# Predicting Inflammatory Bowel Disease from Human Microbiome Data

CS109A Project Statement
Project group 54: Ilia Gelfat, Mei (May) Xue, Tobie Lee, Cheuk (Alan) Leung

## Problem Statement

The human microbiome is the collection of microbes that inhabit our bodies, from the surface of our skin to the depths of our digestive tracts. The microbiome has been implicated in a wide variety of diseases, ranging from obesity to neurological disorders. However, despite its relevance to human health, studying the microbiome has up until recently been an insurmountable challenge, due to its immense complexity. In recent decades, the combined power of Next-Generation DNA sequencing technology and advanced computational methods has finally allowed researchers to begin to effectively tackle this problem [1].

To help advance microbiome research, the National Institutes of Health (NIH) established the Human Microbiome Project (HMP) in 2008 [2]. The first phase of the project - HMP1 - aimed to identify and characterize the microbial strains that comprise the human microbiome, allowing researchers to get a sense of what a healthy microbiome consists of. In 2014, the NIH launched the second phase - the Integrative Human Microbiome Project (iHMP) [3]. Compared to HMP1, the iHMP is much broader in scope, focusing on three types of microbiome-associated medical conditions: preterm birth, inflammatory bowel disease and type 2 diabetes. Both the HMP and iHMP data have been made publicly available to allow anyone to try and contribute to understanding the human microbiome.

In this project we will explore how tractable it is to use the gut microbiota profiles as a form of biomarker for diagnosing inflammatory bowel disease (IBD), which includes Crohn's Disease (CD) and ulcerative colitis (UC), using samples from the Inflammatory Bowel Disease Multi'omics Database (IBDMDB) - one of three major studies which comprise the iHMP.

IBD is a disease characterized by inflammation of the gastrointestinal tract (CD) and the colonic mucosa (UC) and affects more than 1.5 million people in the USA [4, 5, 6, 7]. While the underlying cause of IBD remains unknown, genetic and environmental factors have been shown to contribute to its onset, and recent studies have shown that changes in gut microbiota composition or dysbiosis may be associated with the initiation and perpetuation of IBD [6, 7, 8, 9, 10]. The gut microbiota are known to interact with the host in a variety of ways and can have profound effects on the host’s metabolism and immune response [1, 6, 9].

Given the diversity and dynamic nature of gut microbiota within and between individuals (particularly those with IBD), which can vary with diet, exercise, age, sex, race, and other environmental factors such as medicine or drug usage, it may prove challenging to draw meaningful interpretations from the profiles of gut microbiota alone. Therefore, we expect to restrict and stratify our sample dataset accordingly to minimize the effect of these confounding factors as we build our models.

## Data Resources

1. The Inflammatory Bowel Disease Multi'omics Database: https://ibdmdb.org. Taxonomic data within merged csv file in products of HMP2 (study), 16S (data type), 2018.06 (week): https://ibdmdb.org/tunnel/public/HMP2/16S/1806/products.

## High Level Project Goals

1. Explore the available types of data within the IBDMDB. Download taxonomic and corresponding metadata data from site.
2. Perform exploratory data analysis (EDA) and generate a few hypotheses and questions.
3. Build a statistical model to test the proposed hypotheses. Which features of the microbiome or of the patient are most useful in predicting the disease outcome/diagnosis? What is the relative contributions to the model made by the microbiome compared to other patient factors?
4. Discuss the results and their potential implications on our understanding of the microbiome. Propose experiments that would allow to further study and test the predictions of the model.

## Data Collection and Processing

Initially, the HMP Data Portal was used to download the relevant data. However, upon closer inspection, much of the downloaded files were missing due to incomplete database access. Furthermore, patient metadata files were far less comprehensive than initially described in the HMP documentation. After numerous troubleshooting attempts, as well as contacting HMP support and course staff, the abovementioned IBD database was found.

From this database, we were able to obtain abundance matrices for fecal samples from 178 patients. For each sample, these matrices contained a list of operational taxonomic units (OTUs), each corresponding to a specific microbial strain. Associated with each OTU was a number proportional to the abundance of the corresponding strain within the sample. We proceeded to parse the OTUs into their taxonomic classification (kingdom, phylum, class, etc.), aggregate by phylum, and normalize the signal such that for each patient the abundance values of the 26 present phyla added up to 1. Although this explicitly builds colinearity into the data, we chose to represent the data in this was as it is much more interpretable. The choice to aggregate by phylum was driven by the available data, as not all OTUs contained information about class, order, family or genus.

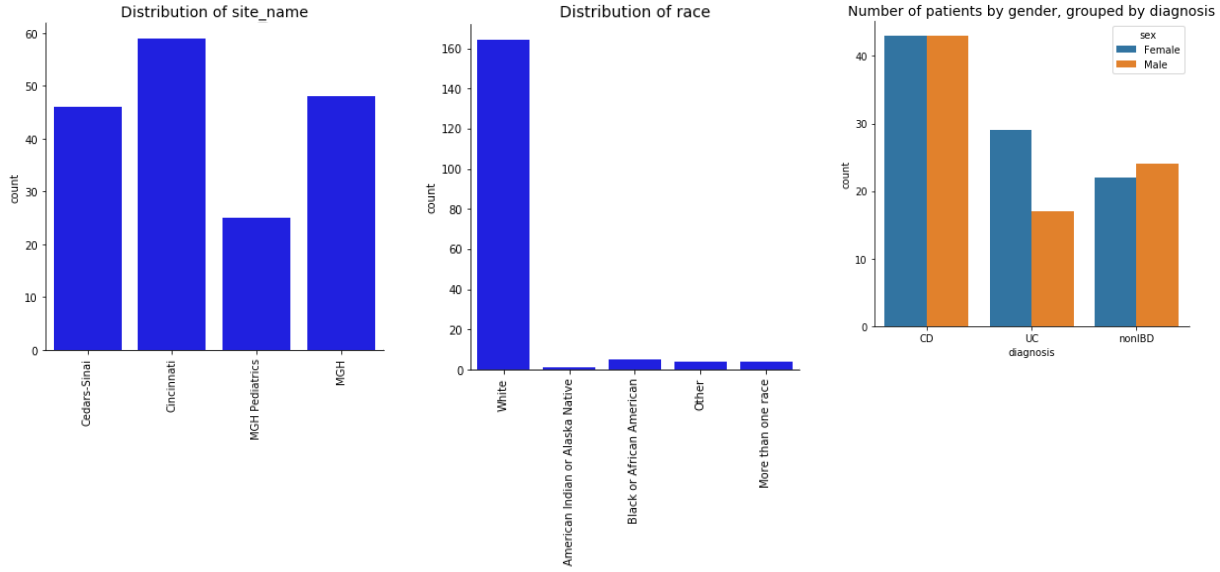
Next, bacterial abundance data was merged with patient metadata. The metadata contained demographic information such as age, gender, race and education level, as well as answers to medical and lifestyle questions. Most importantly, the metadata included our response variable - a diagnosis for each patient: “CD”, “UC” or “nonIBD”.

Closer inspection of the data revealed that most predictors had all missing values (approximately 300/500 predictors). As a first pass, all predictors with more than 10 missing values were dropped. Many of the remaining predictors were subsequently dropped as well, since they contained no information - all patients had the same predictor value (e.g. all patients

replied “no” when asked if they were on antibiotics). For education level, missing values were assigned to the existing category of “Unknown/Not Reported”. For age and biopsy location, missing values were imputed using the median and mode, respectively.

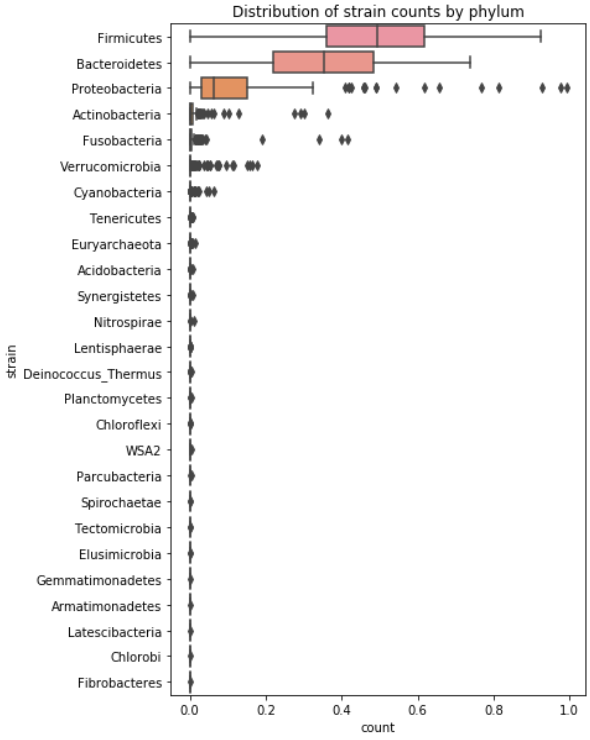
## Summary Statistics

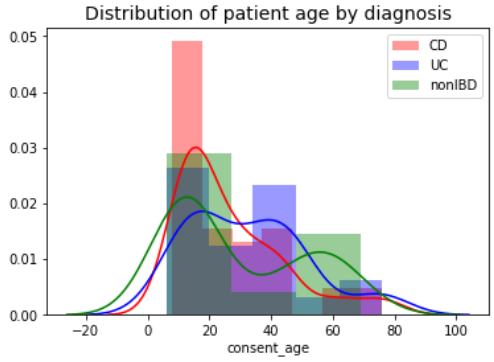
The variables from both the metadata and microbiome abundance were examined. First, we looked at the distribution of the diagnosis. 86 patients were diagnosed with CD, 46 with UC, and 46 with nonIBD. While CD has more patients, the three classes were well represented. We then looked at individual variables sex, education, occupation, biopsy location, age, race, and site of sample collection. Sex, education, age, and collection site were more evenly distributed, while occupation, biopsy location, and race were skewed, with only one or two dominant classes. Representative distributions for site and race are shown below.

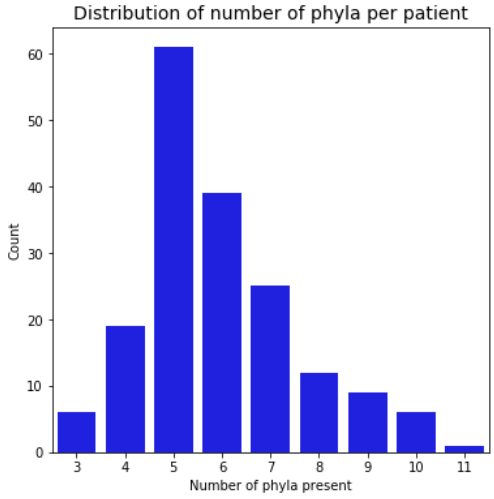



Furthermore, the distributions of sex and age were plotted by diagnosis. For UC, there is a higher proportion of females compared to males, and there are differences in the distributions of age by diagnosis.

In addition, the distributions of each phylum were plotted as boxplots. We see that the data are dominated by three phyla: Firmicutes, Bacteroidetes and Proteobacteria. However, despite the low abundance of other phyla, microbial diversity may turn out to be a relevant factor in predicting disease. Therefore, we also examined the number of phyla observed per patient, which shows a wide distribution of the number of strains. The number of phyla was added as a predictor for modeling.



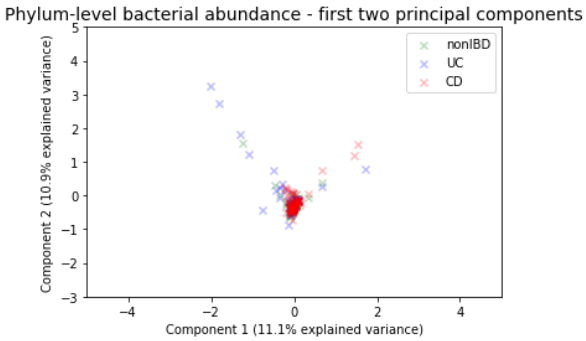




## Exploratory Data Analysis

As a baseline model, a multinomial logistic regression model was used. Categorical variables were converted to binary using one-hot encoding, and the data was split into test and training sets (20:80). Both LASSO and Ridge regularization was used, with 5-fold cross-validation over a range of 21 C-values. Both models yielded a training accuracy of 81-83%, and a test accuracy of 64%. This suggests that overfitting occurred despite applying regularization, suggesting that a simple logistic regression model is not sufficient to capture the complex decision boundaries of required to classify this dataset.

To explore the bacterial abundance data, principal component analysis (PCA) was applied to the 26 phylum-level abundance predictors. The predictors were standardized and plotted as a function of the first two principal components. The data did not appear to cluster by diagnosis, as may be expected from a complex system such as the gut microbiome. The lack of a low-dimensional structure in the data was also supported by the explained variance ratio: the first two principal components explained only 11.1% and 10.9% of the variance, with 16/26 components required to explain 90% of the variance.



# References


1. Cho, Ilseung, and Martin J. Blaser. "The human microbiome: at the interface of health and disease." *Nature Reviews Genetics* 13.4 (2012): 260.
2. “About the Human Microbiome.” *NIH Human Microbiome Project - About the Human Microbiome*, https://hmpdacc.org/hmp/overview/. Retrieved 2019-10-03.
3. “About the Human Microbiome.” *NIH Integrative Human Microbiome Project - About the Human Microbiome*, https://hmpdacc.org/ihmp/overview/. Retrieved 2019-10-03.
4. Chu H, Khosravi A, Kusumawardhani IP, et al. Gene-microbiota interactions contribute to the pathogenesis of inflammatory bowel disease. *Science*. 2016;352(6289):1116–1120. doi:10.1126/science.aad9948
5. The Inflammatory Bowel Disease Multi'omics Database, https://ibdmdb.org. Retrieved 2019-11-13
6. Belkaid Y, Hand TW. Role of the microbiota in immunity and inflammation. *Cell*. 2014;157(1):121–141. doi:10.1016/j.cell.2014.03.011
7. Lloyd-Price, J., Arze, C., Ananthakrishnan, A.N. et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 569, 655–662 (2019) doi:10.1038/s41586-019-1237-9
8. Matsuoka K, Kanai T. The gut microbiota and inflammatory bowel disease. *Semin Immunopathol*. 2015 Jan;37(1):47-55. doi: 10.1007/s00281-014-0454-4. Epub 2014 Nov 25.
9. Wu GD, Lewis JD. Analysis of the human gut microbiome and association with disease. Clin Gastroenterol Hepatol. 2013;11(7):774–777. doi:10.1016/j.cgh.2013.03.038
10. Morgan XC, Tickle TL, Sokol H, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol*. 2012;13(9):R79. Published 2012 Apr 16. doi:10.1186/gb-2012-13-9-r79
11. Mandal RS, Saha S, Das S. Metagenomic surveys of gut microbiota. *Genomics Proteomics Bioinformatics*. 2015;13(3):148–158. doi:10.1016/j.gpb.2015.02.005
12. Shreiner AB, Kao JY, Young VB. The gut microbiome in health and in disease. *Curr Opin Gastroenterol*. 2015;31(1):69–75. doi:10.1097/MOG.0000000000000139
13. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214 (2012)
14. H. M. P. R. N. C. Integrative, The integrative human microbiome project. *Nature* 569, 641–648 (2019).
15. Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci U S A*. 2007;104(34):13780–13785. doi:10.1073/pnas.0706625104