Research into the role that the microbiome plays in human health has helped identify microbes associated with diseases such as obesity and Crohn’s disease. However, current microbiome studies do not account for variations in human anatomy, such as pH changes along the gut, which leads to many spurious associations. Furthermore, these variations increase the overall variance in the microbial abundance profiles, making it harder to distinguish between diseased and healthy microbiota. To correct for these variations, we developed two new methods, “phyllum normalization” and “reduced principal component analysis”, and applied them to the largest pediatric Crohn’s disease dataset containing more than 1,000 diseased and control samples. Each method tackled a different aspect of environmental noise in the data and was used in permutation testing to identify possible disease-causing taxa and for sample diagnosis classification. Phylum normalization, in which relative abundances are obtained by normalizing with respect to phylum taxonomic level counts, reduced the variance between samples—allowing for better classification of disease and control. Reduced principal component analysis, in which a specified number of components causing the greatest data variance are removed, decreased the number of disease-associated taxa, selecting 11 taxa compared to 55 without the method, while preserving classification power. Our methods could have applications in disease diagnosis and in prioritizing follow up studies on potential pathogens. Future work would entail validating results with different datasets and other feature selection methods.

**Objective 1 Results**

The LTPNA performed better in predicting diseased samples using all taxa based on the AUC results from the logistic regression function. These results suggest that environmental factors were most likely accounted for.

**Dataset Normalization**

Phylum Normalization and Reduced Principal Component were used to normalize the data. Both datasets were log 10 transformed after abundances were obtained.

Log Transformed Phylum Normalization Abundances (LTPNA)

\[ Y_j = \frac{Y_j}{\text{log}(a_j + \text{pseudocount})} \]

Where:

\[ a_j = \text{sample phylum count total} \]

\[ Y_j = \text{log abundance} \]

\[ a = \text{abundance} \]

\[ l = \text{current taxon} \]

\[ j = \text{current sample} \]

\[ \text{count} = \text{count of taxon in the current sample pseudocount} = 2(\text{minimum sample count total}) \]

Log Transformed Relative Abundances Using Principal Components (LTRA PC)

- Find the eigenvectors(V) of the data matrix M using singular value decomposition.
- Multiply original matrix by transpose of V to give principal components.
- Set to 0 the specified number of components.
- Multiply again new matrix by the eigenvectors.

The specified number of principal components have been dropped.

**Objective 2 Results**

LTRA PC1 and PC2 selected the fewest significant taxa when permutation tests were performed. The 11 taxa produced similar results to the 27 taxa obtained using log transformed relative abundances suggesting a strong disease association.

**Future Work**

- Validate results with different datasets
- Compare using different feature selection methods

**References and Acknowledgments**

**Logistic Regression**

- A Logistic Regression function from a Python packet was used for determining prediction power and to obtain ROC curves.
- Training was performed in 354 samples and testing on 90 samples.
- Area under the ROC curve, which plots true positive versus false positive rates, was used to determine the predictive accuracy.

**Logistic Regression**

\[ \text{Logit} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots \]

\[ P = \frac{1}{1+e^{-\text{Logit}}} \]

**Logistic Regression**

- A Logistic Regression function from a Python packet was used for determining prediction power and to obtain ROC curves.
- Training was performed in 354 samples and testing on 90 samples.
- Area under the ROC curve, which plots true positive versus false positive rates, was used to determine the predictive accuracy.