Correcting for Environmental Factors in Microbiome Wide Association Studies

Michelle Patino Calero[1,2], Gabriel Birzu[3], Kirill Korolev, Ph.D.[3]
[1]University of Texas at El Paso, [2]Boston University Bioinformatics BRITE REU Summer 2018,
[3]Boston University Physics Department

Abstract

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 microbiotas. To correct for these variations, we developed two new methods, "phylum 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 tests 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.

Background

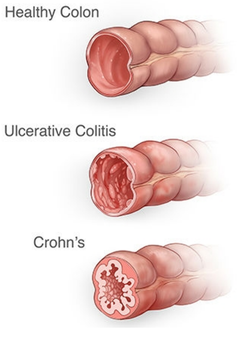


Figure 1. Diagram showing inflammation differences in inflammatory bowel diseases. [2]

Inflammatory bowel diseases (IBD) are characterized by inflammation of the gastrointestial tract and is an umbrella term for Crohn's disease and ulcerative colitis. It currently affects around 1.6 million Americans and there are 70,000 new diagnoses each year. [1]

- Crohn's disease can extend through the entire bowel wall thickness and mainly affects the connection of the small intestine's end and colon's beginning (the ileum). [1]

The **microbiome** is a collection of microorganisms that inhabit a particular environment.

- Current studies show differences in microbiota composition between healthy and diseased samples and that alterations in the interaction between human gut microbiota and the mucosal immune system may have an important role in IBD. [3]

Objectives and Pipeline

The possible difference between healthy and diseased microbiota composition leads us to the following objectives of this project:

1. Reduce environmental noise.
2. Reduce the number of taxa considered "significant".

Data preparation, dataset normalization, permutation tests, significant taxa selection, and receiver operating characteristic curve graph creation are all processes in the pipeline. Our pipeline begins in data preparation where certain taxa and samples are removed if they had missing information or were not present a specified amount.

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)

$$l_{i,j} = log_{10}(a_{i,j} + pseudocount)$$

Where:

$$a_{i,j} = \frac{count}{sample\ phylum\ count\ total}$$

l = log abundance
a = abundance
i = current taxon
j = current sample
count = count of taxon in the current sample
pseudocount = 1/(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.

Permutation Test

Permuation test is used to obtain "significant" taxa from each dataset.

Null Hypothesis: The abundance of taxa in control and diseased is the same (no difference).

$$Taxon_i = [x_1, x_2, \ldots, x_L, x_{L+1}, x_{L+2}, \ldots, x_M]$$

L = # of control
M = # of samples

L control | M-L disease

$Taxon_i$: the current taxa

x_n: the abundance of the taxon in sample n

<x> : the mean of the group

$$\Delta = < x >_{control} - < x >_{disease}$$

$$\Delta_p = < x >_{control} - < x >_{disease}$$

Permutation

$$Taxon_i = [x_3, x_L, \ldots, x_{L+2}, x_{L+4}, x_5, \ldots, x_1]$$

L "control" | M-L "disease"

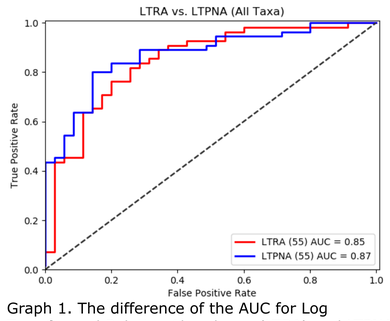
$$p = \frac{|\Delta_p| > |\Delta|}{number\ of\ permutations}$$

Logistic Regression

* A Logistic Regression function from a Python packet was used for determine 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 predictory accuracy.

Objective 1 Results

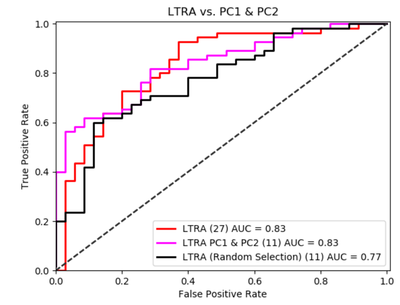
The LTPNA performed better in predicting diseased samples using all taxa based on the AUC results from the logistic regression function. These results suggest that environmental factors were most likely accounted for.



Graph 1. The difference of the AUC for Log Transformed Relative Abundance (LTRA) and LTPNA with all.

Objective 2 Results

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



Graph 2. The difference of the AUC for LTRA and LTRA PC1 and PC2 with respective significant taxa. A random selection of taxa chosen from LTRA significant taxa is obtained for predictive accuracy.

Future Work

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

References and Acknowledgments

References:
[1] Crohn's & Colitis Foundation.(2014). The Facts About Inflammatory Bowel Diseases. Retrieved from: http://www.crohnscolitisfoundation.org/assets/pdfs/updatedibdfactbook.pdf
[2] Children's Hospital of Philadelphia.(n.d) Inflammatory Bowel Diseases. Retrieved from: https://www.chop.edu/conditions-diseases/inflammatory-bowel-disease
[3] Kostic, A. D., Xavier, R. J., & Gevers, D. (2014). The microbiome in inflammatory bowel disease: current status and the future ahead. Gastroenterology, 146(6), 1489-1499.
[4] Gevers, D., Kugathasan, S., Denson, L. A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., ... & Morgan, X. C. (2014). The treatment-naive microbiome in new-onset Crohn's disease. Cell host & microbe, 15(3), 382-392.
[5] Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European journal of epidemiology, 31(4), 337-350.
[6] Wang, F., Kaplan, J. L., Gold, B. D., Bhasin, M. K., Ward, N. L., Kellermayer, R., ... & Dogan, H. (2016). Detecting microbial dysbiosis associated with pediatric Crohn disease despite the high variability of the gut microbiota. Cell reports, 14(4), 945-955.

Research reported in this poster was supported by the National Institute of General Medical Sciences of the National Institutes of Health under linked Award Numbers RL5GM118969, TL4GM118971, and UL1GM118970. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.
This work was funded, in part, by NSF grant DBI-1559829, awarded to the Boston University Bioinformatics BRITE REU program.