INTO THE GUT MICROVERSE

•••

SHOURYA - 2111118
NISSY MILCIA W - 2111078

RECAP

Microbiome is the collection of microbes occupying a habitat and metabolome is the collection of metabolites in a biological space. Changes in the gut microbiome and the metabolome have been associated with human health and diseases. Characterisation of the human faecal microbiome and the associated metabolome could provide the opportunity to develop diagnostic approaches and personalized medicine for human diseases such as IBD, CRC, etc. We intend to study the gut microbiome and metabolome composition across various “healthy” and “disease inflicted” samples and find any patterns regarding the variation, diversity and abundance of the bacteria and their metabolites. This correlation could be used to predict if an individual is at risk for specific diseases.

REFERENCES FOR LITERATURE REVIEW

Article 1: Franzosa EA, Sirota-Madi A, Avila-Pacheco J, et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease [published correction appears in Nat Microbiol. 2019 May;4(5):898]. *Nat Microbiol*. 2019;4(2):293-305. doi:10.1038/s41564-018-0306-4

Article 2: Sinha R, Ahn J, Sampson JN, Shi J, Yu G, et al. (2016) Fecal Microbiota, Fecal Metabolome, and Colorectal Cancer Interrelations. PLOS ONE 11(3): e0152126. https://doi.org/10.1371/journal.pone.0152126

ARTICLE 1

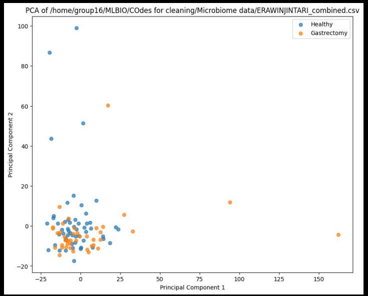
Step	Description
1. Data Preparation	- Gather data on gut microbiota (e.g., species abundance) and metabolite profiles for each participant.
2. Diversity Calculations	- Calculate the Shannon Index for each sample to quantify within-sample diversity (richness and evenness) of microbes or metabolites.
3. Dissimilarity Matrix Creation	- Choose a distance metric like Bray-Curtis distance. - Calculate the distance between every pair of samples based on their microbial/metabolite profiles. This creates a matrix representing overall dissimilarity.
4. CMDS Analysis	- Use CMDS on the distance matrix to project high-dimensional data into a lower-dimensional space (usually 2D or 3D). - Visualize where samples with similar profiles are closer together and distinct profiles are further apart.
6. Visualization and Interpretation	- Plot the CMDS results with each point representing a participant. - Analyze potential clusters or groupings based on gut health profiles. Consider disease status and explore potential differences in positioning.
7. Statistical Tests	- Use statistical tests like PERMANOVA to assess the significance of observed separation between groups in the CMDS plot.
8 Unsupervised Clustering of Metabolites	Group metabolite features with similar abundance patterns across samples (independent of disease status). Analyze co-occurring metabolites for shared origin or function.

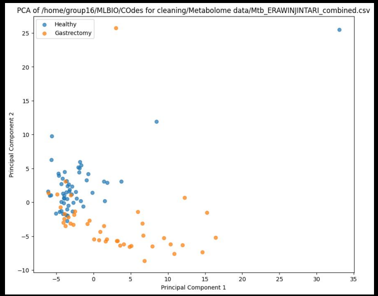
ARTICLE 2

Method	Purpose	Details	Output
Unconditional Logistic Regression	Assess microbe-CRC association	- Binary dependent variable (CRC diagnosis). - Binary independent variable (microbe presence/absence).	Odds Ratio (OR): - OR > 1: Increased odds of CRC with microbe. - OR = 1: No association between microbe and CRC. - OR < 1: Potential protective effect of microbe against CRC.
Stepwise Logistic Regression	Refine microbe-CRC association considering metabolites	- Starts with significant microbes from unconditional analysis. - Adds metabolites one-by-one. - Evaluates if a metabolite changes the OR of a microbe-CRC association.	Identifies metabolites influencing the microbe-CRC relationship.
Correlation Analysis	Explore microbe-metabolite relationships	- Calculates correlation coefficient (r) between variables (e.g., microbe & metabolite).	- r > 0: Positive correlation (variables tend to change together). - r < 0: Negative correlation (variables change oppositely). - r close to 0: Weak or no relationship.
Linear Regression	Explore general microbe-metabolite associations	- Continuous dependent variable (metabolite level). - One or more independent variables (microbe abundance, age, sex, BMI). - Considers factors affecting microbe-metabolite relationships.	Identifies significant associations between metabolites and other variables.
Principal Components Analysis (PCA) & Principal Coordinates Analysis (PCoA)	Analyze complex microbe/metabolite data	- Dimensionality reduction techniques. - Create new composite variables capturing most of the data variation. - Compare correlations between these components in CRC cases vs. controls.	Understands if overall microbe-metabolite relationship changes in presence of CRC.

DATA PREPROCESSING, EXPERIMENTS, TESTS AND RESULTS

- The dataset obtained from https://github.com/borenstein-lab/microbiome-metabolome-curated-data.git consists of patient metadata, the abundance of bacterial genera as well as metabolites and mapping of the metabolites to standard databases such as the Human Metabolome Database (HMDB) across multiple studies.
- Based on our idea, decided to proceed with seven studies that dealt with IBD and gut-related cancers.
- Binning of the age feature. Selected metadata (Study dataset, sample, subject group, and age group).
- Created tables for each dataset that contained the selected metadata and microbiome/metabolome.
- Currently creating new features, such as diversity indices for each sample and mapping the metabolites to HMDB.
- Ran some preliminary tests such as PCA and a basic form of KNN (average accuracy = 0.54). The results are given below. Performed tSNE but the results are yet to be interpreted.





CURRENT ISSUES

- High variation across the datasets: Since the data was obtained from different studies, there is variation regarding certain features, such as genera and metabolite data, in terms of the number of features in the dataset as well as the features themselves.
- Feature reduction: as a consequence of the above issue, the feature reduction process has to be streamlined for each dataset.
- Combining datasets: we are undecided about whether to combine the different datasets and, if yes, how or to proceed with the uncombined data.

FURTHER WORK

- Plan to combine the data, compare accuracy with the average uncombined data, and proceed from there. In the case of individual datasets, decide how to train the model/s.
- Reduce features by performing statistical analyses to check for correlation within the features and between the individual features and the labels.
- Select models and fine-tune them to the data. Models in mind: Random Forest, Logistic regression, XGboost etc.
- Plan to attempt Ensemble Learning.