Infant Microbial Operational Taxonomic Unit Analysis

Fangting Zhou

We coexist with our microbiota as mutualists. High-throughout sequencing technology has been widely used to quantify the microbial composition in order to explore its relationship with human health. Gut microbiota and the host exist in a mutualistic relationship, with the functional composition of the microbiota strongly affecting the health and well-being of the host. Early microbial colonization in infants is critically important for directing neonatal intestinal and immune development, and is especially attractive for studying the development of human-commensal interactions.

Data set

Operational taxonomic units (OTUs) are pragmatic proxies for microbial species at different taxonomic levels and have been the most commonly used units of microbial diversity. The current microbial data set seedLev2Counts was aligned using rapid annotation using subsystem technology against the SEED subsystem database. After aligning to the second level SEED subsystem, there are 162 species of OTUs. The sample size of the data set is 12 with 6 breast-feeding (BF) infants and 6 formula-feeding (FF) infants.

```
InfantID
```

```
##  [1] "BMS8"  "BMS10" "BMS16" "BF3"   "BF4"   "BF6"   "FF2"   "FF3"
##  [9] "FF7"   "FF13"  "FF15"  "FF5"
```

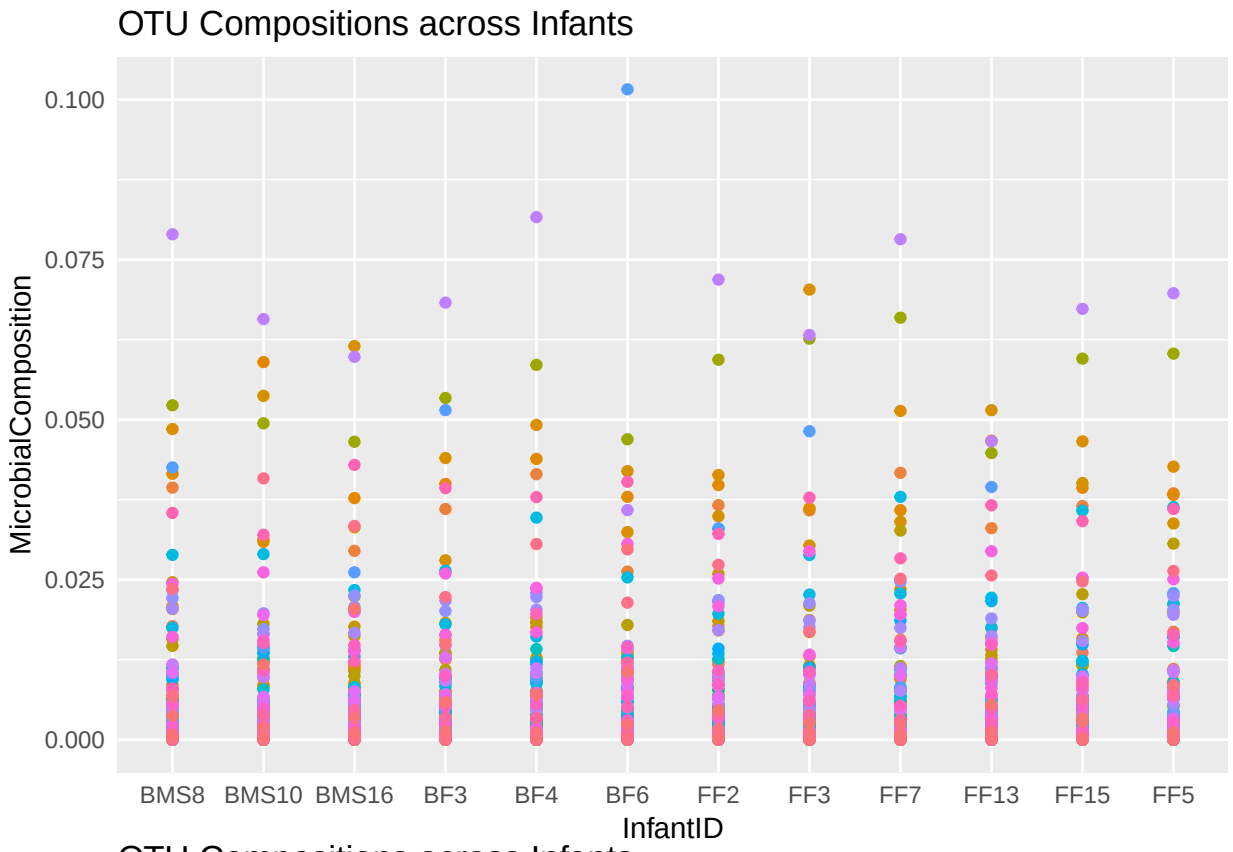
```
head(OTU)
```

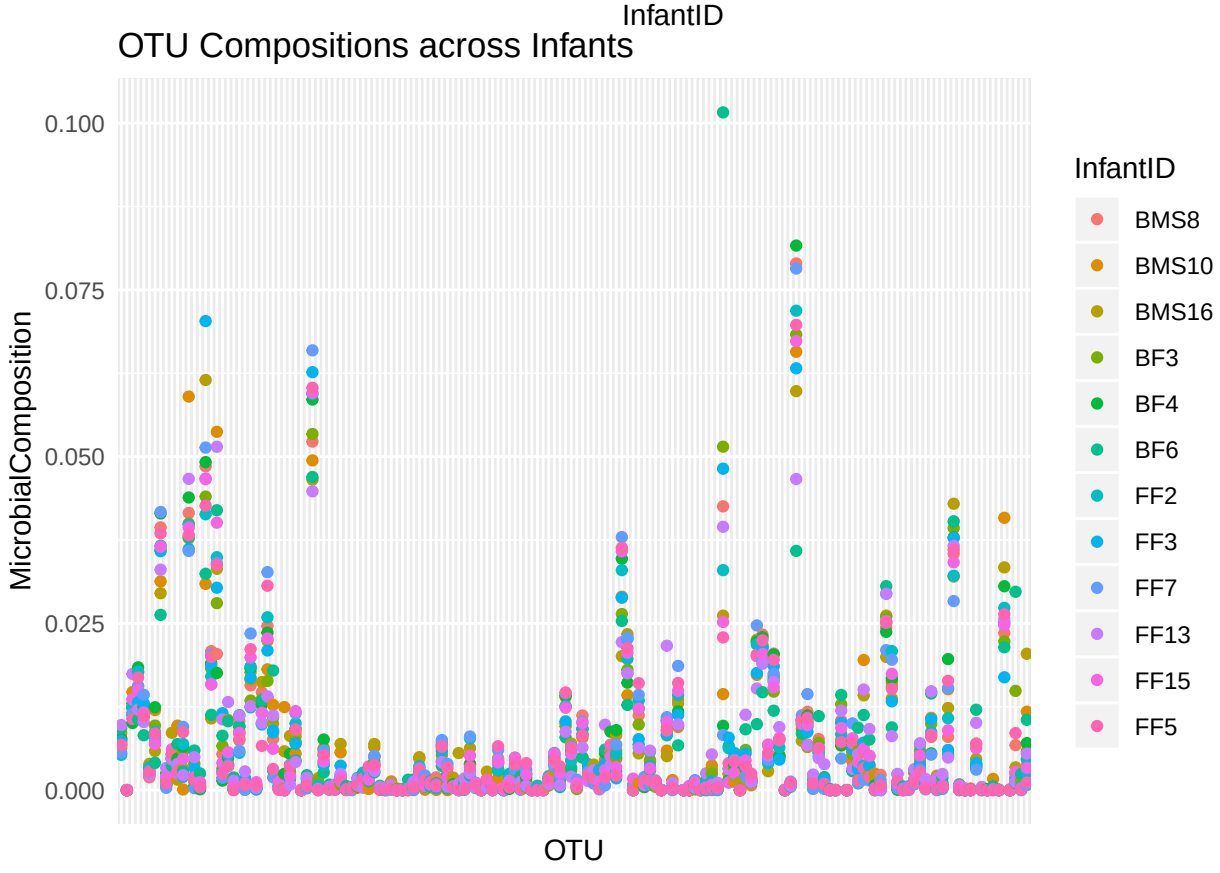
```
## [1] "Amino Acids and Derivatives_Alanine, serine, and glycine"
## [2] "Amino Acids and Derivatives_Amino Acids and Derivatives"
## [3] "Amino Acids and Derivatives_Arginine; urea cycle, polyamines"
## [4] "Amino Acids and Derivatives_Aromatic amino acids and derivatives"
## [5] "Amino Acids and Derivatives_Branched-chain amino acids"
## [6] "Amino Acids and Derivatives_Glutamine, glutamate, aspartate, asparagine; ammonia assimilation"
```

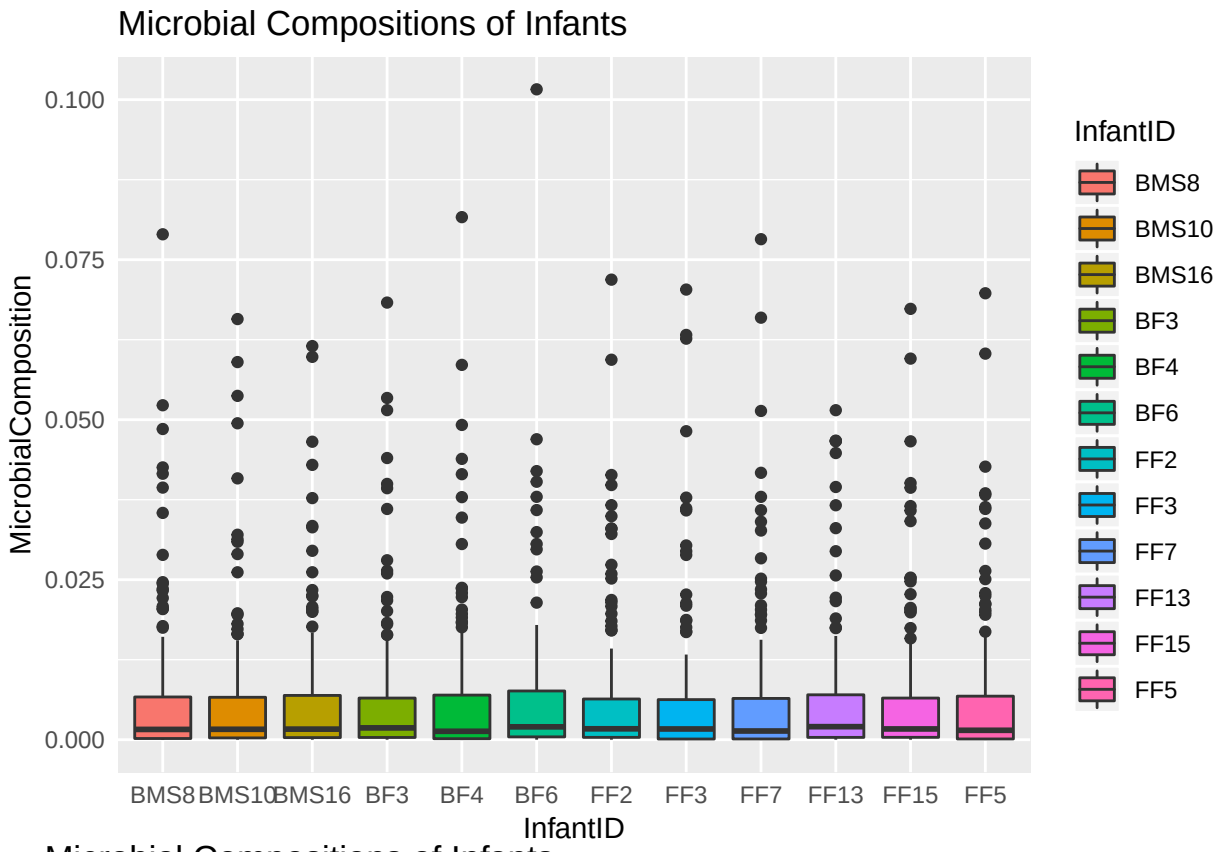
```
tail(OTU)
```

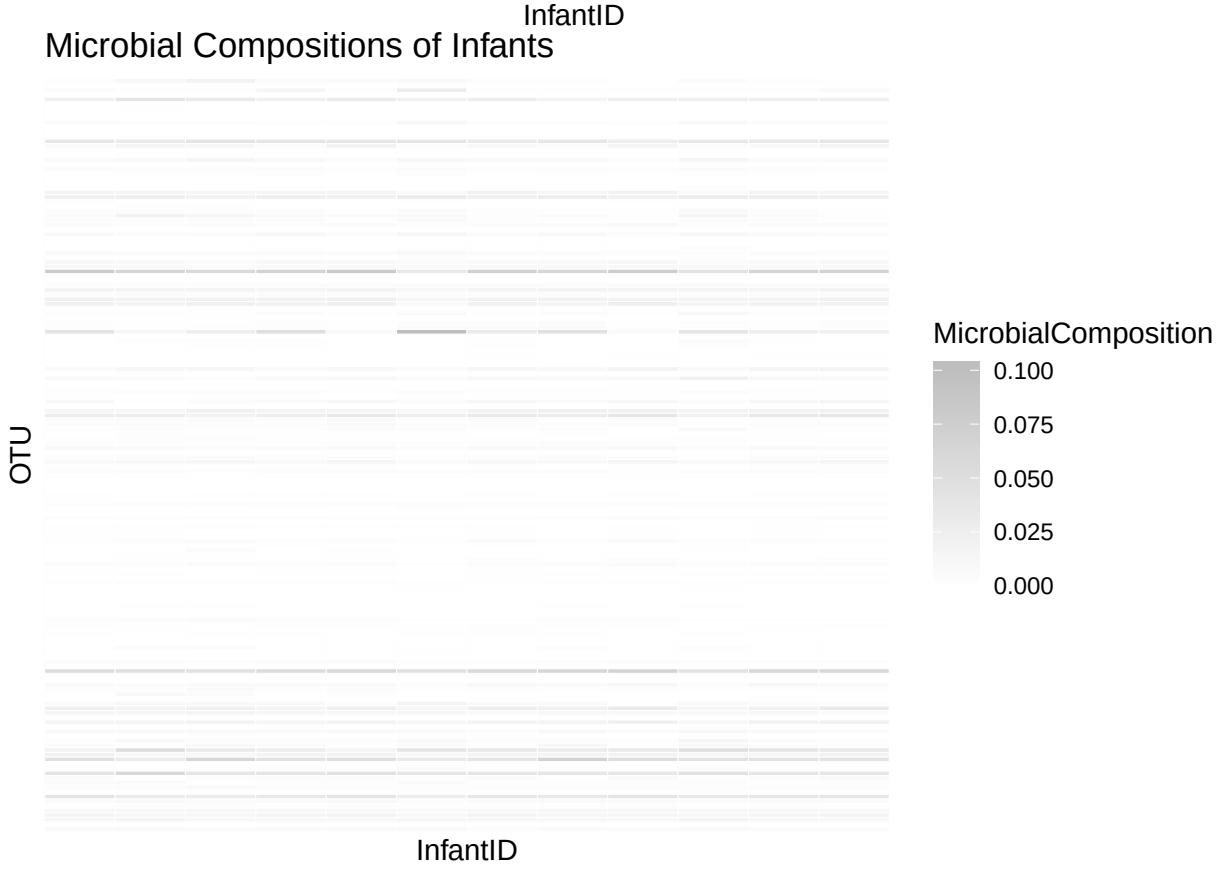
```
## [1] "Virulence_Regulation of virulence"
## [2] "Virulence_Resistance to antibiotics and toxic compounds"
## [3] "Virulence_Toxins and superantigens"
## [4] "Virulence_Type III, Type IV, ESAT secretion systems"
## [5] "Virulence_Type VI secretion systems"
## [6] "Virulence_Virulence"
```

The raw data is strongly right skewed and spare with many zeros.


OTU Compositions across Infants
MicrobialComposition
0.100
0.075
0.050
0.025
0.000
BMS8
BMS10
BMS16
BF3
BF4
BF6
FF2
FF3
FF7
FF13
FF15
FF5
InfantID


OTU Compositions across Infants
MicrobialComposition
0.100
0.075
0.050
0.025
0.000
OTU
InfantID
BMS8
BMS10
BMS16
BF3
BF4
BF6
FF2
FF3
FF7
FF13
FF15
FF5

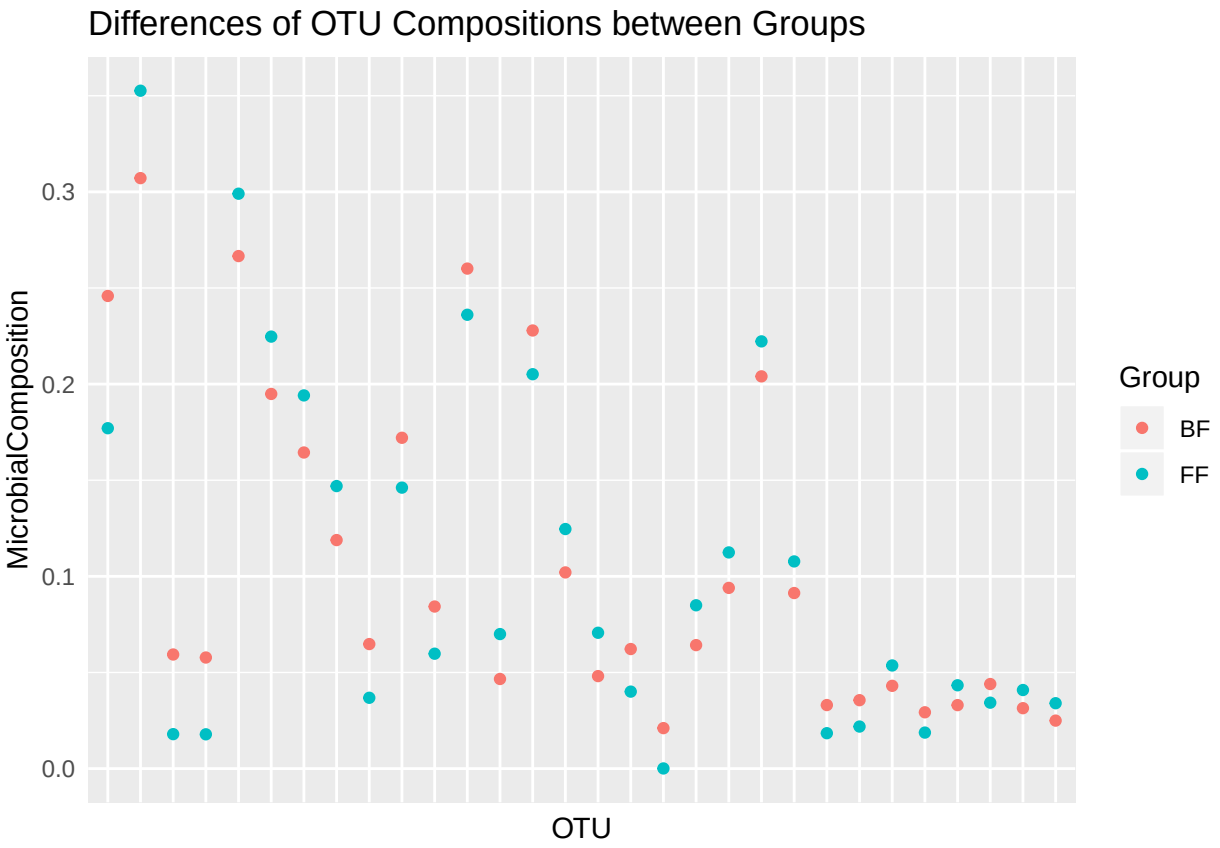




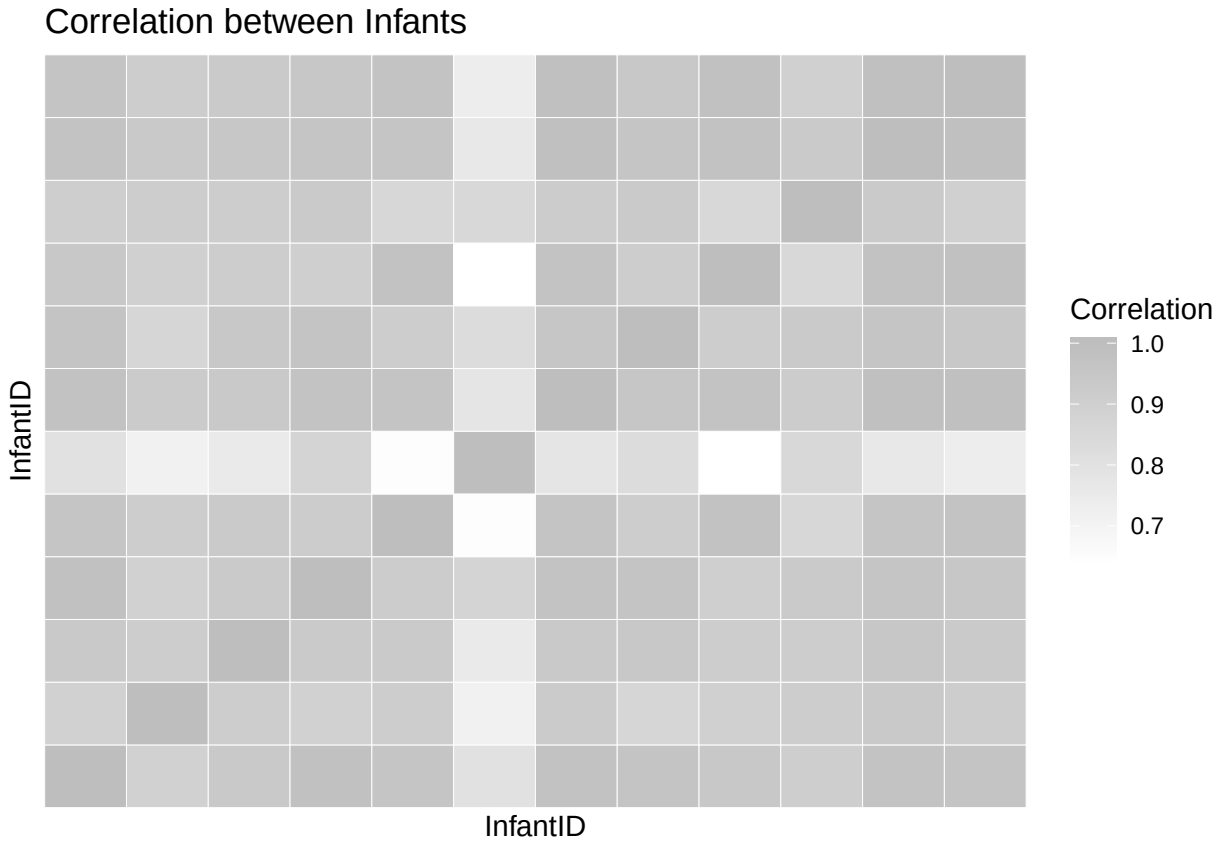
Next we want to explore whether microbial compositions differ between infant groups and show top 30

prominent differences.

```
##  [1] "Miscellaneous_Miscellaneous"
##  [2] "Clustering-based subsystems_Clustering-based subsystems"
##  [3] "Virulence_Virulence"
##  [4] "Virulence_Type III, Type IV, ESAT secretion systems"
##  [5] "Carbohydrates_Di- and oligosaccharides"
##  [6] "Carbohydrates_Monosaccharides"
##  [7] "DNA Metabolism_DNA repair"
##  [8] "Cell Wall and Capsule_Cell Wall and Capsule"
##  [9] "Respiration_Electron donating reactions"
## [10] "Virulence_Resistance to antibiotics and toxic compounds"
## [11] "Cell Wall and Capsule_Capsular and extracellular polysacchrides"
## [12] "Carbohydrates_Central carbohydrate metabolism"
## [13] "Fatty Acids and Lipids_Fatty acids"
## [14] "Unclassified_Unclassified"
## [15] "DNA Metabolism_DNA replication"
## [16] "Membrane Transport_ABC transporters"
## [17] "Cell Wall and Capsule_Gram-Negative cell wall components"
## [18] "Cell Wall and Capsule_Multi-enzyme complex"
## [19] "Membrane Transport_Membrane Transport"
## [20] "Cell Division and Cell Cycle_Cell cycle in Prokaryota"
## [21] "Amino Acids and Derivatives_Lysine, threonine, methionine, and cysteine"
## [22] "Carbohydrates_Fermentation"
## [23] "Carbohydrates_Aminosugars"
## [24] "DNA Metabolism_DNA Metabolism"
## [25] "Clustering-based subsystems_Cell Division"
## [26] "Respiration_Electron accepting reactions"
## [27] "Potassium metabolism_Potassium metabolism"
## [28] "Protein Metabolism_Secretion"
## [29] "Respiration_ATP synthases"
## [30] "Carbohydrates_Carbohydrates"
```



From above plots, we observe that the composition of OTUs behave differently between two groups of BF infants and FF infants. The most significant differences appear to be virulence, carbohydrates and so on. We show correlation structures of infants and OTUs respectively.



The correlation structure between infants seems to indicate that the behavior of BF6 differs greatly from others. If we perform K-means clustering with two centers, all infants except BF6 wiill be grouped into a single cluster.

```
kmeans(WideData, centers = 2)$cluster
```

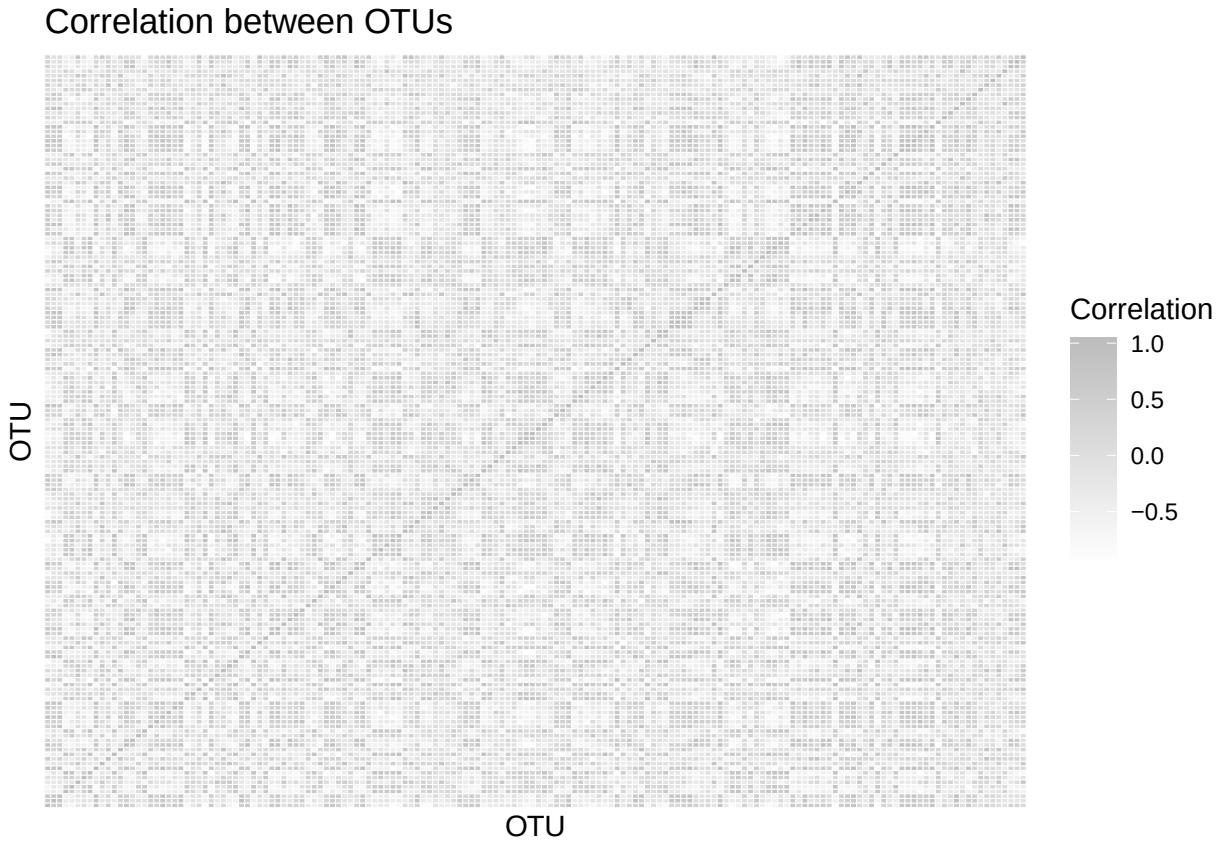
```
##  BMS8 BMS10 BMS16   BF3   BF4   BF6   FF2   FF3   FF7  FF13  FF15   FF5
##     1     1     1     1     1     2     1     1     1     1     1     1
```

```
kmeans(WideData, centers = 3)$cluster
```

```
##  BMS8 BMS10 BMS16   BF3   BF4   BF6   FF2   FF3   FF7  FF13  FF15   FF5
##     1     2     1     1     2     3     2     1     2     1     2     2
```

```
kmeans(WideData, centers = 4)$cluster
```

```
##  BMS8 BMS10 BMS16   BF3   BF4   BF6   FF2   FF3   FF7  FF13  FF15   FF5
##     3     2     2     3     4     1     4     3     4     2     4     4
```



The relationship between OTUs seems vague from the plot.

Model formulation

To explain the variation of microbial compositions between groups and across infants, we propose the Bayesian double feature allocation using the count data matrix. The model will infer latent features that are associated with both OTUs and infants. At the same time, the result can be regarded as overlapping clustering for OTUs and infants simultaneously. Figure 1 illustrates the formation of our model.

Suppose that there exists an OTU-latent matrix $\boldsymbol{A} = (a_{ik}) \in \{0,1\}^{p\times K}$ which is assigned an Indian buffet (IBP) prior. IBP is a distribution over binary matrices with infinitely many columns with a parameter α that controls the sparsity of the matrix. The process is described by imagining an Indian buffet offering an infinite number of dishes. Each customer entering the restaurant chooses the dishes that have been already sampled by other customers with probability proportional to their popularity. Then he also tries a number of new dishes dependent on the parameter α. Customers are exchangeable and dishes are independent. For the current model, customers correspond to OTUs and dishes correspond to latent features. Given the number of columns K of $\boldsymbol{A}$, each elements of the infant-latent matrix $\boldsymbol{B} = (b_{jk}) \in \{0,1\}^{n\times K}$ follows independent Bernoulli distribution Bernoulli(ρ). A Beta(α_ρ, β_ρ) pior is assigned to parameter ρ.

Suppose that we also have a weight matrix $\boldsymbol{W} = (w_{jk}) \in \mathbb{R}_+^{n\times K}$ and a residual vector $\boldsymbol{e} = (e_j) \in \mathbb{R}^n$, each element of which follows independent Gamma($1, \beta_w$) distribution and Normal($0, \sigma_e^2$) distribution respectively. Given $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{W}$ and $\boldsymbol{e}$, each element z_{ij} of the latent matrix $\boldsymbol{Z} \in \{0,1\}^{p\times n}$ is modeled as

$$z_{ij}|\{a_{ik}\},\{b_{jk}\},\{w_{jk}\},e_j \sim \text{logit}\left(\sum_{k=1}^{K} a_{ik}w_{jk}b_{jk} + e_j\right),$$

where $\text{logit}(x) = e^x/(1+e^x)$. Here z_{ij} can be used to indicate that OTU i is relative abundant ($z_{ij} = 1$) and scarce ($z_{ij} = 0$) in infant j.

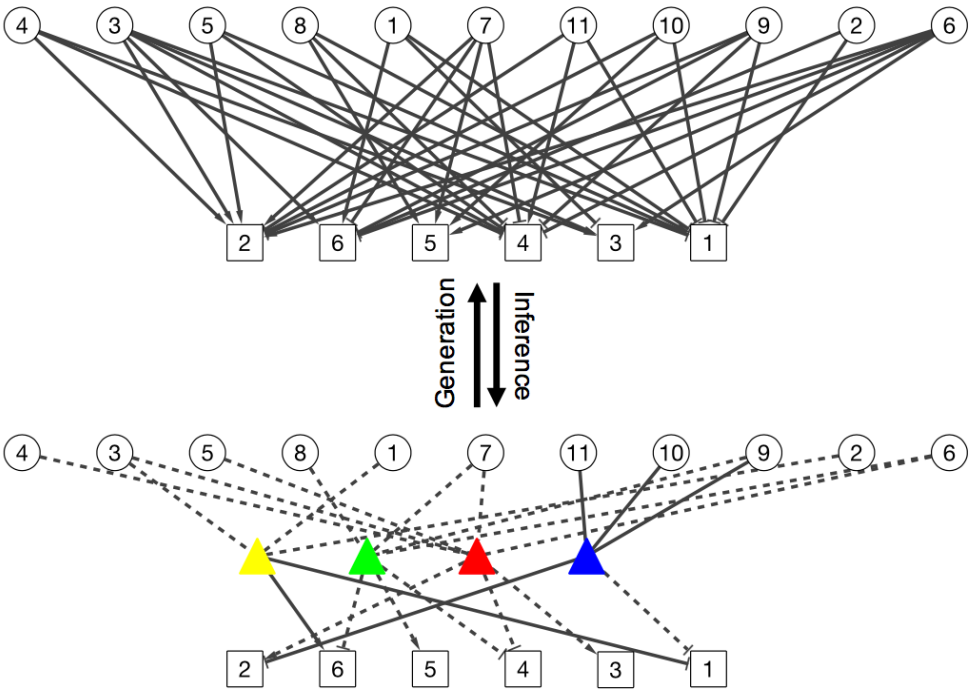


Figure 1: Illustration of the model

In high throughput sequencing, data obtained are count compositions since the capacity of the machine determines the number of reads observed. These reduce to probabilities of observing a feature given the sequencing depth. To this end, our sampling model assumes that each column $\boldsymbol{x}_j$ of the observation matrix $\boldsymbol{X} \in \{\mathbb{R}_+ \cup 0\}^{p \times n}$ follows the multinomial distribution

$$\boldsymbol{x}_j \sim \text{multinomial}(n_j, \boldsymbol{\pi}_j), \quad \boldsymbol{\pi_j} = \frac{\boldsymbol{r_j}}{\sum \boldsymbol{r}_j} = \frac{(r_{1j}, \ldots, r_{pj})}{\sum_{i=1}^{p} r_{ij}},$$

where $n_j = \sum_{i=1}^{p} x_{ij}$. The distribution of r_{ij} in $\boldsymbol{R} \in \mathbb{R}_+^{p \times n}$ depends on the latent indicator

$$r_{ij}|\theta_i, z_{ij} = 1 \sim \text{Gamma}\left(\theta_i + 1, 1\right), \quad r_{ij}|\theta_i, z_{ij} = 0 \sim \text{Gamma}\left(\frac{1}{\theta_i + 1}, 1\right).$$

The prior on $\boldsymbol{\pi}_j$ is then the Dirichlet distribution. We finally put independent Gamma$(\alpha_\theta, \beta_\theta)$ on each $\theta_i \in \boldsymbol{\theta} \in \mathbb{R}_+^p$.

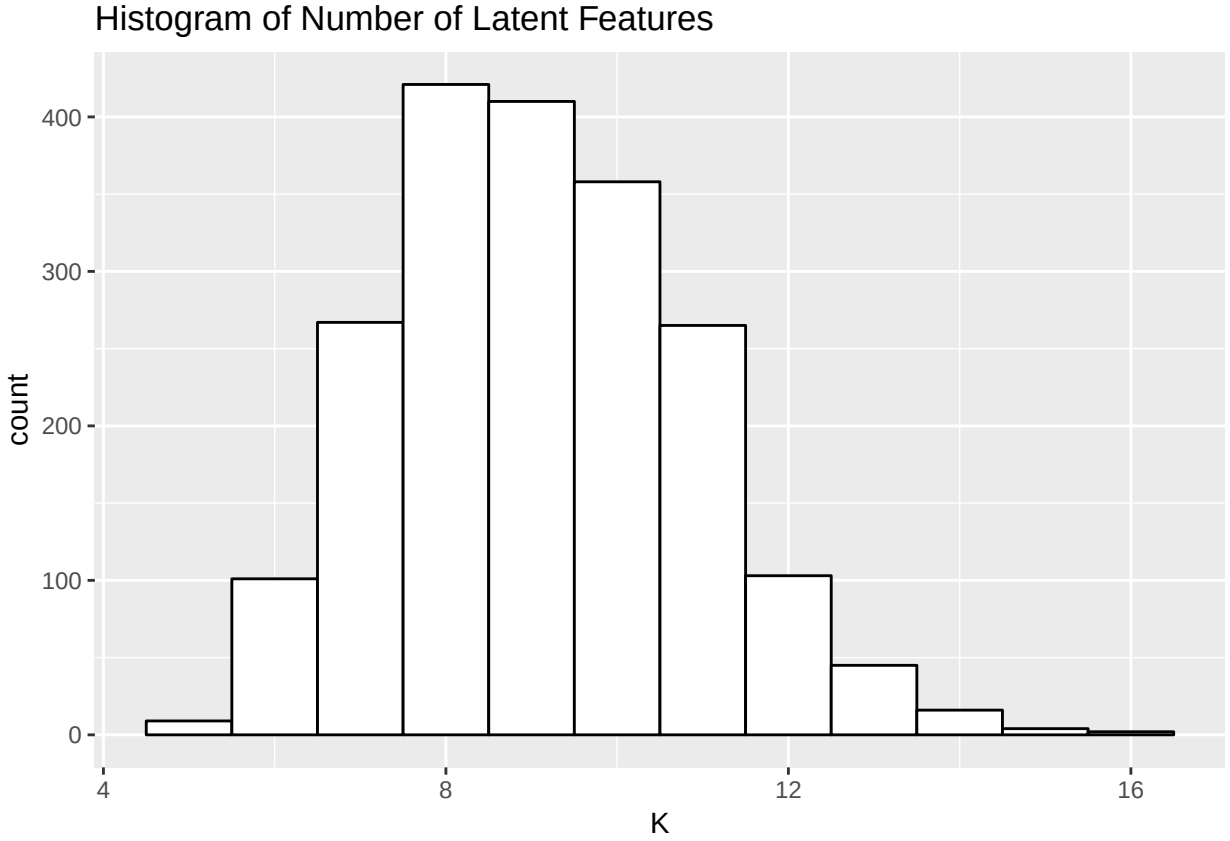
Model inference

The inference procedure is based on markov chain Monte Carlo (MCMC) method, especially Metropolis-Hastings within Gibbs sampling. All parameters expect $\boldsymbol{A}$ can be sampled based on their full conditional distribution or through a Metropolis-Hastings step. Updating $\boldsymbol{A}$ includes sampling existing entries and proposing new latent features based on the Indian buffet construction. The proposed new features are accepted or rejected based on a Metropolis-Hastings step together with associated parameters in $\boldsymbol{B}$ and $\boldsymbol{W}$ drawn from the corresponding prior.

We run the proposed MCMC algorithm for 20000 iteration. The first half samples are discarded as burin-in and posterior samples are retained at every fifth iterations. We remove insignificant features from the result

if it only contains single one element. The posterior mode of the number of latent features occurs at $K = 8$ or $K = 9$ with probability 0.21 and 0.20 respectively. Here we choose $K = 8$.

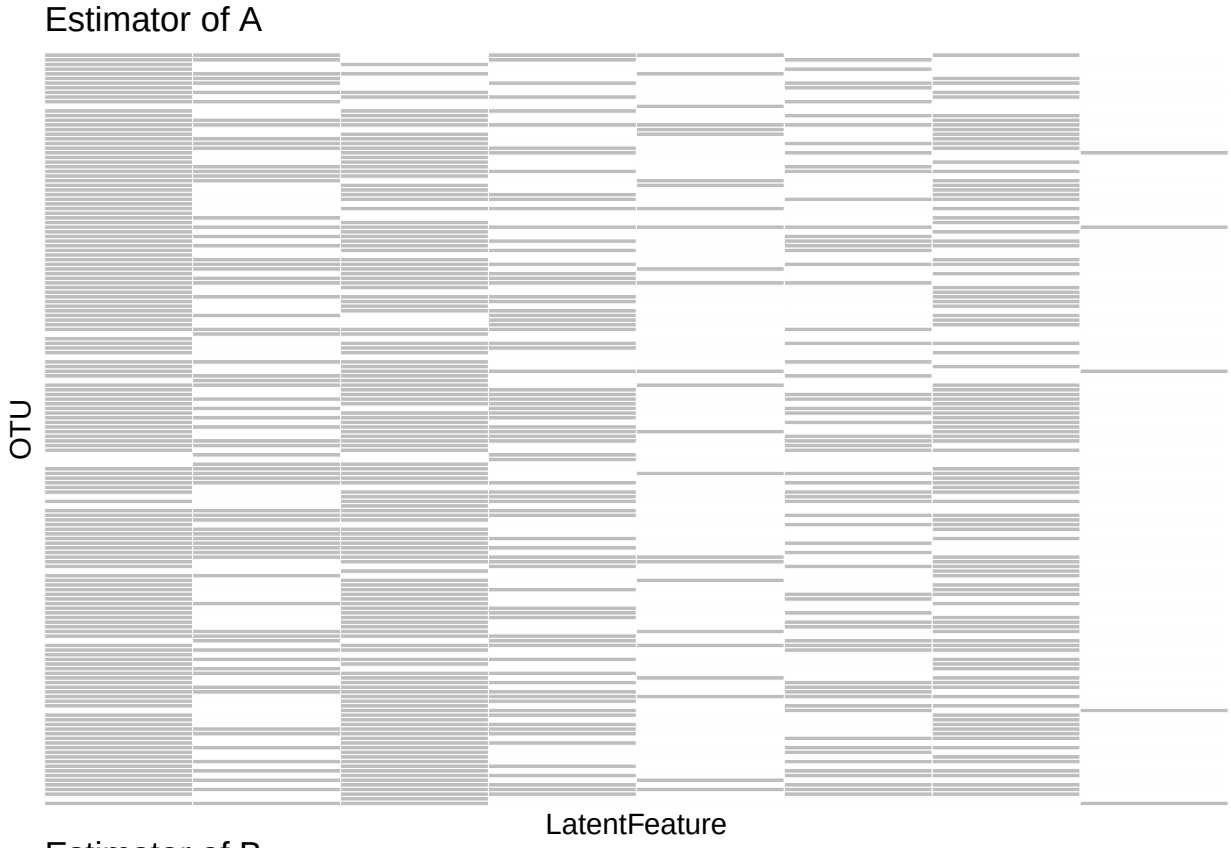
```
## K
##            5            6            7            8            9
## 0.0044977511 0.0504747626 0.1334332834 0.2103948026 0.2048975512
##           10           11           12           13           14
## 0.1789105447 0.1324337831 0.0514742629 0.0224887556 0.0079960020
##           15           16
## 0.0019990005 0.0009995002
```

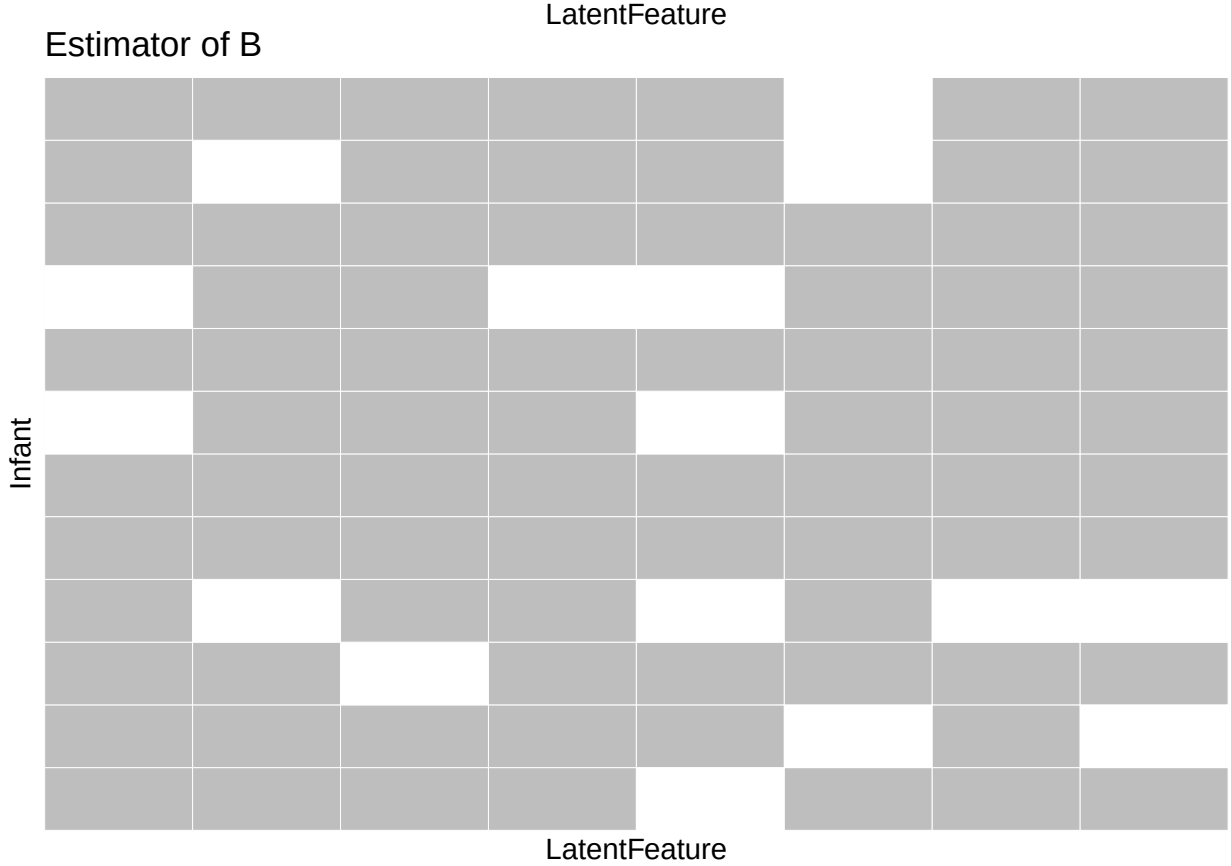


Given K, we find the least squares estimator $\boldsymbol{A}$ by the following procedure. For any two binary matrices $\boldsymbol{A}$ and $\tilde{\boldsymbol{A}}$, we define the distance $d(\boldsymbol{A}, \tilde{\boldsymbol{A}}) = \min_\pi \mathcal{H}(\boldsymbol{A}, \pi(\tilde{\boldsymbol{A}}))$, where $\pi(\tilde{\boldsymbol{A}})$ denotes a permutation of the columns of $\tilde{\boldsymbol{A}}$ and $\mathcal{H}(\cdot, \cdot)$ is the Hamming distance of two binary matrices. A point estimate $\boldsymbol{A}$ is then obtained as

$$\boldsymbol{A} = \arg\max_{\boldsymbol{A}} \int d(\tilde{\boldsymbol{A}}, \boldsymbol{A}) dp(\tilde{\boldsymbol{A}} | \boldsymbol{X}, K).$$

Both, the integral as well as the optimization can be approximated using the available Monte Carlo MCMC samples, by carrying out the minimization over $\tilde{\boldsymbol{A}} \in \{\boldsymbol{A}_t,\ t = 1, \ldots, T\}$ and by evaluating the integral as Monte Carlo average. The posterior point estimators of other parameters are obtained as posterior means conditional on $\boldsymbol{A}$. We evaluate posterior means using the posterior Monte Carlo samples.





If we treat the feature allocation matrix $\boldsymbol{A}$ as the overlapping clustering matrix, that is, when $a_{ik} = 1$, we

assign OTU i to the k-th cluster. Similar explanations can be used to illustrate the result of $\boldsymbol{B}$.

We run several Markov chains with different initializations and the gelman.diag function in R shows the sign of convergence of the number of latent features with the upper limit of potential scale reduction factor close to 1.

Comment and discussion

The recovered latent features are hard to explain from a biological perspective. Currently, there are two main problems in the model. The first one is that the sampling distribution is not quite appropriate for the data at hand, which results in poor recovered latent indicators. The issue is partly due to the high variance of components corresponding to $z = 1$ compared to those corresponding to $z = 0$. The second one is that the number of operational taxonomic units is quite large to form meaningful groups under the Indian buffet prior, which always prefer a few large clusters and many small clusters. Available prior information can be incorporated into the model and the polygenetic Indian buffet process may be used to encourage similar latent features between closer individuals. Moreover, the identification problem of the Indian buffet process may make it hard to explain the data in a meaningful way. We may replace the Indian buffet process with the determinantal point process, which presents a repulsive prior on latent feature components.