clusters of orthologous groups of proteins. These 3,611 clusters that were detected in the individual preterm infant gut communities were contrasted to the 4,055 preexisting orthologous groups identified as the core human microbiome in a study of more than 100 individuals (1). Clusters that were shared, absent,and uniquely present in the infant data were grouped in broad classes [A(RNA) – >Z(Cytoskeleton) as well as those clusters that have not been grouped into a broad class (None)].

1. Qin J, et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464:59–65.

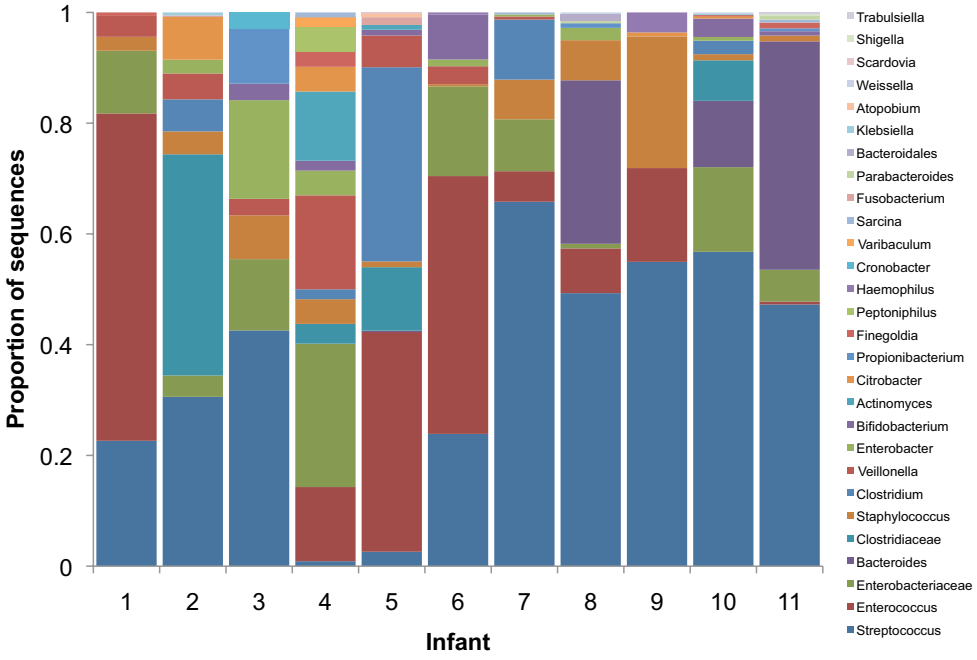

Fig. S8. Taxonomic profiles of gut microbes from hospitalized premature infants receiving either breast milk or infant formula. For a related project involving measurement of gut microbial metabolites in milk-fed and formula-fed infants, fecal samples were collected from five milk-fed and six formula-fed premature infants without major comorbidities. Age at sample collection ranged from 10 to 46 d. Extraction of microbial DNA and analysis of 16S rRNA sequences were performed as described in the text and *SI Materials and Methods*. Shown here is the relative abundance of the dominant bacterial taxa in the 11 individuals. Sequences were classified to the highest taxonomic level to which they could be confidently assigned. The similarity of these gut microbial communities to the communities studied in the main text is represented in Fig. 1*B*.

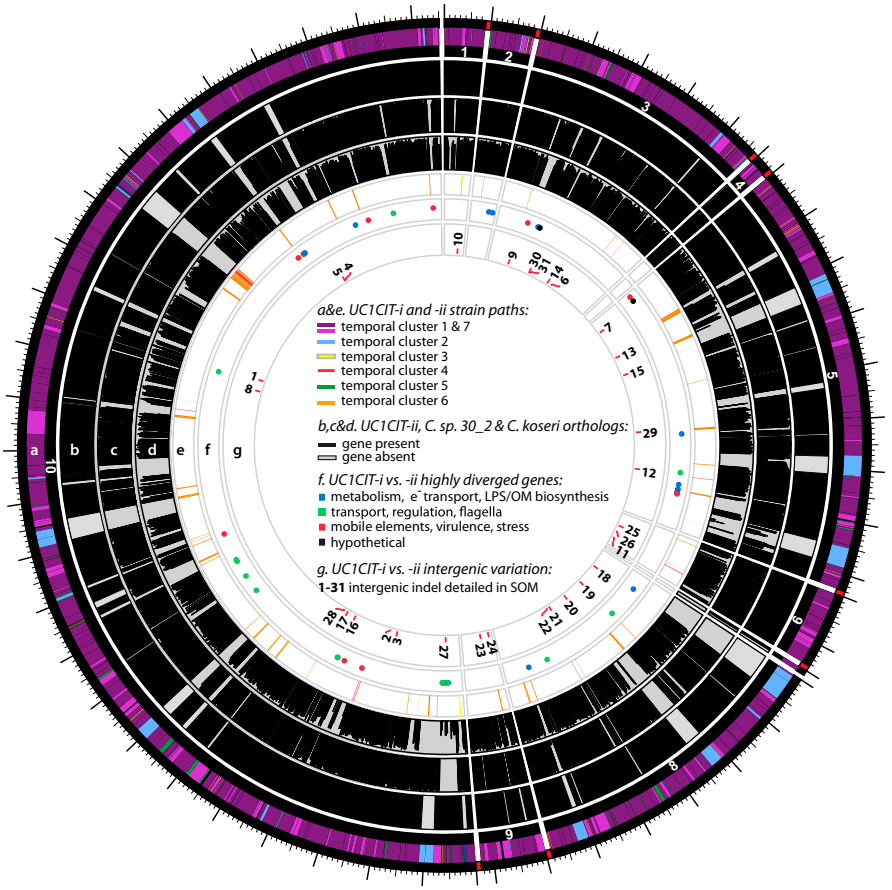

Fig. S9. Copy of Fig. 4*A* enlarged to show detail.

Table S1*A*. Quality screening of Titanium pyrosequencing reads of amplified bacterial 16S rRNA gene sequences

		Quality screens						
Sample	Raw sequences	Length: <350, >550 nt	Ambig. bases: >2	Mean qual. score: <25	Homopol. run: >6	Primer mismatch	High quality sequences	Mean length (nt)
Day 05	1,232	720	35	53	0	45	379	480
Day 06	9,296	8,639	43	72	2	45	495	471
Day 08	3,563	1,206	974	340	18	51	974	475
Day 09	2,483	658	703	242	12	34	834	470
Day 10	732	104	12	16	0	80	520	502
Day 11	706	107	12	25	0	60	502	503
Day 12	1,058	375	10	31	0	35	607	500
Day 13	603	112	10	21	1	21	438	503
Day 15	1,525	1,287	11	19	0	13	195	497
Day 16	1,089	958	8	15	0	14	94	491
Day 17	910	783	16	18	0	11	82	498
Day 18	1,158	1,000	19	17	0	17	105	494
Day 19	975	807	15	26	0	11	116	493
Day 20	834	664	20	20	0	23	107	496
Day 21	961	762	21	31	1	16	130	500

Table S1*B*. Number of short reads, some of which were used to check for taxon bias in length-based screen (Fig. S1*B*)

Sample	Raw sequences	Length: <50 nt	Length*: 50–350 nt	Mean length of 50–350 class (nt)	Classifiable (≥phylum)
Day 05	1,232	491	141	143	88
Day 06	9,296	4,932	3,055	76	59
Day 08	3,563	454	256	210	174
Day 09	2,483	81	230	204	142
Day 10	732	6	64	140	34
Day 11	706	17	34	139	18
Day 12	1,058	250	92	96	24
Day 13	603	21	69	114	36
Day 15	1,525	49	1,146	123	1,024
Day 16	1,089	25	849	125	791
Day 17	910	8	707	126	667
Day 18	1,158	27	882	127	833
Day 19	975	13	726	131	689
Day 20	834	8	586	132	535
Day 21	961	23	632	130	587

*Low quality reads in this length-class were screened out as in Table S1*A*.

Table S1C. Good's coverage and α diversity* using high-quality reads and OTUs picked at 97% sequence identity

Sample	Good's coverage	Observed OTUs	Simpson (1 - *D*)	Shannon	Equitability (evenness)	Phylodiversity (PD)
Day 05	68.3	61	0.98	5.73	0.97	2.89
Day 06	78.6	59	0.98	5.69	0.97	2.94
Day 08	54.0	76	0.99	6.19	0.99	4.58
Day 09	50.7	76	0.99	6.18	0.99	4.02
Day 10	95.4	27	0.92	4.10	0.87	1.01
Day 11	95.0	30	0.92	4.30	0.87	0.96
Day 12	95.1	27	0.92	4.14	0.88	0.92
Day 13	95.2	31	0.94	4.44	0.90	1.04
Day 15	76.4	44	0.96	5.08	0.93	1.33
Day 16	45.7	54	0.97	5.40	0.94	1.99
Day 17	68.3	39	0.94	4.73	0.89	1.59
Day 18	74.3	39	0.95	4.76	0.90	1.69
Day 19	62.9	48	0.96	5.17	0.92	2.15
Day 20	72.0	35	0.93	4.52	0.88	1.42
Day 21	66.2	45	0.95	4.94	0.90	1.82

*Means for 10 random draws of 82 high-quality sequences per sample are shown for each measure of α diversity.

Table S2*A*. Summary table, dominant populations

Bin	N_{reads}	Day 10	Day 16	Day 18	Day 21	$N_{contigs}$	L_{AVG}	L_{MAX}	L_{TOTAL}	Depth	N_{rRNA}	Amino acid identity to closest fully sequenced isolate
Serratia UC1SER	231,922	17	48,355	127,778	55,772	9	558.6 kb	2.36 Mb	5.03 Mb	17 ×	7	97.3% (*Serratia marcescens*, Sanger Institute)
Citrobacter UC1CIT	172,651	26	40,789	73,008	58,828							
UC1CIT-i chromosome	166,688	26	38,672	71,241	56,749	10	490.2 kb	2.55 Mb	4.90 Mb	13 ×*	8	97.5% (*Citrobacter* sp. 30_2/GG657366-83)
UC1CIT-ii anchored paths	3,099	0	1,540	468	1,091	93	2.8 kb	9.1 kb	257.8 kb	4.5 ×	0	n/a
Plasmid	2,864	0	577	1,299	988	2	30.0 kb	57.1 kb	60.0 kb	17 ×	0	n/a
Enterococcus faecalis UC1ENT	39,018	3	5,693	20,907	12,415							
Chromosome	33,783	3	4,762	18,678	10,340	810	3.3 kb	18.4 kb	2.61 Mb	4.7 ×	1	98.7% (*Enterococcus faecalis* V583/AE016830-3)
Plasmid1	1,795	0	223	1,022	550	1	n/a	n/a	68.7 kb	9.6 ×	0	n/a
Plasmid2	147	0	19	74	54	1	n/a	n/a	8.4 kb	6.4 ×	0	n/a
Phage1	135	0	24	73	38	1	n/a	n/a	8.4 kb	5.3 ×	0	n/a
Phage2	3,158	0	665	1,060	1,433	1	n/a	n/a	28.9 kb	40.0 ×	0	n/a

*13× indicates total coverage originating from coassembled reads from strains UC1CIT-i and UC1CIT-ii. However, the assembled chromosome also contains genomic regions only present in the dominant strain UC1CIT-i.

Table S2*B*. Summary table, minor populations

Bin	N_{reads}	Day 10	Day 16	Day 18	Day 21	$N_{contigs}$	L_{AVG}	L_{MAX}	L_{TOTAL}
Pseudomonas	7,733	6,778	698	164	93	613	0.7 kb	2.4 kb	441.5 kb
Staphylococcus	626	169	215	182	60	94	0.7 kb	1.7 kb	69.3 kb
Enterobacter	1,291	0	1,038	172	81	180	1.0 kb	3.3 kb	176.0 kb
Klebsiella	2,826	1	699	1,114	1,012	371	0.9 kb	3.7 kb	346.8 kb
Enterobacteriaceae	14,115	64	5,833	4,446	3,772	1,821	0.8 kb	17.7 kb	1.51 Mb
Firmicutes	41	5	11	11	14	4	1.1 kb	2.7 kb	4.5 kb
Proteobacteria	152	5	60	58	29	26	0.8 kb	1.4 kb	20.0 kb
Plasmid_16230	97	0	5	34	58	1	n/a	n/a	1.7 kb
Plasmid_16231	157	0	13	61	83	1	n/a	n/a	4.0 kb
Plasmid_15087	71	0	64	6	1	1	n/a	n/a	2.5 kb
Plasmid_19143	52	0	8	30	14	1	n/a	n/a	2.3 kb
Plasmid_pHCM2	360	0	106	58	196	21	2.4 kb	5.4 kb	51.2 kb
Plasmid_other	826	2	377	260	187	27	2.9 kb	10.1 kb	77.4 kb
Phage_01020	237	0	38	127	72	1	n/a	n/a	9.2 kb
Phage_02073	78	0	76	2	0	1	n/a	n/a	5.8 kb
(Pro)phage	1,695	0	573	516	606	33	3.6 kb	22.7 kb	117.4 kb
Unassigned	4,621	302	1,678	1,356	1,285	645	0.8 kb	5.2 kb	496.2 kb
Contigs < 500 bp	23,082	4,811	7,026	6,475	4,770	n/d	n/d	n/d	n/d
Unaligned (non-Human)	67,974	18,827	13,992	20,623	14,448	n/a	n/a	n/a	n/d
Unaligned (Human)	101,444	73,950	3,225	20,611	3,658	n/a	n/a	n/a	n/d
Human contigs	14,406	10,141	479	3,265	521	n/d	n/d	n/d	n/d

Other Supporting Information Files

Dataset S1 (XLS)
Dataset S2 (XLS)