BIORANKINGS
TECHNICAL REPORT SERIES

Using Biostatistics to Analyze Microbiome Data

Next-generation sequencing and annotation pipelines produce data tables for microbiome experiments where the rows (or columns) are the samples, and columns (or rows) are the taxa names. The entries in these taxa count tables are the number of sequence reads for a sample (row) by taxon (column) representing the abundances or composition of the taxa in that sample.

Software developed by BioRankings based on formal biostatistical theory makes the analysis of microbiome data automatic, unbiased, reproducible, and interpretable. This Technical Report shows how we present microbiome data summarization, perform hypothesis testing comparing the microbiome across groups, calculate sample size and power for microbiome experiments, and estimate the population, or sample-to-sample, variability (i.e., population diversity).

Readers who would like to test this software on their data and compare the results generated with what they have found by their analyses are invited to do so free of charge. Please email us at contact@biorankings.com and we will arrange for this to happen.

SUMMARIZING MICROBIOME DATA

The Dirichlet-multinomial (DM) probability model applies to microbiome taxa count data. It has 2 parameters that summarize the data - pi (π) and theta (θ) – that define the taxa proportions and sample-to-sample overdispersion or variability (diversity). It is the foundation for a large class of statistical tools such as confidence intervals, hypothesis testing, power calculations, etc.

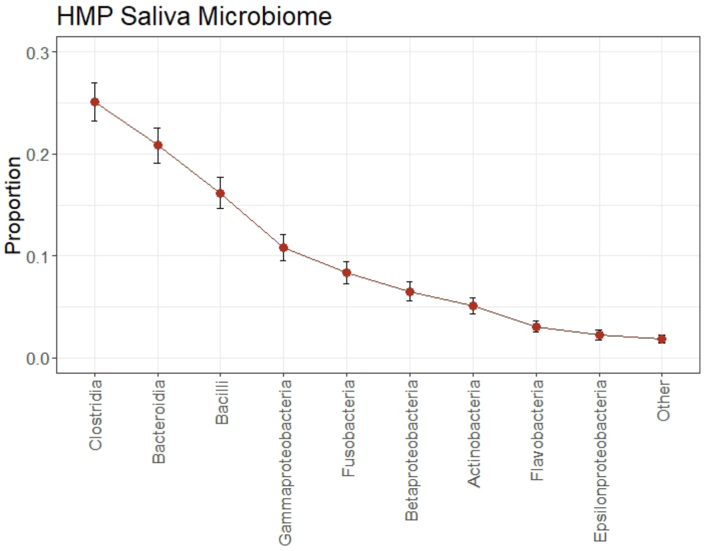

Figure 1

N = 300 saliva samples from the NIH Human Microbiome Project (HMP) study are summarized in Figure 1. The taxa proportions, pi, are displayed on the Y axis along with their 95% confidence intervals (CIs). The CIs are calculated from the theta parameter for all the taxa at the same time and are more accurate than fitting separate CIs to each taxon one at a time. The taxa are labeled on the X axis. From this plot we can see that Clostridia, Bacteroidia, and Bacilli are about 25%, 21% and 16% respectively of all taxa and the remaining taxa are < 15% each. The rare taxa were rolled-up in the analysis into the 'other' category. Theta = 0.037 is small suggesting all the saliva microbiome samples in the HMP were similar to each other.

HYPOTHESIS TESTING (P VALUES)

Deciding if the microbiome is the same or different across groups is done by hypothesis testing. By convention, if P <= 0.05 the null hypothesis (i.e., the microbiomes are the same) is rejected, and we declare the microbiome between groups is statistically different.

Two hypothesis tests are shown here. Figure 2 summarizes the data as above for the first hypothesis test comparing the HMP microbiomes of saliva and stool samples. The CIs do not overlap except for Clostridia and the rare taxa grouped in the 'Other' category, and with P < 0.001 we reject the null hypothesis that the microbiomes are the same. The effect size = 0.508, a measure of how far apart the microbiomes in the two populations are from each other, indicates that the difference between the two groups is large. The lack of overlap in the CIs shows that stool is dominated by Bacteroidia, while saliva is dominated by Bacilli, Gammaproteobacteria, Fusobacteria, Betaproteobacteria, Actinobacteria, Flavobacteria and Epsilonproteobacteria.

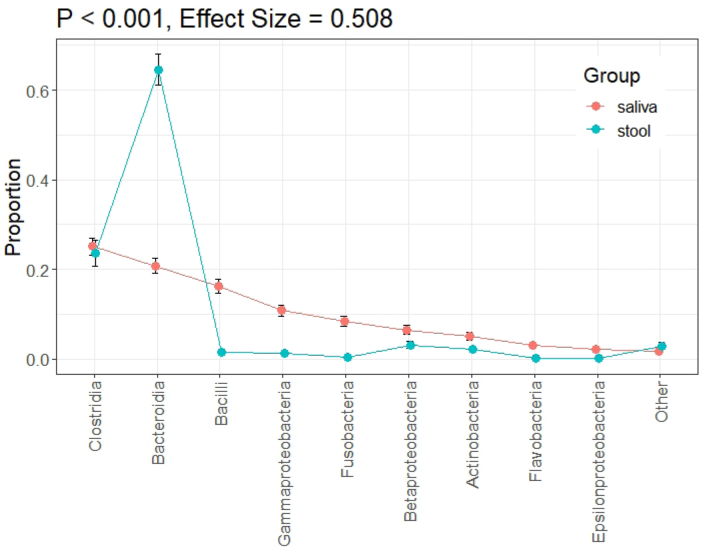

Figure 2

Figure 3 shows the results for the HMP mid vaginal and vaginal introitus microbiomes. The CIs show significant overlap for all the taxa and P = 0.839 indicating no evidence the microbiomes at these two body sites are different. The effect size = 0.102 indicates little difference in composition.

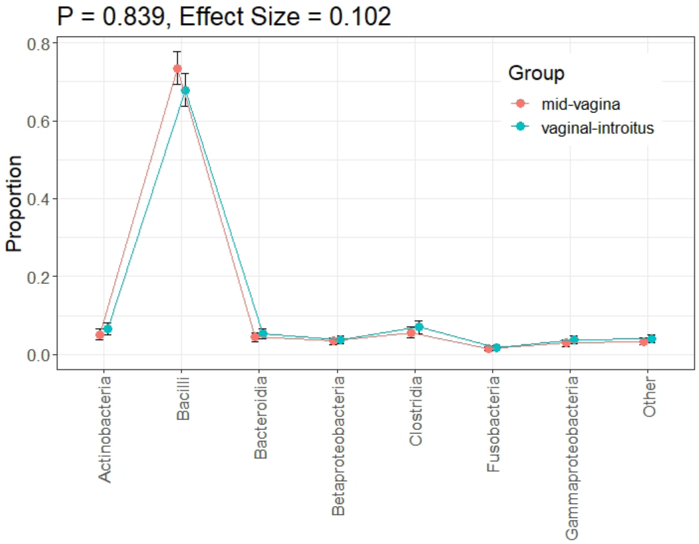

Figure 3

SAMPLE SIZE AND POWER CALCULATION

Data from the Integrative Human Microbiome Project (iHMP) looked at the impact of cytokines on the microbiome. DM-RPart regression analysis suggested high and low leptin levels were associated with changes in taxa composition. To confirm this result a study was designed where the microbiome would be collected and tested from subjects with low and high Leptin. The question needed to design an optimal study, was how many subjects in each group are needed?

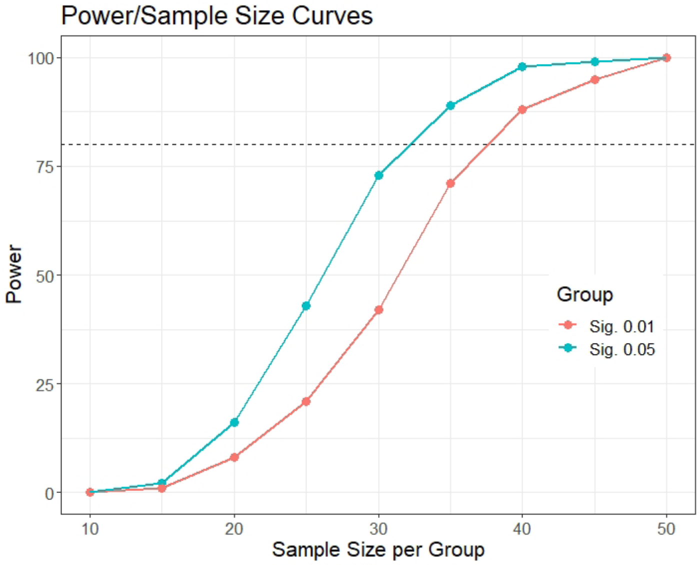

Figure 4

Figure 4 shows the power curves for P = 0.05 and 0.01 (significance levels). The horizontal line is at 80% power and indicates N = 32 and 37 samples per group would be needed for these significance levels, respectively.

THETA (OVERDISPERSION)

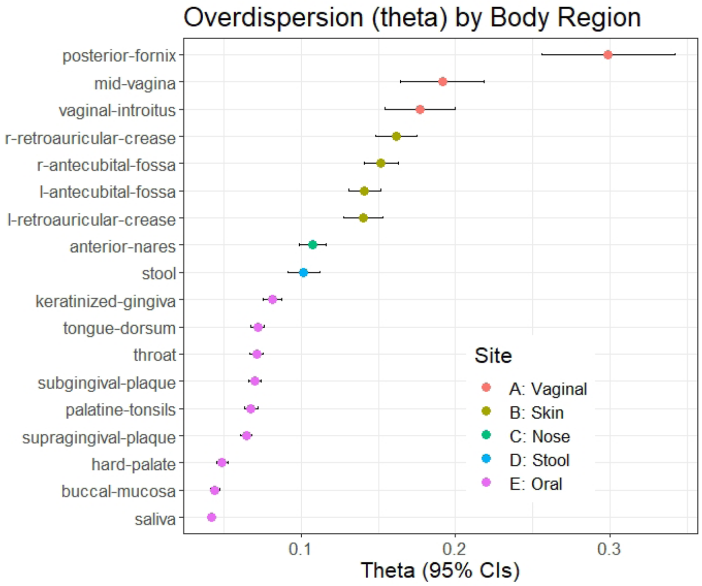

Figure 5

The theta parameter is a measure of how much the samples in a microbiome dataset are spread-out (the sample-to-sample variability). The larger theta is, the more differences in microbial composition between the samples.

Figure 5 shows theta with 95% CIs for each HMP body site at the class level. These are ordered top down from largest to smallest. The consistency in theta within similar body sites should be noted. The 3 most variable body sites in terms of differences in microbiome between subjects are the posterior fornix, mid, and introitus vaginal sites. In decreasing order, next are the 4 skin sites, followed by the nose and stool, with the 9 oral sites being the least variable.

DISCUSSION

Biostatistics presented here have advantages over current methods. Using the DM probability model versus non-parametric histograms and distance methods allows us to automate estimation with confidence intervals, hypothesis testing, power/sample size calculations, and measures of population diversity.

Summarizing microbiome data with pi plots and confidence intervals has the advantage of clearly showing variability which is often masked by other plots such as colored bar charts.

A single hypothesis test comparing all the taxa is generally viewed by statisticians as more powerful compared to testing each taxon separately which requires multiple testing adjustment and is known to result in more false positives. Single taxon testing also masks interactions which are known to occur in many genetic problems (e.g., epistasis).

The use of probability theory also allows researchers access to other tools. Here we presented sample size and population diversity as outcomes of this formal approach. Other methods now become easier to develop, such as regression analysis, building on the last one hundred years of statistical R&D.

REFERENCES

[1] https://pubmed.ncbi.nlm.nih.gov/23316946

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3527355/pdf/pone.0052078.pdf

[3] https://www.sciencedirect.com/science/article/pii/B9780124104723000063

[4] https://www.nature.com/articles/s41598-019-56397-9

For more information about our work in microbiome or any questions, contact Carlos Gonzalez at carlos@biorankings.com