On the Methodological difficulties of analyzing the association between compositional microbiome data and outcomes: An example from an HIV study.

Antonin Colajanni*[1,2], Raluca Uricaru†[3], Patricia Thébault‡[3], and Rodolphe Thiébaut[4]

[1]Bordeaux population health – Université de Bordeaux, L'Institut National de Recherche en Informatique et e n Automatique (INRIA) – France

[2]Laboratoire Bordelais de Recherche en Informatique – Université de Bordeaux (Bordeaux, France) – France

[3]Laboratoire Bordelais de Recherche en Informatique – Université de Bordeaux – France

[4]Bordeaux population health – Université de Bordeaux, Institut de Santé Publique, d'Épidémiologie et de Développement (ISPED), Institut National de la Santé et de la Recherche Médicale, L'Institut National de Recherche en Informatique et e n Automatique (INRIA) – France

Abstract

Microbial translocation occurs when bacterial products access the blood stream from the gut due to gut barrier dysfunction, potentially triggering persistent immune activation and affecting treatment efficacy. Translocation can be studied using RNA-sequencing techniques to analyze the "meta-transcriptome": all the RNA from bacteria, viruses and fungi from whole blood. Building on previous work to repurpose existing human sequencing datasets to extract microbiota-related information, the analyses were conducted on the plasma of patients with a treated HIV infection.
A key challenge with these new data is to determine whether their metatranscriptome accounts for part of the variability observed in the patients' response to the HIV treatment. Specifically, to identify meaningful associations between microbiome-derived explanatory variables and our outcome variable, the CD4+ cell count.
However, analyzing microbiome data presents several challenges due to its compositional nature, the sparsity of the data matrix, and the false positives among the organisms which composes the explanatory variables.

To address these issues and identify the most appropriate analytical approach, we propose a comparative study aimed at evaluating the impact of different transformations and normalization techniques applied to these explanatory variables on both the structure of the dataset itself and the selection of variables used to predict the immunological outcome. Consistently with what has been seen in (1) we observe highly different set of selected variables depending on which normalization method have been used.

To leverage this variability for meaningful biological insights, we looked at the intersection between groups of normalization methods based on several criteria: the context in which they

*Speaker

†Corresponding author: raluca.uricaru@u-bordeaux.fr

‡Corresponding author: patricia.thebault@u-bordeaux.fr

were originally developed (RNA-seq, microbiome data, ...), whether they involve a transformation or a scaling step, and the preprocessing performed before normalization, such as expressing counts as proportions. Then, to better understand the impact of transformation methods on raw counts and on the overall dataset, we examined the correlation structure among the explanatory variables,

In order to find how robust these selected variables can be, we can leverage the hierarchical structure of the taxonomy: repeating the analyses starting from the order levels down to genus level. This will have the advantage of finding potential taxonomic nodes (i.e. in fungus) at which it is not necessary to characterize at a deeper level to avoid having noisy data, and find at which taxonomic level, the signal is clearer for the association between certain organisms and the immunological outcome.

1. Karwowska et al. Effects of data transformation and model selection on feature importance in microbiome classification data. Microbiome 2025;13:2.

Keywords: Transcriptomics, Microbiome data, Metatranscriptomics