A fully automated pipeline for the analysis of Liquid Chromatography-Mass Spectrometry (LC-MS) based metabolomics experiments

Franceschi, P.[1,*]; Scholz, M.[1]; Shahaf, N.[1]; Wehrens, R.[1]

[1]Biostatistics and Data Management, Fondazione Edmund Mach

*Presenting author: **Pietro Franceschi** (pietro.franceschi@fmach.it)

Background

The recent improvement in analytical technologies has been the ground for the advent of high throughput metabolomics. This member of the "omics" family has the objective of fully characterizing biological systems at a comprehensive metabolic level. As in many other "omics" cases, however, metabolomics datasets show a high level of complexity and the development and the optimization of dedicated data preprocessing and analysis tools is of paramount importance to guide biological interpretation and biomarker identification.

Methods

In this communication we will present the fully automated data analysis pipline for LC-MS based metabolomics, which have been recently set-up at the FEM Research and Innovation Centre. The pipeline has been developed in R and aims at a seamless integration of data preprocessing, quality assessment, feature annotation and data analysis within a unified framework. The pipeline has been integrated in a web based application which can be directly used by the scientists lacking a specific bioinformatic background.

Data preprocessing and feature extraction is performed by the widely diffuse xcms package, running with a set of parameters tailored on the technological platform installed at FEM. After this step, the assessment of the data quality relies on set of visualization tools based on Principal Component Analysis. As a third step, experimental features are "annotated" with the objective of assigning them a "chemical" and "metabolic" identity. Since this last step represents undoubtedly the critical stage in metabolomics, a strong effort has been put into the annotation modulus, both by implementing annotation against an in-house developed database and also by developing a new tool to adaptively calculate the mass tolerance used for database search.

Results

Even if the pipeline is still under active development, preliminary results clearly indicates that:

1) R can be used as an effective environment to develop data analysis tools suitable for the use by a wide scientific community;

2) the solutions implemented to perform quality control allow the fast identification of critical/bad samples and also the early identification of drifts in the analytical pipeline;

3) the implementation of the adaptive mass tolerance window for database search guarantees an improvement in the quality of the annotation both in terms of a reduced number of false positive and of a better identification of low concentration compounds.