Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/5/12/eaax5727/DC1

Supplementary Materials for

Necrotizing enterocolitis is preceded by increased gut bacterial replication, *Klebsiella*, and fimbriae-encoding bacteria

Matthew R. Olm, Nicholas Bhattacharya, Alexander Crits-Christoph, Brian A. Firek, Robyn Baker, Yun S. Song, Michael J. Morowitz, Jillian F. Banfield*

*Corresponding author. Email: jbanfield@berkeley.edu

Published 11 December 2019, *Sci. Adv.* **5**, eaax5727 (2019)
DOI: 10.1126/sciadv.aax5727

The PDF file includes:

Fig. S1. Metagenomic characterization of 1163 samples from 160 premature infants.
Fig. S2. Fecal samples taken before NEC diagnosis have a higher abundance of plasmids from specific bacterial taxa.
Fig. S3. PCA is unable to separate pre-NEC and control samples.
Fig. S4. ML feature importance values reveal organismal associations with NEC.
Legends for tables S1 to S6

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/5/12/eaax5727/DC1)

Table S1 (.csv format). Metagenomic sequencing depth and read quality information.
Table S2 (Microsoft Excel format). Patient metadata.
Table S3 (Microsoft Excel format). Information about dereplicated secondary metabolite clusters, de novo–assembled genomes, and genome-wide importances of genomes.
Table S4 (Microsoft Excel format). Accuracy of ML algorithms and protein clustering algorithms and mapping-based abundances of bacterial taxa.
Table S5 (Microsoft Excel format). Full feature table provided to the ML classifier and importances of all features resulting from the ML classifier.
Table S6 (Microsoft Excel format). Proteins enriched in genomes of interest and identified fimbrial genes.

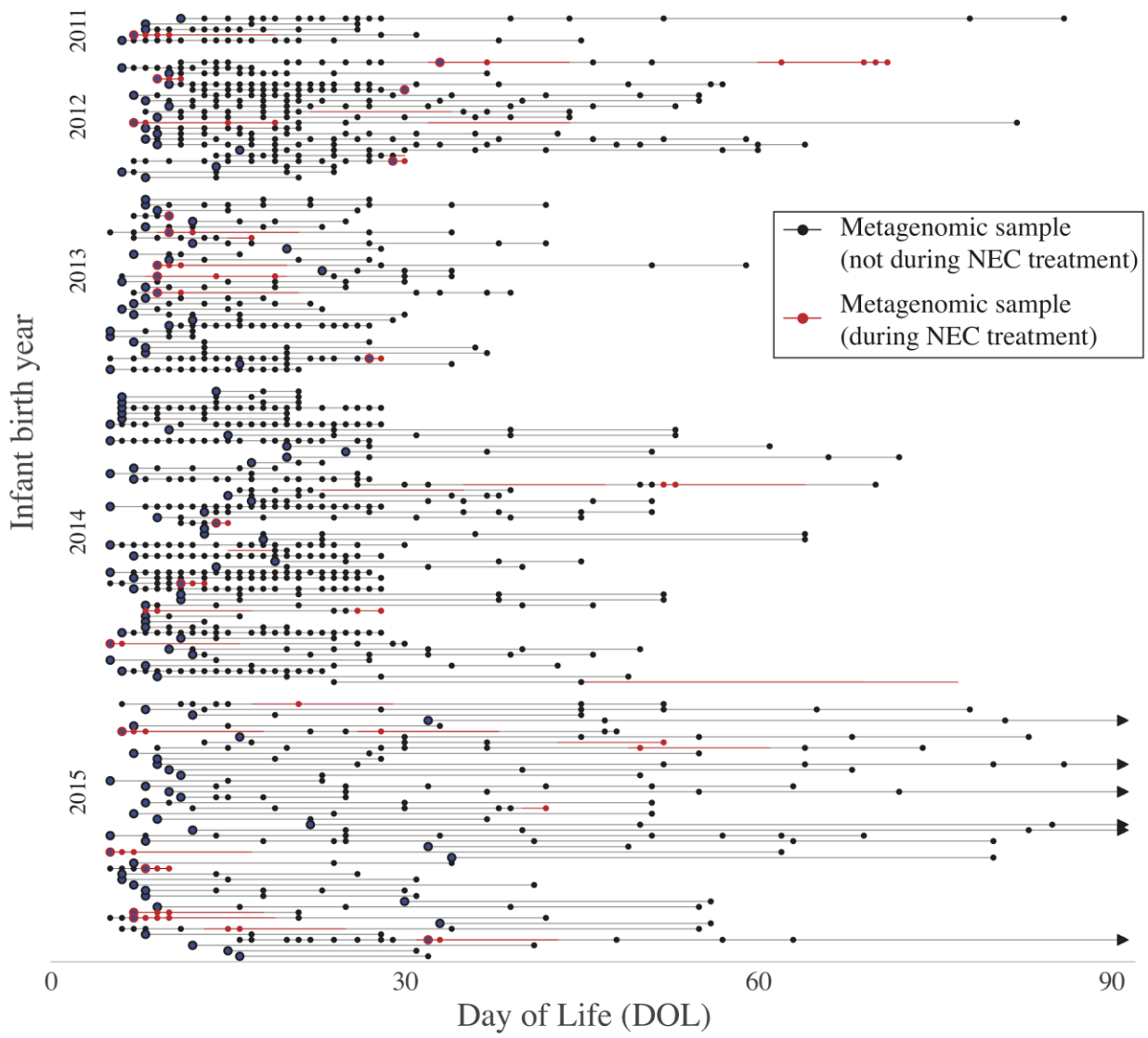

Fig. S1. Metagenomic characterization of 1163 samples from 160 premature infants. Each infant is represented by a horizontal line, and dots on the line represent sequenced metagenomic samples. Red sections indicate periods in which the infant was undergoing treatment for necrotizing enterocolitis. For some statistical tests one sample was chosen for each infant (pre- NEC and control samples); these samples are marked with larger circles.

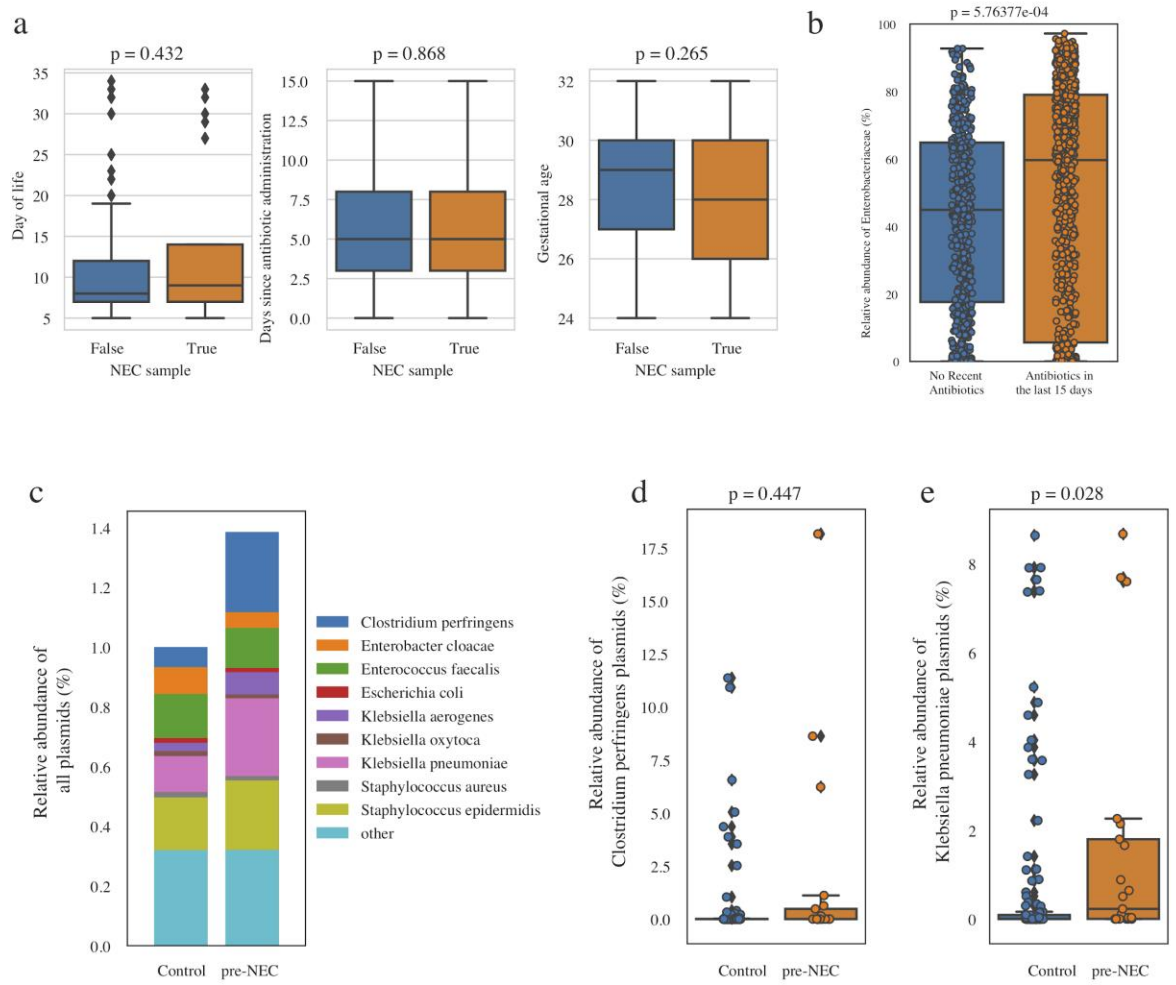

Fig. S2. Fecal samples taken before NEC diagnosis have a higher abundance of plasmids from specific bacterial taxa. (**a**) Distribution of clinical metadata in metagenomic samples used for statistical tests. There are 21 pre-NEC samples (all within 2 days of NEC diagnosis) and 126 control samples. Distributions were compared using the Wilcoxon rank sums test, with the overall p-value reported. (**b**) Relative abundance of Enterobacteriaceae in samples within 15 days of antibiotic administration vs. other samples. (**c-e**) Total plasmid content in pre-NEC vs. control samples. (**c**) Taxonomic distribution of plasmids in control and pre-NEC samples. Height of bars represents the average relative abundance of plasmids of each species-level taxa. (**d, e**) Difference in the total abundance of *C. perfringens* plasmids (**d**) and *Klebsiella pneumoniae* (**e**) plasmids in control and pre-NEC samples. *P*-values listed above each plot are from Wilcoxon rank-sums test.

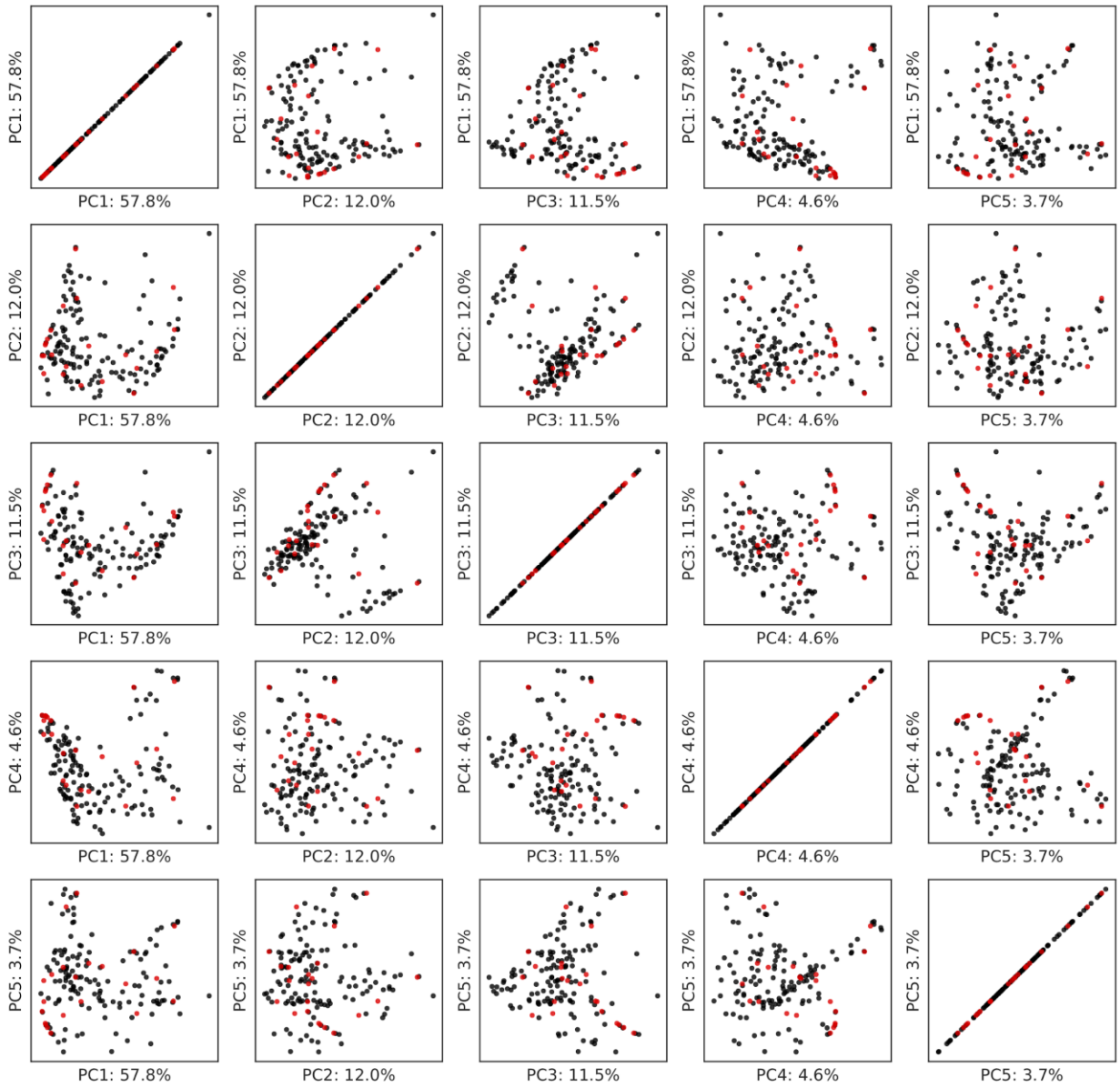

Fig. S3. PCA is unable to separate pre-NEC and control samples. Visualization of metadata as it relates to principal component analysis. The first sheet shows pre-NEC (red) and control (black) samples across the top 5 principal components. Subsequent sheets show the first two principal components of both matched and all samples, colored by 8 different pieces of health metadata.

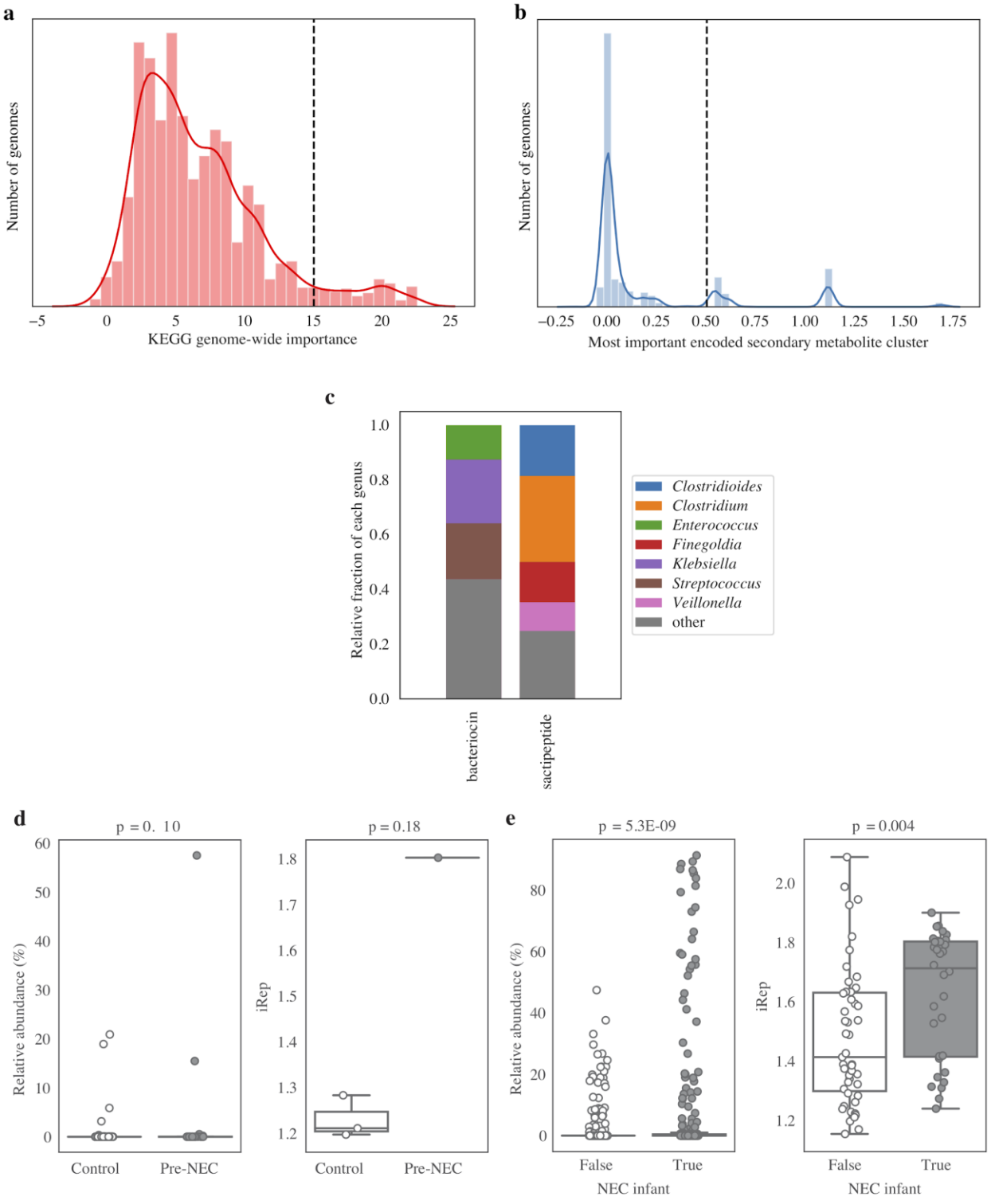

Fig. S4. ML feature importance values reveal organismal associations with NEC. (**a, b**) Genome-wide associations with important KEGG modules and important secondary metabolite clusters. (**a**) The distribution of total genome KEGG module importances. Those with overall importances above 15 were considered "metabolically important. Total genome KEGG module importance was calculated by taking the sum of the importances of all KEGG modules encoded by each genome. (**b**) The distribution of the

most important secondary metabolite cluster of each genome. Genomes without a secondary metabolite cluster are not included. Genomes encoding a secondary metabolite cluster with an importance over 0.5 were considered to encode an important secondary metabolite clusters. (**c**) Genus-level taxonomic makeup of genomes encoding two types of secondary metabolite clusters enriched in NEC infants. Taxa at less than 6% relative abundance are put into the "other" category. (**d, e**) NEC association with genomes which are not organisms of interest but encode fimbriae cluster 49. (**d**) Comparing the abundance and iRep of these organisms in pre-NEC and control infants does not achieve statistical significance, seemingly because there are not enough data-points to compare. (**e**) Comparing the abundance and iRep of these organisms in all samples from NEC infants vs. all samples from control infants does achieve statistical significance. P-values from Wilcoxon rank-sums test.

Table S1. Metagenomic sequencing depth and read quality information. Metagenomic sequencing information for all samples. Samples marked as day of life 0 are co-assemblies.

Table S2. Patient metadata.

Table S3. Information about dereplicated secondary metabolite clusters, de novo–assembled genomes, and genome-wide importances of genomes. Based on secondary metabolites and KEGG modules.

Table S4. Accuracy of ML algorithms and protein clustering algorithms and mapping-based abundances of bacterial taxa.

Table S5. Full feature table provided to the ML classifier and importances of all features resulting from the ML classifier. Feature names are coded using the format "category $ type of data $ value".

Table S6. Proteins enriched in genomes of interest and identified fimbrial genes.