# SIMPER Evaluation Metric

**Overview**

Similarity of Percentages (SIMPER) is a nonparametric multivariate analysis used to identify which entities (taxa) contribute the most to the variation between two sets of samples. The analysis utilizes the summarized taxonomic abundance information to calculate the percentage contribution of each taxon to the overall dissimilarity between the groups. The Bray-Curtis distance metric is used to estimate the distance (contribution) between samples. Comparisons are made between sample sets to compute the average percent contribution of each taxon to the overall dissimilarity measure (average taxa contribution/overall contribution). In cases where there are more than two groups to be compared the comparison is done in a pair-wise manner.

From literature, we understand that SIMPER results are hard to interpret as we cannot distinguish if the dissimilarity is due to within group or between group differences in abundance. In order to test whether the percentages generated by SIMPER are largely related to the within group abundance of taxa we used HMP Stool data (divided equally into two sets). To find the variance around the percentages generated by SIMPER we bootstrapped the samples. We observe that the analysis singles out one taxon which supports the finding in literature. We also computed the Spearman’s Rank Correlation Coefficient with 95% confidence interval between the ranked percentage contribution of each taxon and the ranked percentage abundance (composition) of each taxon and found high correlation.

**Usage**

./SIMPER.r

-i <summary table>

-f <factors>

-b <number of bootstraps>

[-o <output filename root>]

A tab delimited summary file and factors file are needed for running this script along with the required number of bootstraps.

**Example**

**Input**

./SIMPER.r –i v35.16sTaxa.TotFilt\_1000.Stool.summary\_table.xls \

-f v35.16sTaxa.TotFilt\_1000.Stool.factors.txt -b 200

**Output**

v35.16sTaxa.TotFilt\_1000.Stool.summary\_table.pdf v35.16sTaxa.TotFilt\_1000.Stool.summary\_table.confidence\_interval95.xls

The output pdf file contains three plots.

1. Box plotprovides the variation around the percentage contribution to dissimilarity generated by SIMPER for the top 20 taxa. This is produced as a result of **bootstrapping** the samples in the given summary file. Each box in the box plot represents the first quartile (lower horizontal line), median (middle bold line), and upper quartile (upper horizontal line) of the percent contribution generated by SIMPER for each taxon. The circles represent the outliers present in the bootstrapped data.
2. Bar plotrepresents the **observed** top 20 taxa based on their composition/abundance in the sample.
3. Bar plot represents the **observed** top 20 taxa contributing to the dissimilarity as reported by SIMPER

The bar plots are colored by taxa. The *Top 20 Taxa: Observed Composition* barplot is arranged in decreasing order of taxa composition. The *Top 20 Taxa: Observed Contribution to Dissimilarity* barplot follows the same taxa composition rank as the *Top 20 Taxa: Observed Composition* barplot but its heights represent the percentage contribution to dissimilarity generated by SIMPER. The bars on both barplots, *Top 20 Taxa: Observed Composition* and *Top 20 Taxa: Observed Contribution to Dissimilarity,* are labeled with the taxon rank corresponding to observed composition and observed contribution to dissimilarity respectively.

The output xls file contains the lower bound and upper bound 95% confidence intervals around the percentages generated by SIMPER for each taxon.