PharmaSUG 2018 - Paper AA-20

Statistical Analysis and Visualization of Microbiome data in Clinical Trials

Athira Sudhakaran, Limna Salim, Genpro Life Sciences, Thiruvananthapuram, India

ABSTRACT

Radical changes across Microbiome research in Clinical Trials, has materialized with the widespread application of next-generation sequencing technologies. This is empowered by high-throughput profiling of the genetic contents of microbial communities. With the advent of big data technologies and high throughput computing resources, large complex microbiome data sets can be easily analyzed. Research is ongoing to identify new methods and models suitable for analyzing microbiome data. Some of the areas that need attention includes

a. Analysis of microbiome data and tools to explore its composition
b. Longitudinal data analysis
c. Causation Analysis

Since data is extremely complex, it remains a challenge for statisticians to develop tools required for such analysis. This paper explores the possibilities of developing a dynamic software framework using Angular JS, SAS®, R, Python® and Rasa NLU for easy and effective analysis of microbiome data sets.

This paper specifically details five different statistical/machine learning methods that help researchers to understand microbiome data structure, analyze it and visualize it with ease. Some of the features includes

- Tabular presentations and visualization of the abundance and richness of Operational Taxonomic Units (OTU)
- Identify significant Operational Taxonomic Unit within each group
- Graphical comparison between different groups of microbiome data based on the significant Operational Taxonomic Unit
- Comparison and Visualization of microbiome data between two groups based on the distance matrix
- Minimizing bias introduced to the analysis of blinded microbial data among different groups, using an unsupervised learning technique such as cluster analysis

INTRODUCTION

The gut microbiota, the ecosystem of about 10^{14} microbial cells in the intestine, is now regarded as a symbiotic essential organ of the human body responsible for functions that human cells cannot carry out (Moya and Ferrer 2016). The gut microbiota plays a significant role in human health and disease by affecting nutrient utilization, colonization resistance, development of the immune system, modulating host metabolism, and other diverse aspects of the host's physiology (Sommer and Bäckhed 2013). Changes in the microbiota, called dysbioses, are correlated with a variety of "lifestyle diseases," such as obesity, metabolic syndrome, diabetes, and even allergies and asthma (Patel and DuPont 2015).

Human gut microbiome studies helps researchers to understand the microbiome community composition along with the interactions among microbiomes belonging to different groups. Here, the statistical tests to summarize the microbiome data and to compare different groups are discussed

Planned analysis also include, descriptive reports that will provide an overview of the nature of data in hand.

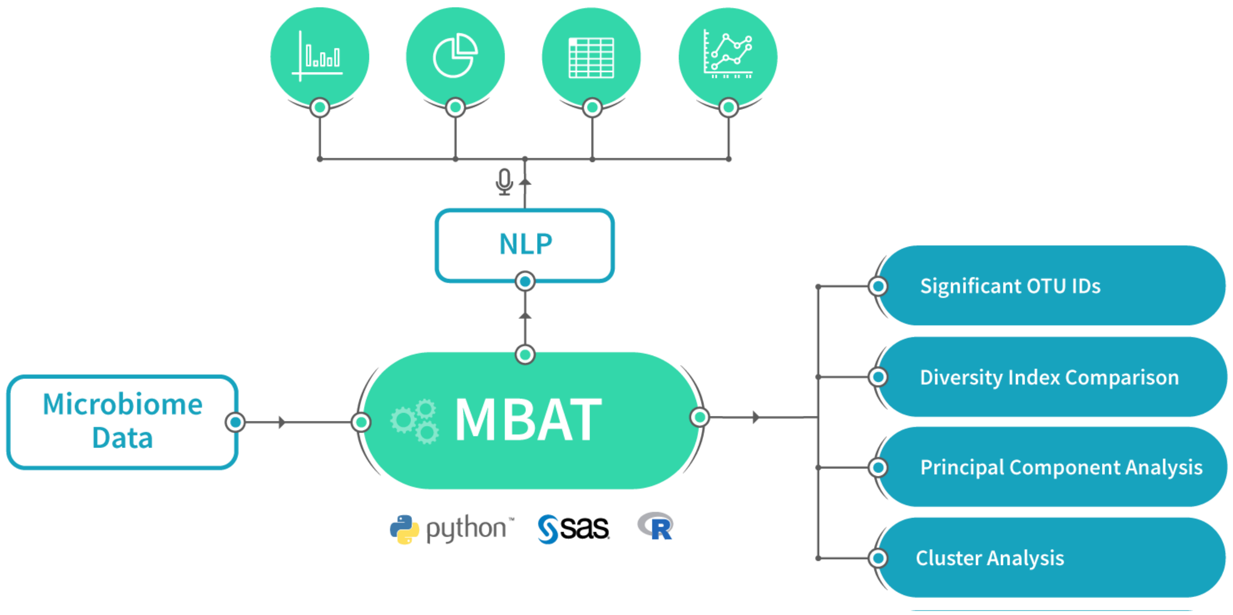

Figure 1.Graphical representation for the analysis

As explained in Figure 1, MBAT (Microbiome Analysis Tool kit) is a web based application which will combine the features of Angular JS, SAS, R, Python and Rasa NLU. This application will feature all the statistical methods that are explained in the paper. MBAT allows access to the microbiome data as a csv file. It allows the user to perform the following analysis

1. Identification of Significant OTU IDs
2. Diversity Index Comparison
3. Principal Component Analysis
4. Cluster Analysis

This analysis will be performed based on different taxonomic rank (Kingdom /Phylum/Class/Order/Family /Genus/Species).

SIGNIFICANT OPERATIONAL TAXONOMIC UNIT

This section will explain the statistical methods that are employed to identify significant operational taxonomic units within each group.

A cross over study data of 5 subjects is considered to compare the effect of treatment and placebo group with respect to microbial signature from gut. As a first step to this analysis, the treatment and placebo groups are compared with respect to all the OTU ids in the gut and it is observed that there is a significant difference among the microbial signatures between the two. Wilcoxon test is used to compare the effect of means of two samples.

Further, a second stage analysis was performed to identify the individual OTU ids which were significantly different between the treatment and placebo groups. Wilcoxon test was used to compare treatment and Placebo group with respect to microbial signature in the gut.

Below is sample code that illustrates the syntax:

```
PROC NPAR1WAY DATA=microbiome WILCOXON;
  CLASS treatment;
  VAR microbe;
  BY OTUid;
  EXACT WILCOXON;
RUN;
```

The distribution of each significant OTU id is then graphically investigated using a bar diagram. Figure 2 is an example of the graphical representation of significant OTU ids. The adjoining stacked bar diagram shows the proportion of the OTU id named OTU1000 in the treatment and placebo group along with the p value indicating significance. For this particular OTU id, which is under the Bacteroidetes Phyla level, the proportion in placebo is much higher than the proportion in treatment group.

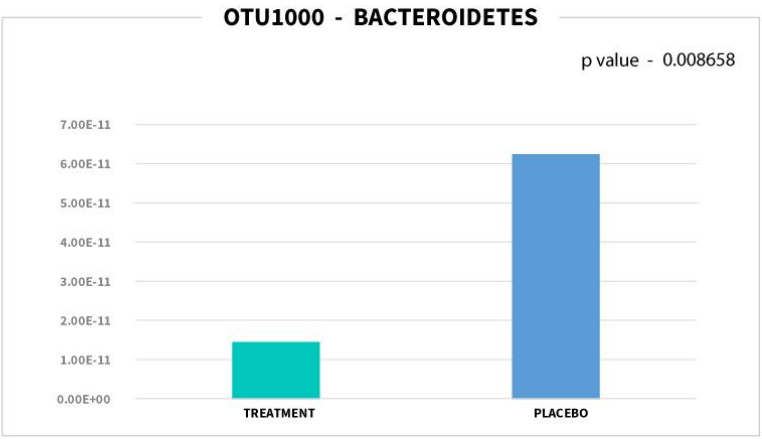

Figure 2. Graphical Representation of significant OTU id "OTU1000"

ABUNDANCE AND RICHNESS OF TAXONOMIC UNIT

This section will detail the tabular presentation and visualization of the abundance and richness of operational taxonomic units. Some of the key metrics that we have captured as part of the analysis include

1. Richness
2. Abundance
3. Dominance
4. Diversity index

Different graphs in the form of bar diagrams or pie charts will be presented detailing the above mentioned characteristics for comparing all/significant OTU ids.

Richness is the number of species in each sample. Figure 3 explains the richness of different phyla for the entire OTU ids (6696) and for the significant OTU ids (288)

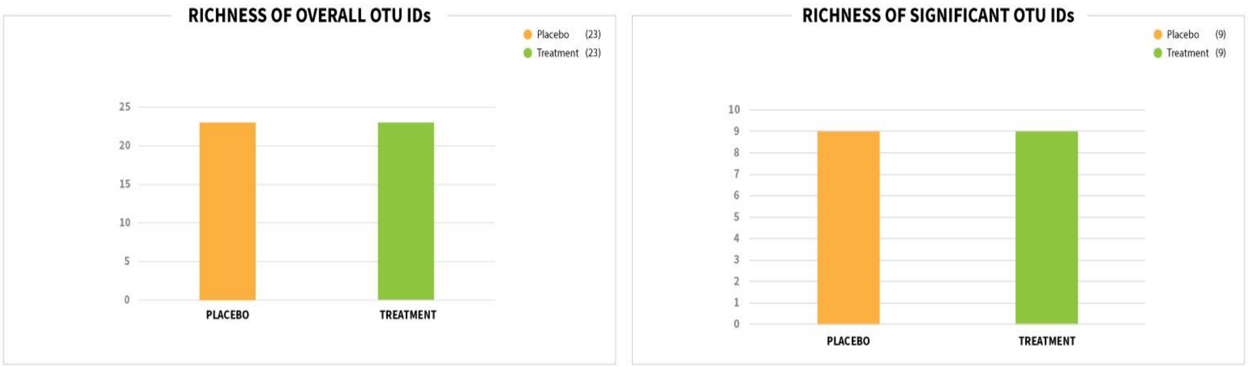

Figure 3. Graphical Representation for Richness

Figure 3 depicts that the richness for the entire OTU ids between Treatment and Placebo group are the same (23%). Also the richness for the significant OTU ids between Treatment and Placebo are the same (9%). This indicates that there is no significant difference in the species status between treatment and placebo group.

Figure 4 explains the abundance of different phyla for the entire OTU Ids (6696) and for the significant OTU Ids (288) along with the phyla level. Abundance between treatment and placebo group is also the same. The most abundant Phyla within the entire significant OTU ids is Firmicutes (33.49%).The most abundant Phyla within the significant OTU ids is also Firmicutes (32.98%)

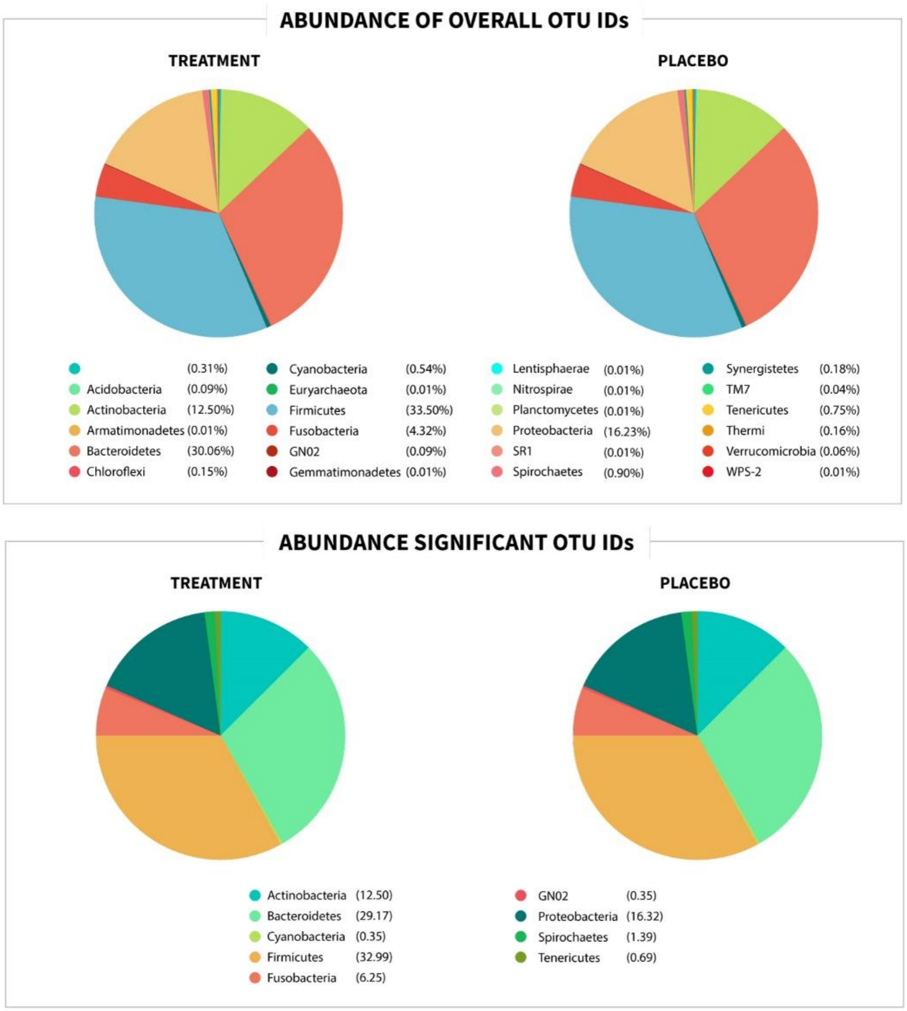

Figure 4. Graphical Representation for Abundance

Figure 5 explains the Dominance of different phyla for the entire OTU ids (6696) along with the phyla level. Dominance is different between the treatment and placebo group. When you consider the entire OTU ids the most dominant Phyla within the treatment group and the placebo group is Firmicutes (28.85% for Treatment and 42.74% for Placebo).

Figure 6 explains the Dominance of different phyla for the significant OTU ids (288) along with the phyla level. Dominance is different between treatment and placebo group. Within the significant OTU ids the most dominant Phyla is Tenericutes in treatment group (41.65%) and Firmicutes (72.94%) in Placebo group

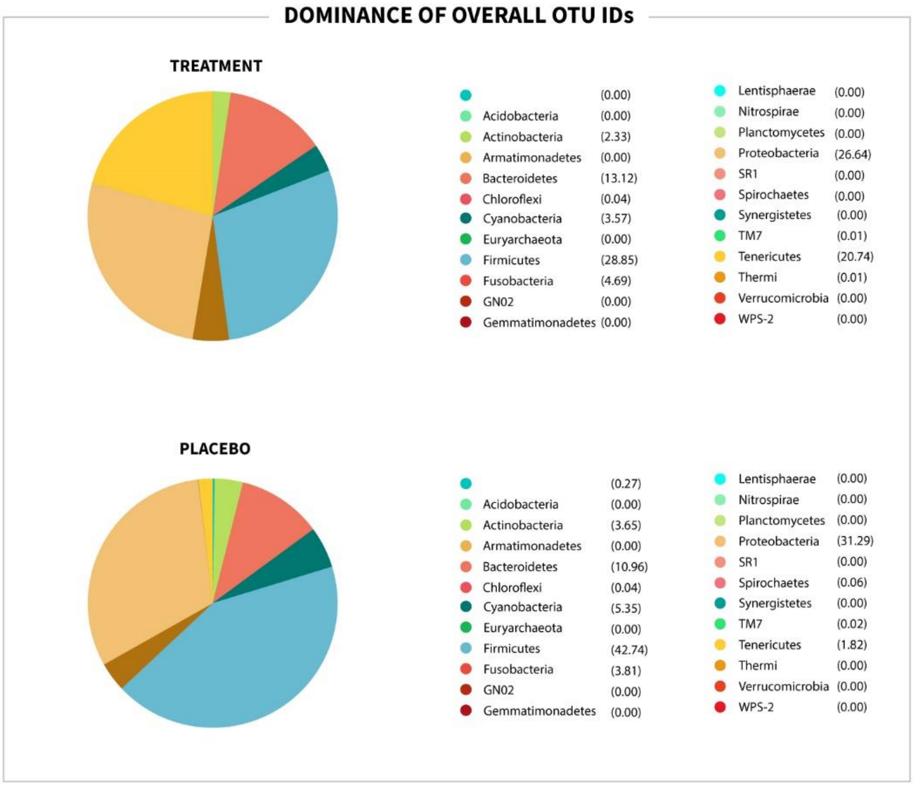

Figure 5. Graphical Representation for Dominance of Overall OTU IDs

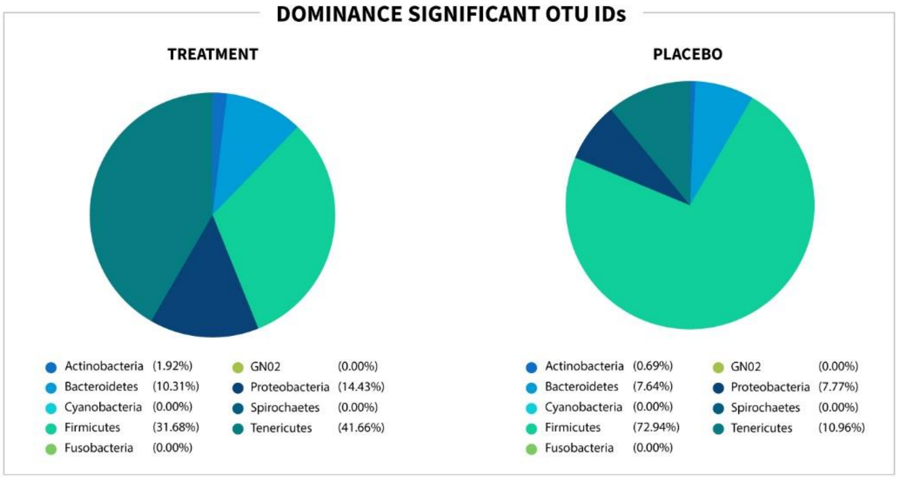

Figure 6. Graphical Representation for Dominance of Significant OTU IDs

Figure 7 explains the diversity index of different treatment groups for the significant OTU ids (288). Diversity is a weighted average of the proportions of each species present. In this case we are using the Shannon Index.

From Figure 7 it is clear that the treatment group is slightly more diverse than the placebo group. The diversity index for treatment group is 1.29 and the diversity index for placebo group is 1.05.

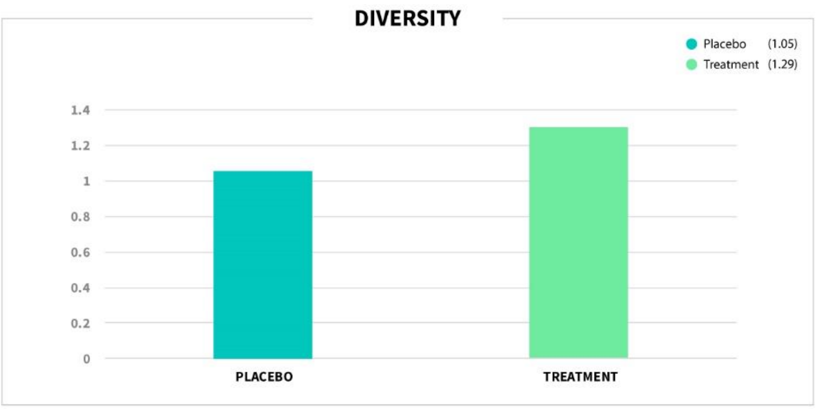

Figure 7. Graphical Representation for Diversity

COMPARISON BASED ON THE DISTANCE MATRIX

This section explains the comparison and visualization of microbiome data between two groups based on the distance matrix.

Comparison of microbiome data for each subject is performed using distance matrix principal coordinate analysis plot. This analysis takes distance matrices as input with corresponding points and transforms the second coordinate set, by rotating, scaling, and translating it to minimize the distances between corresponding points for the treatment and placebo group.

Based on the available microbiome data the corresponding Euclidean distance matrix was calculated using the following syntax:

```
PROC DISTANCE DATA=microbiome OUT=Distance_Matrix METHOD=Euclid;
   VAR interval (OTU0--OTU999 / std = Std);
   ID treatment;
RUN;
```

This distance matrix was then used to calculate the different principal coordinate values:

```
PROC PRINCOMP OUT= Distance_Matrix plots= score (ellipse ncomp=3);
  ODS OUTPUT Eigenvectors=pca Eigenvalues=proportion;
RUN;
```

These coordinates were plotted as a 3 dimensional figure, with bars connecting the corresponding treatment and placebo groups of each subject. This will explain the distance or change in microbial signatures between two groups of each subject.

This analysis helps to determine whether the same beta diversity conclusions can be derived regardless of which metric is used to compare the samples.

Figure 8 explains the Human Microbiome diversity between treatment and placebo group of 6 subjects.

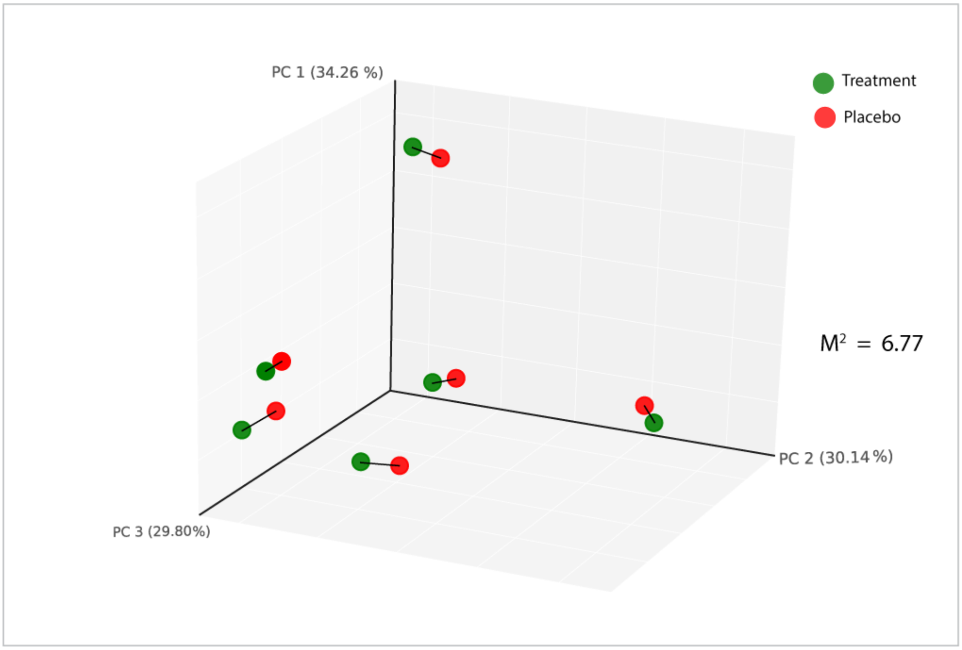

Figure 8. Graphical Representation for comparison between two groups

Figure 8 is a graphical explanation to understand the nature of the data per subject. As shown in figure, 94.2 percentage of variance is explained by the first three principal components together.Figure 8 clearly says that there is no much difference (distance) between the treatment and placebo group of each subject. Mean square value (6.77) associated with it also helps to confirm the result.

MINIMIZING BIAS

This section explains how to minimize bias introduced in the analysis of blinded microbial data among different groups, using an unsupervised learning analysis such as cluster analysis.

Computational biologists performing unsupervised learning analysis will be blinded to the study product assignment until after sample clustering is performed.

Using cluster analysis, the microbial community composition from the most recent non-missing samples of each period will be used to assess whether samples can be clustered into groups representing the two study product groups. The clustering will be conducted prior to informing the computational biologist about the assignment of subjects into groups (no grouping information will be provided). K-means clustering was used to cluster the samples within each treatment group.

The clustering will be based on Euclidean distances. So the cluster centers are based on least squares estimation within each of the treatment groups. The following statements standardize the variables and perform a cluster analysis on the standardized data:

```
PROC STDIZE DATA=AD_tran OUT=Stand METHOD=STD;
  VAR OTU0--OTU999;
RUN;

PROC FASTCLUS DATA=Stand out=Clust MAXCLUSTERS=2;
  VAR OTU0--OTU999;
RUN;
```

After clustering is performed, the computational biologist will be informed of the subject treatment groupings and assess the accuracy with which the clusters recapitulate the treatment groups.

CONCLUSION

In this paper the possibilities of using SAS to put together an analysis framework for microbiome data sets are explored. The five different statistical/machine learning methods explained will help researchers to understand microbiome data structure, analyze it and visualize it with ease. With the help of MBAT, the researchers in the microbiome field can easily handle the microbiome data using interactive visualization. We will be able to further add more analysis using the dynamic software framework created.

REFERENCES

Moya A, Ferrer M, 2016 May 24, Functional Redundancy-Induced Stability of Gut Microbiota Subjected to Disturbance, (5):402-413

Sommer F, Bäckhed, 2013 Apr 11, The gut microbiota--masters of host development and physiology, (4):227-38

Patel R, DuPont HL, 2015 May 15, New approaches for bacteriotherapy: prebiotics, new-generation probiotics, and synbiotics

Sharma NS, Wille KM, Athira S, Zhi D, Hough KP, Diaz-Guzman E, Zhang K, Kumar R, Rangarajan S, Eipers P, Wang Y, Srivastava RK, Rodriguez Dager JV, Athar M, Morrow C, Hoopes CW, Chaplin DD, Thannickal VJ, Deshane JS, Distal airway microbiome is associated with Immunoregulatory myeloid cell responses in lung transplant recipients, 2017 Jul 15

ACKNOWLEDGMENTS

The content, ideas and recommendations presented in this paper are all developed from experiences in our career. These experiences come through previous companies, various industry leaders, colleagues, mentors, conferences and direct experience.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Athira Sudhakaran
Company: Genpro Life Science India Private Ltd
Address: Technopark, India
City / Postcode: Thiruvananthapuram, 695581
Work Phone: 781-373-8455
Email: athira.sudhakar@genproindia.com
Web: www.genproindia.com

Author Name: Limna Salim
Company: Genpro Life Science India Private Ltd
Address: Technopark, India
City / Postcode: Thiruvananthapuram, 695581
Work Phone: 781-373-8455
Email: limna.salim@genproindia.com
Web: www.genproindia.com