# An unsupervised approach to find connections between gut microbiome data and other omics-data describing a common sample set

## Introduction

The intricate relationship between hosts and their gut microbiome has always been of paramount interest to researchers. The gut-microbiome, teeming with diverse bacterial species, plays a pivotal role in human health and disease. In recent years, attempts to comprehend these interactions have adopted various analytical frameworks. Wu et al. [Lia+18], for instance, introduced the ecological concept of a ”guild” to the realm of gut microbiome analysis. Within this framework, members of a guild display co-abundance patterns, fluctuating synchronously, irrespective of their taxonomic affiliations, in response to resource availability. Simply put, guild members tend to ”vary” together when faced with specific conditions.

Bridging the gap between microbiome patterns and host responses, Priya et al. [Pri+22] combined Sparse CCA and LASSO to uncover links between host genes and microbial profiles. Their study delved into paired host transcriptomic and gut microbiome samples from colonic mucosal tissues of patients, elucidating intricate connections across a spectrum of conditions including colorectal cancer, inflammatory bowel disease, and irritable bowel syndrome.

Moreover, the push for a more integrative, multivariate approach to data analysis has been evident in recent literature. Liang et al. [Lia+18] introduced a novel variable reduction technique, the hierarchical clustering method (HCM), specifically designed for the joint analysis of multiple phenotypes. This method supports the belief that analyzing multiple phenotypes together can unearth insights that might be elusive to a singular, univariate analysis.

Inspired by these efforts, we present a novel approach in this study. Our method not only seeks connections between gut microbiome data and other multi-omics datasets, upholds the ecological essence of guilds, and harness the power of hierarchical clustering as a dimension reduction tool.

## A Referesher on Ward’s Method in Hierarchical Clustering Algorithm

Given two clusters $U$ and $V$ with $n_U$ and $n_V$ elements respectively, the Ward’s linkage criterion for merging is:

$$d(U, V) = \frac{n_U n_V}{n_U + n_V} \times \|\mu_U - \mu_V\|^2$$

Where $\mu_U$ and $\mu_V$ are the centroids of clusters $U$ and $V$ respectively, and $\|\mu_U - \mu_V\|^2$ is the squared Euclidean distance between the centroids.

## User-Driven Hierarchical Clustering for Gut Microbiome Data

Let $M$ be a matrix where each row $r_i$ represents a patient or sample, and each column $c_j$ represents a specific microbe. The element $m_{ij}$ denotes the abundance (or presence) of microbe $j$ in sample $i$.

1. For each pair of microbe columns $c_j$ and $c_k$ in $M$, compute the Euclidean distance to obtain a microbe-microbe distance matrix.
2. Using Ward's method, perform hierarchical clustering on the microbes based on the computed distance matrix.
3. Present a dendrogram (or other appropriate visualization) of the microbe clustering to the user.
4. Allow the user to choose a cut-off distance, $d_{\text{cut-off}}$, which will determine the number of "main" microbial clusters.

## Creating "Main" Clusters by Computing Algorithmic Means for Defined Microbial Clusters

Given main microbial clusters $M_1, M_2, M_3, \dots$ (denoted as $M_n$) previously defined by the user, for each cluster $M_n$:

1. Extract all measures within the cluster for the each sample $r_i$.
2. Compute the algorithmic mean across these measures to obtain a single value $\mu_{ni}$.
3. The result will be a vector of means $m_n = [\mu_{1n}, \mu_{2n}, \dots, \mu_{in}]$ representing or corresponding to cluster $M_n$ itself.

## Creating "Main" Clusters for Other Multi-omics Datasets

Suppose we also have other type of omics- datasets describing the same sample set $r_i$. Apply the processes above, we have:

- Metabolomics main clusters: $E_1, E_2, E_3, \dots, E_k$ as well as their corresponding vector of means $e_1, e_2, e_3, \dots, e_n$
- Transcriptomics main clusters: $T_1, T_2, T_3, \dots, T_m$ as well as their corresponding vector of means $t_1, t_2, t_3, \dots, t_m$
- Proteomics main clusters: $P_1, P_2, P_3, \dots, P_l$ as well as their corresponding vector of means $p_1, p_2, p_3, \dots, p_l$

This method enables a summarized representation of various omics datasets, where omics-measures, under the same main clusters, are represented by a corresponding vector of means.

## Spearman Correlation Computation

Given various vectors of means representing main clusters from omics datasets that were previously user-defined:

1. Rank the values within each vector.
2. Compute the Spearman correlation between:
   - (a) Vectors of means within the same omics data type (e.g., between $m_1$ and $m_2$, between $t_1$ and $t_2$, etc.).
   - (b) Vectors of means from different omics data types (e.g., between $m_1$ and $t_1$, between $m_2$ and $p_1$, etc.).

This approach provides insights into the relationships between different clusters / vectors of means across these datasets.

## Network Construction based on the Spearman correlation coefficient

1. Set a threshold $\tau$ for the Spearman correlation coefficient.
2. Construct a network where:
   - (a) Each node represents a main cluster (e.g., $M1, T1, P1$, etc.).
   - (b) An edge between two nodes exists if the absolute value of their Spearman correlation calculated from their corresponding vectors of means surpasses $\tau$.

Through this approach, users obtain a network visualization that showcases significant relationships between different omics clusters based on their correlations.