nature methods

SUPPLEMENTARY INFORMATION

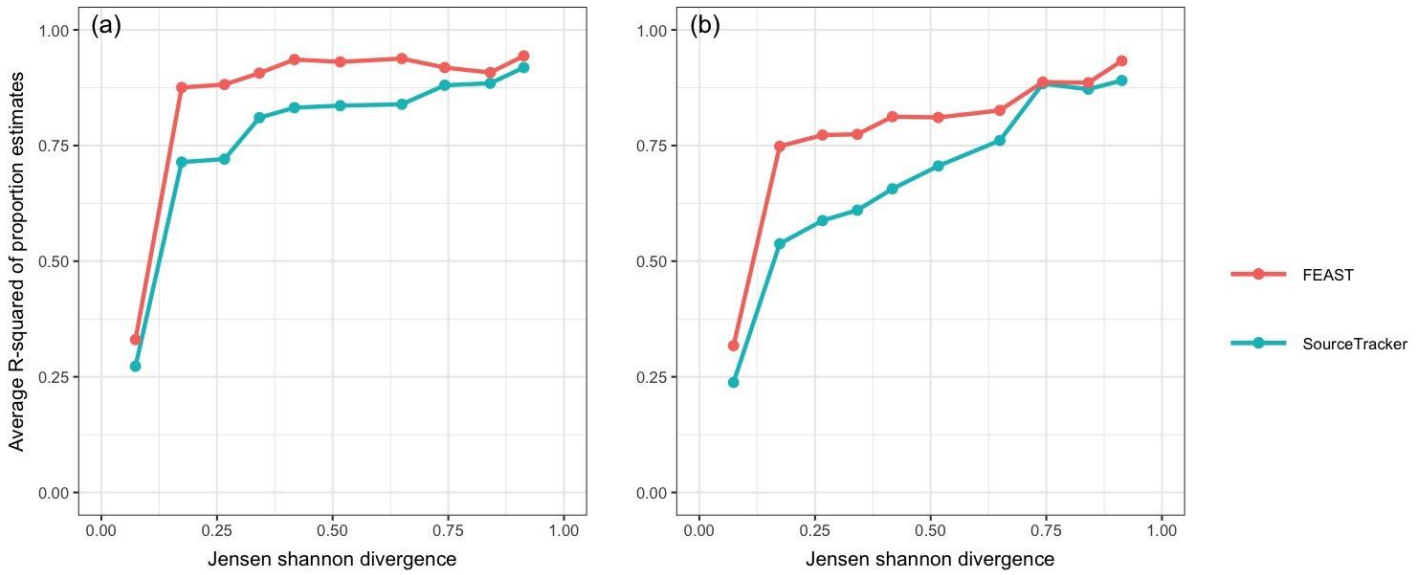
https://doi.org/10.1038/s41592-019-0431-x

In the format provided by the authors and unedited.

FEAST: fast expectation-maximization for microbial source tracking

Liat Shenhav[1], Mike Thompson[2], Tyler A. Joseph[3], Leah Briscoe[2], Ori Furman[4], David Bogumil[4], Itzhak Mizrahi[4], Itsik Pe'er[3] and Eran Halperin[1,2,5,6]*

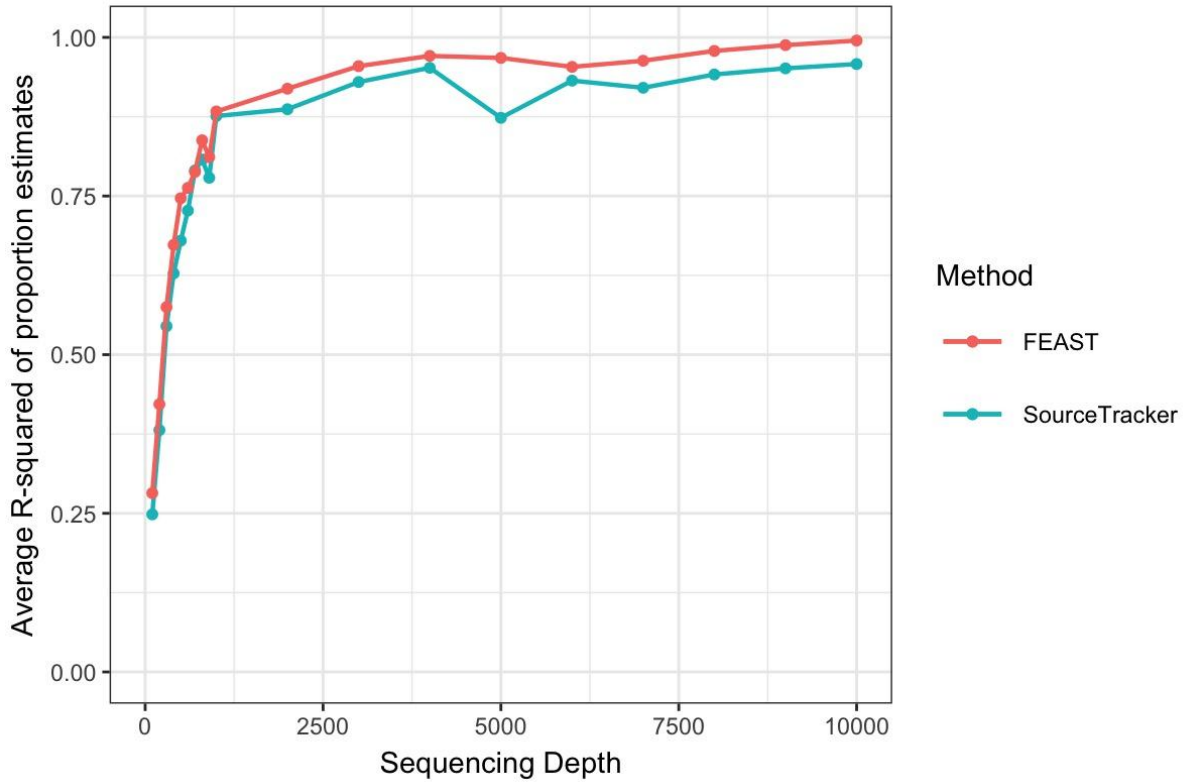
[1]Department of Computer Science, University of California Los Angeles, Los Angeles, CA, USA. [2]Department of Human Genetics, University of California Los Angeles, Los Angeles, CA, USA. [3]Department of Computer Science, Columbia University, New York, NY, USA. [4]Life Sciences, Ben Gurion University, Be'er Sheva, Israel. [5]Department of Anesthesiology and Perioperative Medicine, University of California Los Angeles, Los Angeles, CA, USA. [6]Department of Computational Medicine, University of California Los Angeles, Los Angeles, CA, USA. *e-mail: ehalperin@cs.ucla.edu



Supplementary Figure 1

The accuracy of FEAST and SourceTracker using data-driven synthetic mixtures

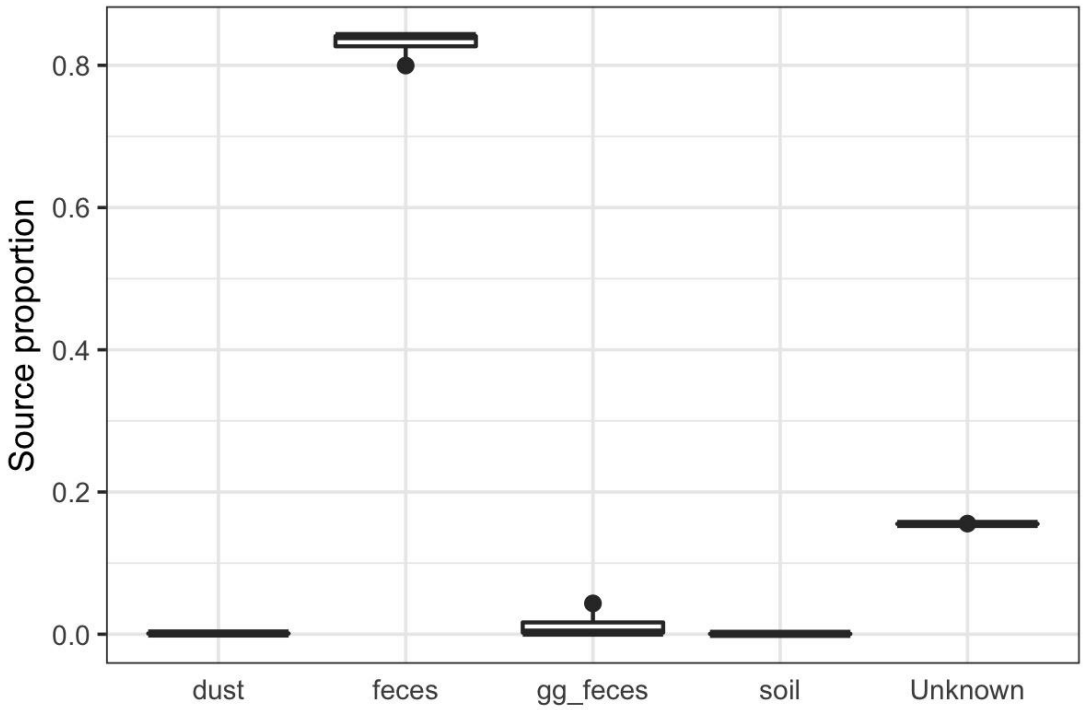
The accuracy of FEAST and SourceTracker on simulated data. Each simulation was performed using 10 real source environments and simulated sinks. The x-axis is average Jensen-Shannon divergence value across known sources. The y-axis represents correlation across all source environments between true and estimated mixing proportions, measured by (a) the squared Pearson correlation coefficient averaged across sources, and (b) the squared Spearman correlation coefficient averaged across sources.



Supplementary Figure 2

Evaluation of FEAST and SourceTracker through varying levels of sequencing depth

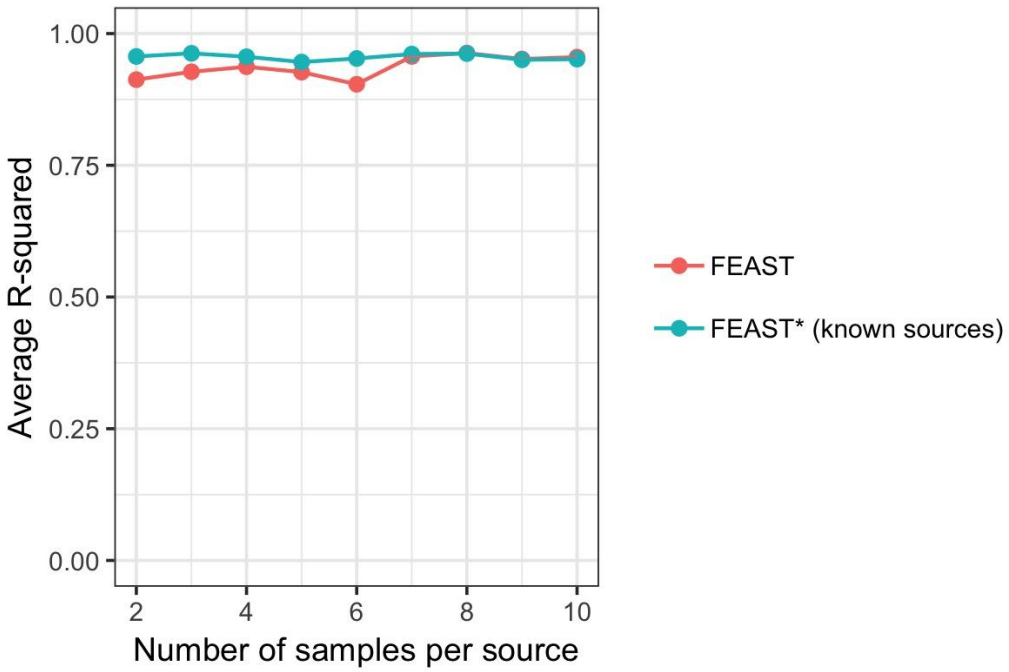
Evaluation of FEAST and SourceTracker through varying levels of sequencing depth. Similarity of sequences remained constant (Jensen-Shannon divergence = 0.95, trivial to disambiguate), while sequencing depth was set to vary in the range 100-10,000.



Supplementary Figure 3

The expected variance in FEAST's output

The expected variance in FEAST's output using the dataset from McDonald et al. We used the gut microbiome of one, randomly selected, ICU patient as a sink, and the sources considered by McDonald et al. : 126 healthy controls, 126 samples of mammalian corpse decomposition, 126 samples of the gut from healthy children, and 126 samples from indoor house surfaces. By repeating this analysis 100 times and calculating the standard deviation of each source we demonstrate that the variance in FEAST's output is very small (i.e., sd(dust) = 7.7e-05, sd(healthy adults' feces) = 0.01, sd(healthy children's feces) = 0.01,sd(soil) = 5e-05, sd(unknown) = 8.5e-05).



Supplementary Figure 4

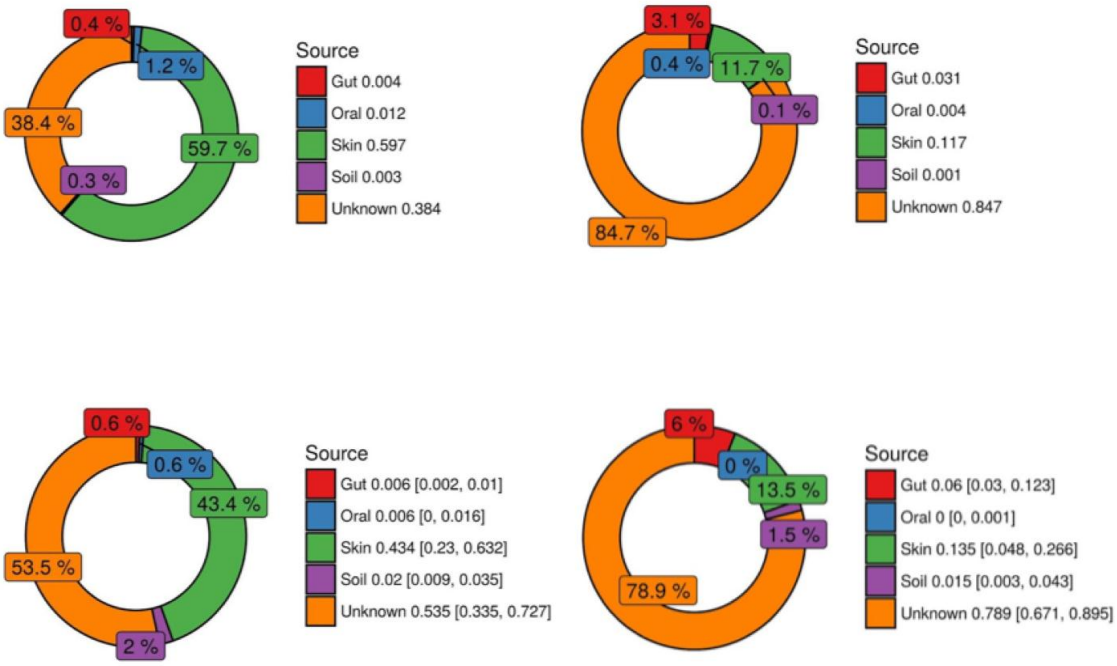
The effect of noisy samples among sources on prediction accuracy

The effect of noisy samples among sources on prediction accuracy (i.e., estimation of the known and unknown sources). As we increase the number of samples per source, FEAST's prediction accuracy improves, however this effect is moderate (squared Pearson correlation ranges from 0.9 - 0.99, Jensen-Shannon divergence values range from 0.87-0.92).

Supplementary Figure 5

The source proportions using SourceTracker

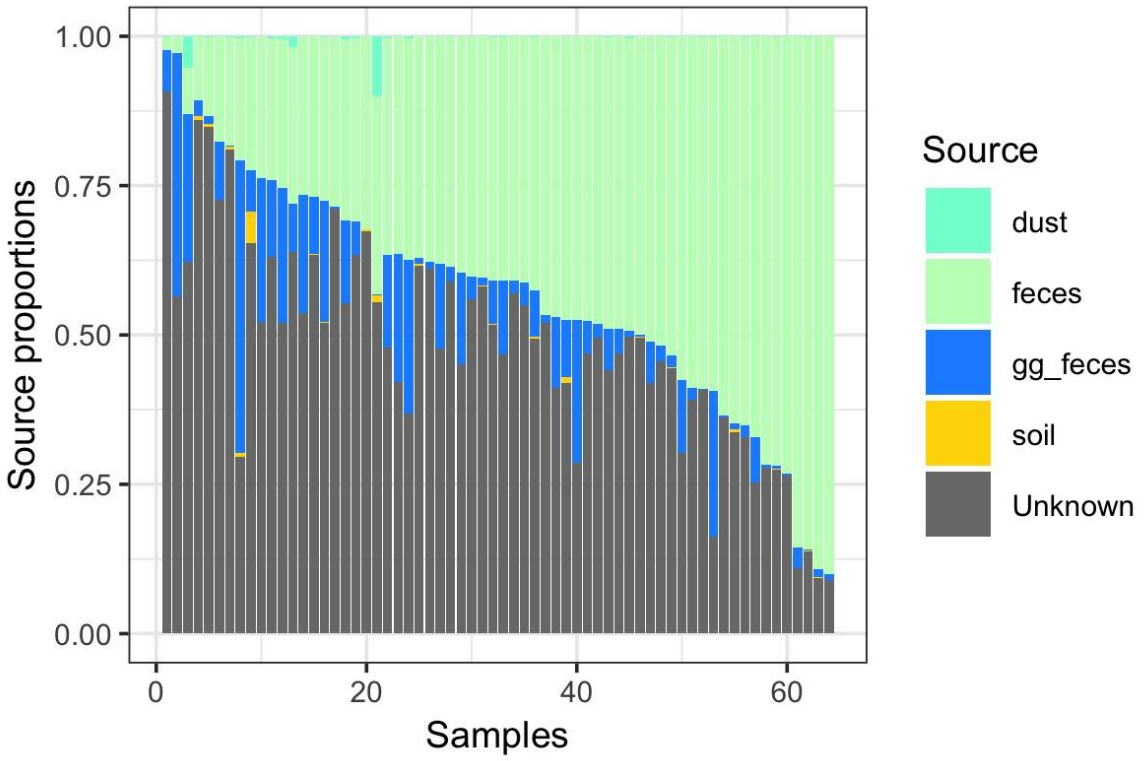
SourceTracker estimations of source contribution (the gut microbiome of mother, infant at 4 months and infant at birth) to the gut microbiome of 12-month-old infants. According to SourceTracker differences between C-section (n = 15) and Vaginally-delivered (n = 83) infants in terms of maternal contribution are not significant (two-sided t-test p-value = 0.6408). Box plots indicate the median (central lines), interquartile range (hinges), and the 5th and 95th percentiles (whiskers).



Supplementary Figure 6

Detecting contamination in lab-settings

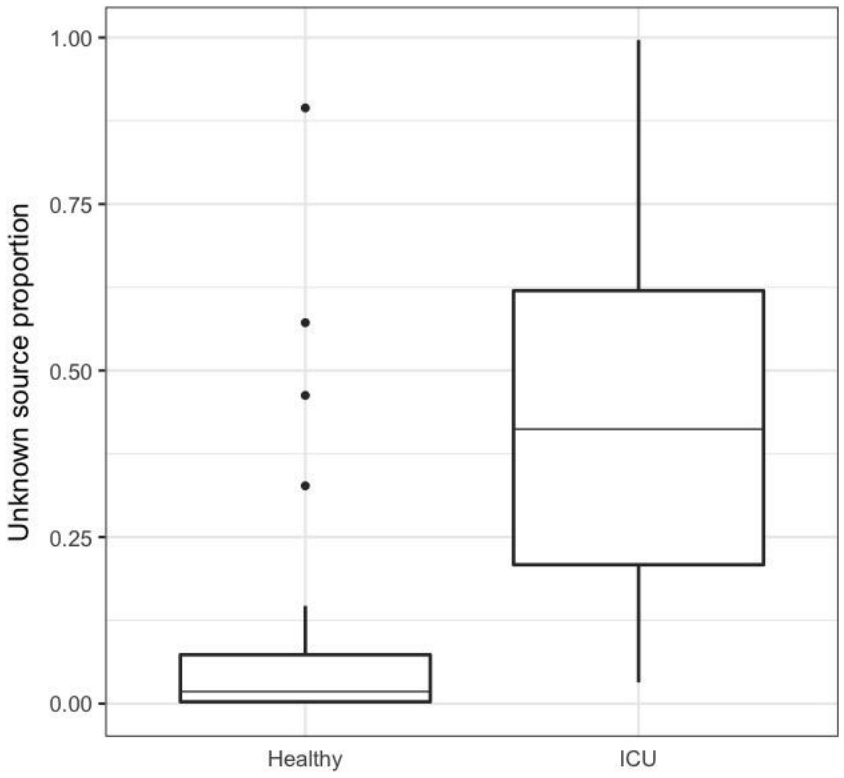
FEAST and SourceTracker report consistent proportions of contamination, despite minor discrepancies in a lab-setting (left: keyboard, right: Counter). Estimates on the top row were reported by SourceTracker and estimates on the bottom row were reported by FEAST.



Supplementary Figure 7

Gut microbiome samples from ICU patients are not reminiscent of gut samples from healthy individuals

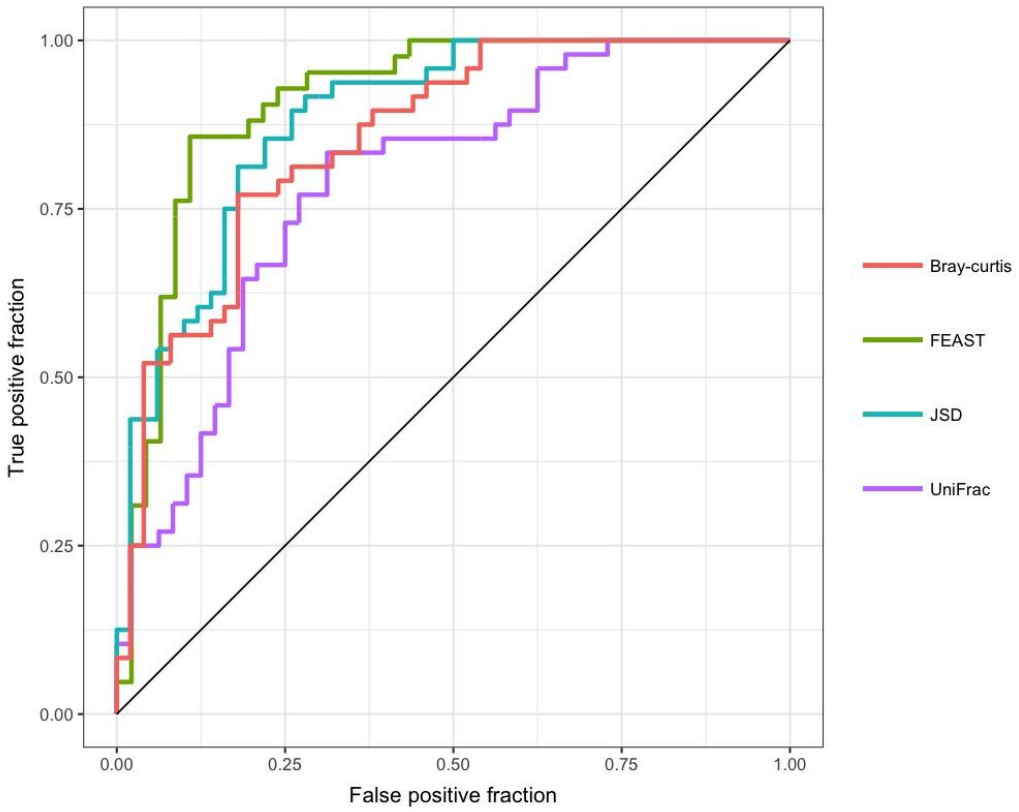
Gut samples from ICU patients are not reminiscent of gut samples from healthy individuals. We used the gut microbiome of each ICU patient (at discharge or after 10 days) as a sink, and the sources considered by the original study (McDonald et al. 2016): 126 samples from the American Gut Project (healthy controls), 126 samples of mammalian corpse decomposition, 126 samples of the gut from healthy children (Global Gut study) , and 126 samples from indoor house surfaces.



Supplementary Figure 8

Unknown source distribution across sink samples (ICU patients vs. healthy individuals)

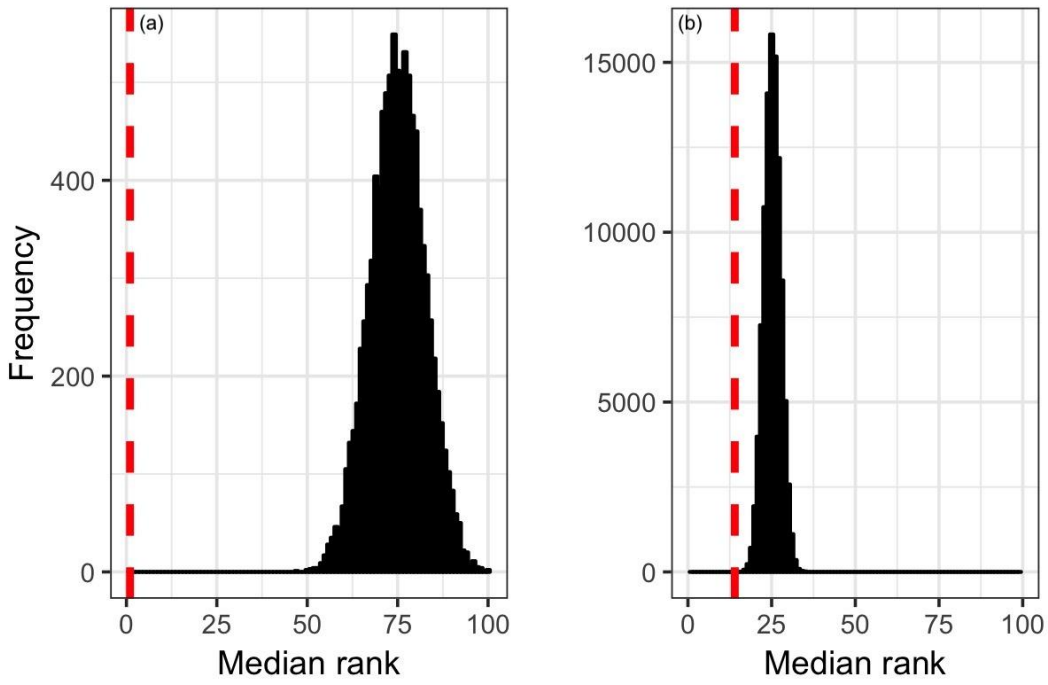
The distribution of the unknown source across sink samples - healthy individuals and ICU patients (n = 100).



Supplementary Figure 9

Distinguishing between ICU patients and healthy individuals

The receiver operating characteristic curve (ROC curve) using FEAST, Weighted UniFrac, Bray-curtis and Jensen Shannon divergence to classify healthy individuals and ICU patients with dysbiosis. FEAST AUC = 0.91, Weighted UniFrac AUC = 0.78, Jensen Shannon divergence AUC = 0.87, Bray-curtis AUC = 0.86.



Supplementary Figure 10

The source contribution across maternal samples

Distribution of the median random maternal rank in two scenarios: (a) all maternal and early infant samples (from all the infants in the study) were considered as potential sources (n = 293 sources), and (b) only the maternal samples were considered as potential sources (n = 98 sources). In both scenarios samples taken from infants at age 12 months were considered as sinks (n = 98 sinks). The red vertical line in each figure corresponds to the actual median rank of the maternal contribution.

FEAST: Fast Expectation-Maximization for Microbial Source Tracking

Liat Shenhav[1], Mike Thompson[2], Tyler A. Joseph[3], Leah Briscoe[2], Ori Furman[5], David Bogumil[5], Itzhak Mizrahi[5], Itsik Pe'er[3], and Eran Halperin[1,2,4,6]

[1]Department of Computer Science, University of California Los Angeles, Los Angeles, CA, USA

[2]Department of Human Genetics, University of California Los Angeles, Los Angeles, CA, USA

[3]Department of Computer Science, Columbia University, New York, NY, USA

[4]Department of Anesthesiology and Perioperative Medicine, University of California Los Angeles, Los Angeles, CA, USA

[5]Life Sciences, Ben Gurion University

[6]Department of Computational Medicine, University of California Los Angeles, Los Angeles, CA, USA

Supplementary Information

Supplementary Table: Running time comparison

Number of sources	5	10	50	100	500	1000
SourceTracker	00 : 05 : 18	00 : 09 : 06	05 : 39 : 00	11 : 43 : 02	54 : 34 : 02	71 : 07 : 00
FEAST	00 : 00 : 09	00 : 00 : 36	00 : 07 : 42	00 : 15 : 54	00 : 47 : 46	01 : 35 : 30

Table S1. Running time (hh:mm:ss) comparison across multiple source environments, randomly sampled from the Earth Microbiome Project. Sequencing depth is 10,000 reads per source.

Supplementary Note: Main simulation study In order to examine the accuracy of *FEAST*, we used multiple source environments with varying degrees of overlap in their distribution by randomly sampling from the Earth Microbiome Project. Each source environment was sub-sampled to contain $10,000$ reads. In each iteration of our simulation we sampled $K+1$ known environments and used them to build a synthetic sink, with different mixing proportions. In order to simulate an unknown source, we use only K source environments as our sources.

The simulation procedure was as follows. For each $l = 1 : T_1$ (different Jensen Shannon divergence values):

1. Draw $K+1$ samples $S_1, \ldots, S_{K+1}$, from a selected data set.
2. Draw noisy realization of $S_1, \ldots, S_{K+1}$ from the Multinomial distribution (denoted $\tilde{S_k}$).
3. For each $i = 1 : T_2$ (different mixing proportions):
   (a) Generate random mixing $m \sim Pareto(\alpha > 0)$, where $\sum m = 1$.
   (b) Set the sink sample abundances to $\sum_{k=1}^{K+1} m_k S_k$ per taxa.
   (c) Estimate the known source proportions in the sink using $\tilde{S_1}, \ldots, \tilde{S_K}$.
   (d) Estimate the unknown source proportions in the sink.
4. Calculate the squared Pearson correlation (r^2) between the estimated and the true mixing proportions per source and average across sources.

5. Calculate the average Jensen-Shannon divergence of m (based on the pairwise Jensen-Shannon divergence).

In the simulations presented we used $T_1 = 10$, $T_2 = 30$, $K = 20$.

Supplementary Note: Sequencing depth simulations In order to examine the robustness of *FEAST* to varying levels of sequencing depth, we used multiple source environments from the Earth Microbiome Project while varying their sequencing depth. In each iteration of our simulation we sampled environments (with median Jensen-Shannon divergence of 0.95) and used them to build a synthetic sink, with different mixing proportions and a set sequencing depth ranging from 100 through 10,000. Notably, by choosing a median Jensen-Shannon divergence of 0.95 we wanted to emphasize that even under the scenario in which the sources are non-overlapping and thus trivial to disambiguate, the sequencing depth will have an effect. Additionally, in these simulations, we only varied the sequencing depth of the sources. However, since the sink samples are a linear combination of the sources, these samples are also, indirectly, affected. To simulate an unknown source, only K source environments are designated as known sources.

The simulation procedure was as follows. For each $l = 1 : D_1$ (different sequencing depth values):

1. Draw $K + 1$ samples $S_1, \ldots, S_{K+1}$, from a selected data set.
2. Draw noisy realization of $S_1, \ldots, S_{K+1}$ from the Multinomial distribution (denoted $\tilde{S_k}$).
3. For each $i = 1 : D_2$ (different mixing proportions) :
   (a) Generate random mixing $m \sim Pareto(\alpha > 0)$, where $\sum m = 1$.
   (b) Set the sink sample abundances to $\sum_{k=1}^{K+1} m_k S_k$ per taxa.
   (c) Estimate the known source proportions in the sink using $\tilde{S_1}, \ldots, \tilde{S_K}$.
   (d) Estimate the unknown source proportions in the sink.

4. Calculate the squared Pearson correlation (r^2) between the estimated and the true mixing proportions per source and average across sources.

5. Calculate the average Jensen-Shannon divergence of m (based on the pairwise Jensen-Shannon divergence).

In the simulations presented we used $D_1 = 19$, $D_2 = 30$, $K = 20$.

Supplementary Note: Unknown source simulations In order to evaluate *FEAST*'s ability to estimate the contribution of the unknown source, we used real source environments from Lax et al. (2014) [1] where disambiguation of sources is challenging, and created synthetic sink communities. Given that any source not sampled should, theoretically, be accounted for in the unknown source, realistic values of the unknown source can therefore span the range of percentages occupied by the observed sources. Specifically, there are scenarios in which the known sources comprise the entirety of the sink (unknown source contribution = 0), or on the other hand, scenarios in which the known sources did not contribute any taxa to the sink (unknown source contribution = 1). Therefore, the unknown source contribution values in our simulation ranges from 0 to 1. As a measure of accuracy, we used the squared Pearson correlation between the estimated mixing proportions and the true mixing proportions for each individual source across repeated simulation runs for the same scenario as the measure of accuracy.

The simulation procedure was as follows. For each $l = 1 : U_1$ (different unknown source proportions):

1. Set the unknown proportion u to $U_1[l]$.

2. Generate random mixing $m - 1 \sim Pareto(\alpha > 0)$, where $\sum m - 1 = 1 - u$.

3. For each $i = 1 : U_2$ (different Jensen-Shannon divergence $\in (0.5 + \epsilon, 0.5 - \epsilon)$:

   (a) Draw $K + 1$ samples $S_1, \ldots, S_{K+1}$, from a selected data set.

(b) Draw noisy realization of $S_1, \ldots, S_{K+1}$ from the Multinomial distribution (denoted $\tilde{S}_k$).

(c) Set the sink sample abundances to $\sum_{k=1}^{K} m_k S_k + S_k$ per taxa Draw $K+1$ samples $S_1, \ldots, S_{K+1}$, from a selected data set.

(d) Estimate the unknown source proportions in the sink.

4. Calculate the squared Pearson correlation (r^2) between the estimated and the true mixing proportions of the unknown source.

In the simulations presented we used $U_1 \in (0, 1)$, $T_2 = 30$, $K = 4$, $\epsilon = 0.2$

Supplementary Note: The effect of noisy samples among sources on prediction accuracy We used $K+1$ distinct source environments randomly sampled from the Earth Microbiome Project (i.e., soil, fresh water, feces etc.), where each source was represented by 10 different samples (e.g., $soil_1$, $soil_2$, etc). We then amalgamated these 10 samples (per source environment) and used them to build a synthetic sink, with different mixing proportions. In each iteration of our simulation we sampled $k \in 1, \cdots, 10$ samples from each source environment in order to estimate the corresponding mixing proportions of the amalgamated sources. To simulate an unknown source, we use only K source environments as our known sources. Indeed, we observed that as we increase the number of samples per source, *FEAST*s prediction accuracy improves, however this effect is moderate (squared Pearson correlation ranges from $0.9 - 0.99$, Jensen-Shannon divergence values range from $0.87 - 0.92$).

The simulation procedure was as follows. Draw 11 sources $S_1, \ldots, S_K$, from the Earth Microbiome Project. From each source S_i draw 10 different samples.

1. Draw K sources $S_1, \ldots, S_{K+1}$, from the Earth Microbiome Project. From each source S_i draw 10 different samples.
2. Amalgamate the 10 samples per source environment and create new sources $\tilde{S}_1, \ldots, \tilde{S}_{K+1}$
3. Generate random mixing $m \sim Pareto(\alpha > 0)$, where $\sum m = 1$.

4. Set the sink sample abundances to $\sum_{k=1}^{K+1} m_k \tilde{S}_k$ per taxa.

For each $L = 1 : 10$ (different number of samples representing the sources):

1. Draw L samples from each source $S_{L1}, \ldots, S_{L(K+1)}$,
2. Draw noisy realization of $S_{L1}, \ldots, S_{L(K+1)}$ from the Multinomial distribution (denoted $\tilde{S}_{Lk}$).
3. For each $i = 1 : T_2$ (different mixing proportions) :
   (a) Estimate the known source proportions in the sink using $\tilde{S}_{L1}, \ldots, \tilde{S}_{LK}$.
   (b) Estimate the unknown source proportions in the sink.
4. Calculate the squared Pearson correlation (r^2) between the estimated and the true mixing proportions per source and average across sources.
5. Calculate the average Jensen-Shannon divergence of m (based on the pairwise Jensen-Shannon divergence).

In the simulations presented we used $K = 10, T_2 = 30$;

Supplementary Note: Using all maternal and early infant samples as potential sources

In this analysis we used the infants at their last time point as sink samples i.e., infant $i \in \{1, \ldots, 98\}$ at 12 months of age. First, we considered all maternal and early infant samples (from all the infants in the study) as potential sources. We used *FEAST* to rank the contribution of each source as compared to all other sources and found that the median contribution of the corresponding maternal sample across all sinks is 1. We performed a permutation test in which the ranks are randomly assigned for each sink, and the p-value is calculated as the number of permutations in which the median of the maternal contributions rank is smaller than the original median. We used 100,000 iterations and obtained a p-value < 0.0001 (Figure S10 (a)). Notably, the top 5 contributing sources included the corresponding infants family 83% of the time (43% of the cases,

the corresponding family ranked 1st, in 21% it ranked 2nd, 4% 3rd, 10% 4th and in 5% it ranked 5th). Next, We repeated these experiments by considering only the maternal samples as potential sources. In this set of sinks (i.e., infants at 12 months of age), the median maternal contribution was 14, and a similar permutation test as the one described above shows that this finding is statistically significance (p-value = 0.00017, Figure S10 (b)). Notably, the gut microbiome of healthy individuals is relatively similar. We therefore removed the samples with low Jensen Shannon divergence value to reduce noise in our estimations. To do this, for each sink_j, we calculated the Jensen Shannon divergence values (1) between mother_j and all other mothers (2) infant-at-birth_j and all other infants at birth (3) $\text{infant-at-4-months}_j$ and all other infants at 4 months, and calculated the median Jensen Shannon divergence for each of these source environments. We then removed samples whose Jensen Shannon divergence fell below their respective median.

Supplementary Note: Expectation-Maximization - derivation Here we derive the full EM algorithm for *FEAST* in detail. Recall that the observed data consist of the sink vector $x = (x_1, x_2..., x_N)$, and source vectors $y_i = (y_{i1}, y_{i2}, ..., y_{iN})$. for $1 \leq i \leq K$. The j-th component of each vector denotes the observed abundance of taxa j in the sink and sources respectively. Denote the total number of observations in each source by $C_i := \sum_{j=1}^{N} y_{ij}$ and total number of observations in the sink by $C := \sum_{j=1}^{N} x_j$. For each source, we have a vector $\gamma_i = (\gamma_{i1}, \gamma_{i2}, ..., \gamma_{iN})$ denoting the unobserved relative abundances of each source y_i. Further, there is assumed to be one unknown, unsampled, source–say $K+1$–with relative abundances $\gamma_{K+1} = (\gamma_{(K+1)1}, \gamma_{(K+1)2}, ..., \gamma_{(K+1)N})$.

Based on the source proportions, each source observation is assumed to have been generated by drawing a random sample from the source with replacement. Thus,

$$y_{ij} \sim \text{Multinomial}(C_i, \gamma_i) \tag{1}$$

For the sink we assume the following generative model. We draw C observations. For each observation $c = 1, ..., C$, we pick a source z^c with the probability of choosing source i given by α_i.

The vector $\alpha = (\alpha_1, \alpha_2, ..., \alpha_{K+1})$ gives the proportion of the sink derived from each source. Once the source is chosen, we pick taxa x^c from source z^c based on the relative abundances γ_{z^c}. Hence

$$z^c \sim \text{Multinomial}(1, \alpha) \tag{2}$$

$$x^c | z^c \sim \text{Multinomial}(1, \gamma_{z^c}) \tag{3}$$

where we denote $z^c = i$ as having drawn sample c from source i, indicating that the multinomial observation $z^c = (0, ..., 1, ..., 0)$ has 1 in its i-th component and 0s elsewhere. If we marginalize out source assignments z^c, we obtain

$$p(x^c = j) = \sum_{i=1}^{K+1} p(x^c = j | z^c = i) p(z^c = i) = \sum_{i=1}^{K+1} \gamma_{ij} \alpha_i.$$

Hence the marginal distribution of x^c is Multinomial$(1, (\beta_1, ..., \beta_N))$, where $\beta_j = \sum_{i=1}^{K+1} \alpha_i \gamma_{ij}$.

We can therefore rewrite the model as:

$$\beta_j = \sum_{i=1}^{K+1} \alpha_i \gamma_{ij} \quad \text{for } j = 1, ..., N \tag{4}$$

$$y_i \sim \text{Multinomial}(C_i, (\gamma_{i1}, \ldots, \gamma_{iN})) \quad \text{for } i = 1, ..., K \tag{5}$$

$$x \sim \text{Multinomial}(C, (\beta_1, \ldots, \beta_N)) \tag{6}$$

The expected complete log likelihood As demonstrated above, the log likelihood is given by

$$\log p(x, y_1, y_2, ..., y_K | \alpha, \gamma) = \sum_{j=1}^{N} x_j \log \left(\sum_{i=1}^{K+1} \alpha_i \gamma_{ij} \right) + \sum_{i=1}^{K} \sum_{j=1}^{N} y_{ij} \log(\gamma_{ij}) + \text{const} \tag{7}$$

Using the notation separating each draw from the sink, the complete log likelihood is given by

$$\log p(x^1, ..., x^C, z^1, ..., z^C, y_1, ..., y_K | \alpha, \gamma) = \sum_{c=1}^{C} \sum_{i=1}^{K+1} z_i^c (\log \gamma_{ix^c} + \log \alpha_i) + \sum_{i=1}^{K} \sum_{j=1}^{N} y_{ij} \log(\gamma_{ij}) + \text{const} \tag{8}$$

where $x^c = j$ denotes that observation c corresponds to taxa j. Taking expectations and collecting terms, the expected complete log likelihood is given by

$$Q = \sum_{i=1}^{K+1} \sum_{j=1}^{N} x_j p(i|j) \cdot \log(\alpha_i \gamma_{ij}) + \sum_{i=1}^{K} \sum_{j=1}^{N} y_{ij} \log(\gamma_{ij}) + \text{const} \tag{9}$$

where

$$p(i|j) = \frac{\alpha_i^{(t)} \gamma_{ij}^{(t)}}{\sum_{i=1}^{K+1} \alpha_i^{(t)} \gamma_{ij}^{(t)}} \tag{10}$$

The remainder of the derivation follows the main text.

Table S2. An example of *FEAST*'s output, using the infants dataset from Bäckhed et al. 2015 [2], which includes the top 50 pairs of taxa shared between a vaginally-delivered infant at 12 months of age (sample ERR525717, sink) and its corresponding maternal sample (sample ERR525720, source) (an optional setting)

Class	Order	Family	Genus	Species	Sink	Source
Acidimicrobiia	Acidimicrobiales	AKIW874	NA	NA	0.19639	0.06038
Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	coleohominis	0.17838	0.07551
Bacilli	Bacillales	Staphylococcaceae	Staphylococcus	equorum	0.11066	0.01407
Gammaproteobacteria	Alteromonadales	211ds20	NA	NA	0.10158	0.01238
Bacilli	Bacillales	Planococcaceae	Lysinibacillus	odysseyi	0.06719	0.10009
Bacilli	Bacillales	Bacillaceae	Bacillus	horneckiae	0.04117	0.10399
Bacilli	Bacillales	Planococcaceae	Planococcus	maitriensis	0.02739	0.0308
Bacilli	Lactobacillales	Enterococcaceae	Melissococcus	plutonius	0.0245	0.11209
Actinobacteria	Actinomycetales	Actinosynnemataceae	Actinokineospora	diospyrosa	0.02243	0.00341
Actinobacteria	Actinomycetales	NA	NA	NA	0.01857	0.00605
Bacilli	Bacillales	Sporolactobacillaceae	Bacillus	racemilacticus	0.01553	0.02153
Actinobacteria	Actinomycetales	Actinomycetaceae	NA	NA	0.01425	0.01886
Acidimicrobiia	Acidimicrobiales	koll13	NA	NA	0.01308	0.00929
Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	mucosae	0.01271	0.04955
Bacilli	Bacillales	Thermoactinomycetaceae	Mechercharimyces	mesophilus	0.01264	0.00328
Bacilli	Bacillales	Listeriaceae	Brochothrix	NA	0.01232	0.00968
Gammaproteobacteria	Oceanospirillales	Oleiphilaceae	NA	NA	0.01165	6.00E-05

Bacilli	Bacillales	[Exiguobacteraceae]	Exiguobacterium	NA	0.01011	0.01914
Actinobacteria	Actinomycetales	Corynebacteriaceae	Corynebacterium	variabile	0.00981	0.00441
Actinobacteria	Actinomycetales	Actinosynnemataceae	NA	NA	0.00662	0.00128
Bacilli	Bacillales	Staphylococcaceae	Jeotgalicoccus	NA	0.00641	0.02238
Solibacteres	Solibacterales	Solibacteraceae	CandidatusSolibacter	NA	0.00609	0.00239
Actinobacteria	Actinomycetales	Frankiaceae	NA	NA	0.00558	0.00175
Actinobacteria	Actinomycetales	Dermabacteraceae	Dermabacter	NA	0.00521	0.00388
Actinobacteria	Actinomycetales	Brevibacteriaceae	Brevibacterium	aureum	0.00331	0.00185
Gammaproteobacteria	Alteromonadales	Alteromonadaceae	ND137	NA	0.00255	0.00028
Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	pontis	0.0025	0.00418
Acidimicrobiia	Acidimicrobiales	Microthrixaceae	NA	NA	0.00209	0.00043
Clostridia	Clostridiales	Peptococcaceae	Desulfosporosinus	NA	0.00202	0.01059
Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	delbrueckii	0.00198	0.01541
Actinobacteria	Actinomycetales	Brevibacteriaceae	Brevibacterium	casei	0.00182	0.00058
Gammaproteobacteria	Oceanospirillales	Oceanospirillaceae	Nitrincola	NA	0.00172	4.00E-05
Actinobacteria	Actinomycetales	Actinospicaceae	NA	NA	0.00142	0.00083
Bacilli	Bacillales	Bacillaceae	Geobacillus	NA	0.0014	0.00512
Bacilli	Lactobacillales	Leuconostocaceae	Weissella	NA	0.00136	0.0013
Clostridia	Clostridiales	Clostridiaceae	Caminicella	NA	0.00136	0.00034
Acidimicrobiia	Acidimicrobiales	Acidimicrobiaceae	NA	NA	0.00135	0.00017
Bacilli	Bacillales	Planococcaceae	Planomicrobium	NA	0.00133	0.00341
Gammaproteobacteria	Chromatiales	Ectothiorhodospiraceae	Thioalkalivibrio	NA	0.00133	0.00015
Bacilli	Bacillales	Bacillaceae	Bacillus	marisflavi	0.00129	0.01611
Gammaproteobacteria	Alteromonadales	Shewanellaceae	Shewanella	benthica	0.00127	9.00E-05
Clostridia	Clostridiales	Ruminococcaceae	Oscillospira	guilliermondii	0.0012	0.00124
Actinobacteria	Actinomycetales	Actinomycetaceae	Mobiluncus	NA	0.00106	6.00E-05
Clostridia	Clostridiales	Clostridiaceae	Caloranaerobacter	NA	0.00101	4.00E-05
Bacilli	Bacillales	Bacillaceae	Bacillus	badius	0.00099	0.05367
Gammaproteobacteria	Alteromonadales	Alteromonadaceae	Marinimicrobium	NA	0.00096	0.00015
Actinobacteria	Actinomycetales	Dietziaceae	Dietzia	NA	9.00E-04	0.00047
Bacilli	Lactobacillales	Aerococcaceae	Lacticigenium	naphtae	0.00089	0.00543
Bacilli	Bacillales	[Exiguobacteraceae]	NA	NA	0.00081	0.00661
Bacilli	Bacillales	Planococcaceae	Solibacillus	NA	0.00078	0.00077

References

[1] Simon Lax, Daniel P Smith, Jarrad Hampton-Marcell, Sarah M Owens, Kim M Handley, Nicole M Scott, Sean M Gibbons, Peter Larsen, Benjamin D Shogan, Sophie Weiss, et al. Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science*, 345(6200):1048–1052, 2014.

[2] Fredrik Bäckhed, Josefine Roswall, Yangqing Peng, Qiang Feng, Huijue Jia, Petia Kovatcheva-Datchary, Yin Li, Yan Xia, Hailiang Xie, Huanzi Zhong, et al. Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell host & microbe*, 17(5):690–703, 2015.