# Balance selection in microbiome studies

**J. Rivera - Pinto[1,2] , Vera Pawlowsky-Glahn[3] , Juan José Egozcue[4] , Roger Paredes[1,2,5,6] , M. Noguera - Julian[1,2,5] and M. Luz Calle[2]**

[1] Institut de recerca de la SIDA IrsiCaixa, Badalona (Spain)
[2] Universitat de Vic - UCC, Vic (Spain)
[3] Universitat de Girona, Girona (Spain)
[4] Universitat Politècnica de Catalunya, Barcelona (Spain)
[5] Universitat Autònoma de Barcelona, Barcelona (Spain)
[6] HIV Unit & Lluita Contra la SIDA Foundation, Badalona (Spain)

jrivera@irsicaixa.es; vera.pawlowsky@udg.edu; juan.jose.egozcue@upc.edu; rparedes@irsicaixa.es; mnoguera@irsicaixa.es; malu.calle@uvic.cat

**Abstract**

The analysis of the microbiome has recently become the topic of research for different diseases due to the availability of high-throughput sequencing technologies. Microbiome data is typically a matrix of counts corresponding to the number of DNA sequence reads assigned to each feature, which can be either taxonomic units (e.g. species) or functional units (e.g. genes). Microbiome data is intrinsically compositional since it is derived from count measures on microbial population samples. Thus, treating microbiome data through absolute counts or relative abundances, through proportions, may provide inconsistent or invalid results. Conversely, due to its compositional nature, information contained within ratios of microbiome components is methodologically more suitable and contains consistent information.

One of the most important goals in microbiome studies is the identification of differentially abundant features across different conditions or the detection of those associated to a specific continuous characteristic of interest. For instance, detection of microbial species that are differentially abundant across healthy and diseased individuals or detection of microbial species whose abundance is associated to inflammatory markers, is of major interest.

Acknowledging the compositional nature of microbiome data, we propose to search for groups of features whose balances are differentially abundant or associated to a variable of interest as an alternative to the usual analysis of differentially abundant individual features.

In this work we propose an algorithm for finding the groups of features (e.g. microbial species) whose balance is most associated with a dependent continuous or categorical variable. We take as objective function the explained variation of a linear or logistic regression model.

The iterative algorithm starts searching for the pair of variables with the highest correlated balance with the response variable and continues adding individual variables into the balance until there is no additional improvement in the explained variation when adding any other remaining variable. To evaluate the robustness of the selected balance of variables, a cross-validation procedure is performed.

We illustrate the methodology with data from the Barcelona HIV metagenome project (MetaHIV) that analyses how the gut microbiome influences the ability of HIV-1 infected individuals to achieve adequate immune reconstitution, control HIV-1 replication and limit chronic inflammation. The MetaHIV project is conducted at the Institute for AIDS research IrsiCaixa in Spain.