BIORANKINGS®
TECHNICAL REPORT SERIES

Microbiome Research: Moving from Exploratory to Regulatory

If you plan on going to a regulatory agency like the FDA or USDA for your microbiome product, you will need to have biostatistics in order to meet their requirements. This paper shows how to move from exploratory R&D to formally designed experiments to test hypotheses about microbiome data.

How do cytokines change microbiome composition?

We were asked to find how cytokines change microbiome composition. The data included serum cytokine levels and stool 16s microbiome taxa counts on 128 subjects. We developed a new regression model that takes the microbiome composition as the dependent variable and the cytokines as the independent variables.

Five subgroups splitting on 4 different cytokines (**Figure 1**) were found. The top node of the tree (node #1) contains all N = 128 microbiome samples. These were split into two subgroups (nodes #2 and #13) based on if the subject's serum leptin is less than or greater than 1476. N = 57 subjects ended up in node 2, and N = 71 in node 13. The same process is applied independently to the N = 57 subjects in node 2 and the N = 71 subjects in node 13, and then to the subgroups from these splits, and repeated until there are few samples in the nodes. To avoid overfitting the large tree is pruned back to find the best tree.

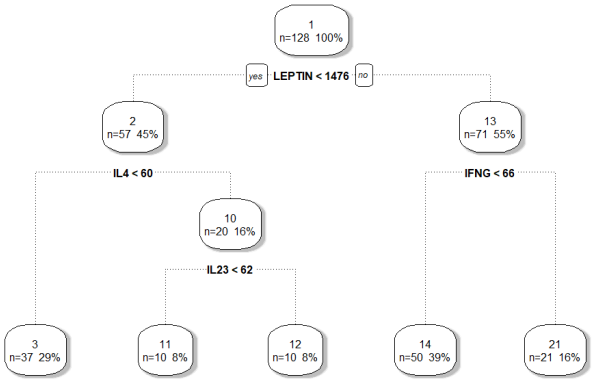

Figure 1

Each terminal node is defined by the splits leading from the top node. Node 3, for example, are the subjects with leptin < 1476 and IL4 < 60, and node 21 are the subjects with leptin > 1476 and IFNG > 66. The taxa distributions for the 5 terminal nodes (**Figure 2**) are different.

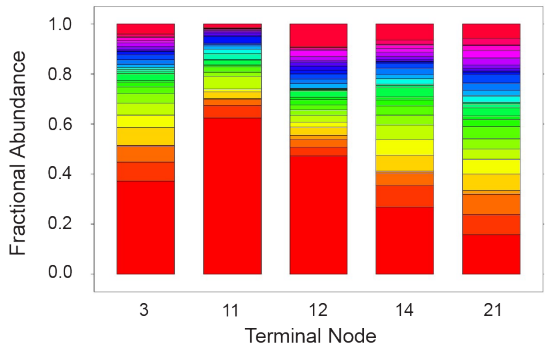

Figure 2

Suppose investigators decided nodes 3 and 21 were clinically interesting based on differences in cytokines and microbiome. It might be tempting to statistically test this pattern, but that would be wrong since the pattern was found in an exploratory big data analysis. What is needed to test for statistical significance is a follow-up validation study.

How do we do the validation study?

We want to test if the microbiome compositions in nodes 3 and 21 are different.

Assuming the same eligibility criteria as was used to collect data on the above 128 subjects, measure each subject's serum leptin IL4, and IFNG. These will be enough to find subjects that fall into nodes 3 and 21. These subjects will have their microbiome measured.

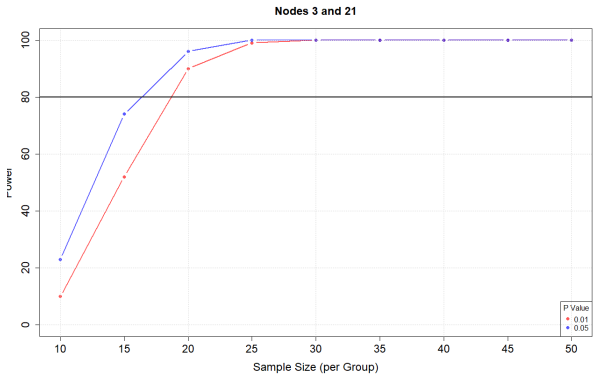

Figure 3

Using methods and software we developed previously, the sample sizes to test the null hypothesis that groups 3 and 21 are the same for P < 0.05 and P < 0.01 for 80% power are shown in **Figure 3**. Along the X axis are sample sizes per group and along the Y axis is power to detect a difference. The horizontal line at power = 80% is a typical level for experiments.

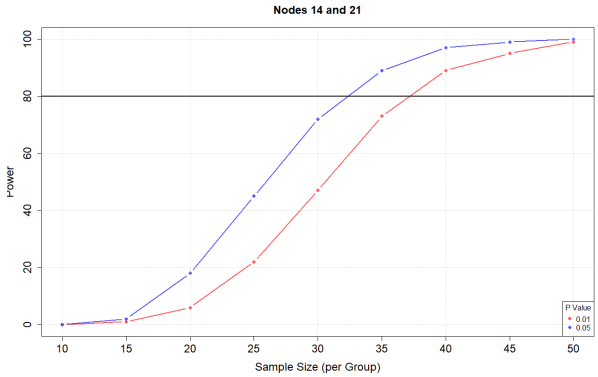

Figure 4

In this validation study, 17 subjects per group are needed for P < 0.05, and 19 subjects per group for P < 0.01. These numbers come from where the blue and red power curve lines cross the 80% power line.

For comparison, suppose investigators want to test nodes 14 and 21, which have more similar taxa distributions. The power curves indicate sample sizes per group of 33 and 38 for P < 0.05 and P < 0.01, respectively (**Figure 4**).

Microbiome patterns can be validated formally in follow-up studies

"Research with big data looks for patterns in datasets that do not originate from a planned sample or designed experiment." Once interesting patterns are found in exploratory research, they need to be confirmed in a hypothesis testing experiment.

As microbiome research matures and companies develop products, it is necessary for FDA and USDA approval processes that research moves from exploratory to hypothesis testing. This Case Study shows how to do this in a microbiome project. Since the statistical machinery is already developed for specifying testable hypotheses, determining the right sample size, and calculating a P value, this change in R&D focus is straight forward.

BioRankings' Technical Report Series

BioRankings' mission is to help biomedical researchers move their technology from the lab to clinical applications using statistically valid analytical tools for efficient study designs, correct data analyses and conclusions, and rigorous and objective decision making for designing follow-up studies and eventual FDA approval.

To help achieve its mission, BioRankings publishes a Technical Report series focused on applying various statistical methods to real data analyses. Written for understanding by scientists and administrators, these reports will provide an intuitive understanding of the analyses leaving the statistical details to other publications.

For more information, contact BioRankings at 314-704-8725 or bill@biorankings.com.

References

"The integrative human microbiome project." Nature, 569:641-648. 2019.

"Longitudinal multi-omics of host-microbe dynamics in prediabetes." Nature, 569:663-671. 2019

"New statistical method identifies cytokines that distinguish stool microbiomes." Nature Scientific Reports, 9:20082. 2019.

"Hypothesis testing and power calculations for taxonomic-based human microbiome data." PLoS ONE, 7(12). 2012.

"Leadership in statistics: Increasing our value and visibility." The American Statistician, 73(2), 109-116. 2018.