Projet BioBanaTom "Identification of new biomarkers of banana and tomato intake"

Protocol ID: 2016-A00153-48

Statistical Analysis for urinary and plasma metabolome changes after banana and tomato intake

Version 2 – 25/06/2018

The statistical analysis is centralized and performed by the INRA investigators, in close collaboration with the statistician of the Metabolomics platform (Natalia Vazquez INRA, Mélanie Pétéra, PFEM: Plateforme d'Exploration du Métabolisme).
The dataset resulting for, the urine samples of the clinical trial, ion intensities is normalized according to the volume collected for each urine fraction. A PCA model is applied to the normalized dataset to detect possible outliers in the dataset. Q primary PLSDA model is built using the three interventions as "Y" variables. The dataset are also assessed separately for banana versus control and tomato versus control.

To better discriminate the ions according to the intake of banana, an OSC-PLSDA model is built with the dataset containing control and banana samples (n=24). Statistical analysis is performed with SIMCA-P+ software (version 14.0, Umetrics, Umea, Sweden). The OSC filter is built with log transformed and Pareto scaled data using treatment (banana vs control; tomato vs control) as correction factor. A k-fold cross validation method is used with 4 subsets to predict the ability of the model to correctly classify samples to the corresponding dietary treatment. Q^2 is used to assess the prediction power and to determine the reliability of the model a permutation test (n= 500) is carried out.

Simultaneously, the difference in intensity of the extracted features between dietary treatments is assessed by applying a Student T test with Benjamini Hochberg correction (BH) for multiple comparisons using Metaboanalyst 3.0 [20]. Significance level is established at p value BH <0.05. Those features with a variable importance of projection (VIP) higher than 2 in the OSC-PLS-DA model, a T-test with BH p value <0.05 and which present in higher intensity in the urinary samples of the banana group are prioritized for identification as candidate biomarkers of banana intake**.** The same strategy is applied for the tomato dataset.

The dataset resulting from the analysis of plasma samples is assessed separately for banana and tomato versus control. Principal component analysis is performed for the detection of outliers in the dataset. For a first selection of ions, a one way ANOVA with a post hoc between the different time points is

applied using SPSS IBM version 22. Ions that were higher and significantly different than time point 00h (p value <0.01) are selected to be analyzed using a PLSDA model using SIMCA-P (version 14.0, Umetrics, Umea, Sweden). Ions with a VIP>2 and in higher intensity in banana or tomato are selected as plausible biomarkers of intake.