Analysis of gut *E. coli* composition in light of the whole gut microbiome

Jimmy Mullaert[1,2] (PhD), Olivier Tenaillon[1] (PhD), France Mentré[1,2] (MD, PhD)

1. Infection, Antimicrobials, Modelling, Evolution (IAME) UMR 1137, Inserm and Université Paris Diderot
2. Service de Biostatistiques - HUPNVS - AP-HP

Corresponding author: Jimmy Mullaert (jimmy.mullaert@inserm.fr)

Abstract:

High-throughput sequencing of the gut bacterial content (microbiome) is a promising approach to better understand the host-microbiome relationship, especially during and after an antibiotic-induced dysbiosis. Basically, sequencing pipelines produce a table of read counts, describing for each sample the relative abundance of so-called Operationnal Taxonomic units (OTU). Efficient statistical methods are thus needed to handle the specificity of such data.

In most of the published literature however, statistical methods applied to this table only rely on a quantitative parameter derived from the OTU table (e.g. relative abundance of some OTU or phyla, ratio of abundance between Bacteroides and Firmicutes, gene diversity or richness). This approach reduces the whole OTU table and its underlying phylogenic structure to a single quantitative variable and, hence, loses a lot of information. In addition, the justification for using one parameter or another is often missing and leads to an uncontrolled type-I error, unless a multiple testing correction is used.

Strategies to study associations between this kind of data and specified outcomes should take into account the underlying phylogenetic structure to avoid a loss of statistical power. We propose to use the Unifrac distance to derive a similarity matrix that serves as basis for a low dimensional MDS data reconstruction. Then, a distance-based regression allows the microbiota composition to be part of a fully general regression model, possibly including covariates and a continuous dependent variable.

In this presentation, we aim at analyzing data from the twinsUK cohort, which includes targeted 16S sequencing data as well as basic phenotypes (age, sex, BMI, monozygous/dizygous twin) of 465 twins, using the above-described methodology. In addition, counts for *E. coli* species are available for the same sample and can also be analyzed with the same methods. The two datasets (16S and *E. coli* specific) are also compared together to find relationships between the whole bacterial diversity and parameters derived from the *E. coli* table count (proportion of B2 type, diversity) and, conversely, associations between the *E. coli* diversity and parameters derived from the 16S table.