Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jprot

Enhancing metabolomics research through data mining☆,☆☆

Ibon Martínez-Arranz[a], Rebeca Mayo[a], Miriam Pérez-Cormenzana[a], Itziar Mincholé[a], Lorena Salazar[b], Cristina Alonso[a], José M. Mato[c,*]

[a]OWL, Parque Tecnológico de Bizkaia, Derio, Bizkaia, Spain
[b]Osarten kooperatiba elkartea, Mondragón, Guipúzcoa, Spain
[c]CIC bioGUNE, CIBERehd, Parque Tecnológico de Bizkaia, Derio, Bizkaia, Spain

ARTICLE INFO

Available online 7 February 2015

Keywords:
Metabolomics
Inter-batch normalization
Statistical assumptions
Aging
MANOVA
Linear regression

ABSTRACT

Metabolomics research, like other disciplines utilizing high-throughput technologies, generates a large amount of data for every sample. Although handling this data is a challenge and one of the biggest bottlenecks of the metabolomics workflow, it is also the clue to accomplish valuable results. This work has been designed to supply methodological data mining guidelines, describing systematically the steps to be followed in metabolomics data exploration. Instrumental raw data refinement in the pre-processing step and assessment of the statistical assumptions in pre-treatment directly affect the results of subsequent univariate and multivariate analyses. A study of aging in a healthy population was selected to represent this data mining process. Multivariate analysis of variance and linear regression methods were used to analyze the metabolic changes underlying aging. Selection of both multivariate methods aims to illustrate the treatment of age from two rather different perspectives, as a categorical variable and a continuous variable.

Biological significance

Metabolomics is a discipline involving the analysis of a large amount of data to gather relevant information. Researchers in this field have to overcome the challenges of complex data processing and statistical analysis issues. A wide range of tasks has to be executed, from the minimization of batch-to-batch/systematic variations in pre-processing, to the application of common data analysis techniques relying on statistical assumptions. In this work, a real-data metabolic profiling research on aging was used to illustrate the proposed workflow and suggest a set of guidelines for analyzing metabolomics data.
This article is part of a Special Issue entitled: HUPO 2014.

Abbreviations: UPLC, ultra performance liquid chromatography; MS, mass spectrometry; MANOVA, multivariate analysis of variance; ANOVA, analysis of variance; GGT, gamma glutamyl transferase; ALT, alanine transaminase; QC, quality control; RSD, relative standard deviation; Q–Q plot, quantile–quantile plot; VIF, variable inflation factor.

☆ This article is part of a Special Issue entitled: HUPO 2014.
☆☆ Financial Support: Supported by Spanish Plan Nacional I1D SAF 2011-29851 (J.M.M.), ETORTEK-2010 Gobierno Vasco (J.M.M.), Educación Gobierno Vasco 2012 (J.M.M.), BBVA Foundation (J.M.M.), MINECO-ISCiii PIE14/00031 (J.M.M.), Ministerio Economía y Competitividad IPT-010000-2010-013 (I.M.-A., R.M., I.M. and C.A.), Gobierno Vasco, Dpto. Industria, Innovación, Comercio y Turismo IG-2012/0000346 (I.M.-A., I.M. and C.A.). CIBERehd is funded by ISCiii.
* *Corresponding author at:* CIC bioGUNE, Parque Tecnológico de Bizkaia, 48160 Derio, Bizkaia, Spain. Tel.: +34 944 061300; fax: +34 944 0611301.
E-mail address: director@cicbiogune.es (J.M. Mato).

http://dx.doi.org/10.1016/j.jprot.2015.01.019

1. Introduction

Metabolomics, a new discipline in the group of 'omics' sciences, involves the study and characterization of small molecules in different biological matrices, such as biofluids, tissues, or cells. The metabolite fingerprint resembles a snapshot of a metabolic state; comparisons of different metabolic profiles let us distinguish between different physiological stages of an individual, or different individual stages [1–3]. Thus, it is not surprising that the number of metabolomics publications has been exponentially increasing over the last decade. Metabolomics is involved in many different areas of biomedical research such as cell biology [4–7], toxicology [8,9], nutrition [10–12], oncology [13–15], biomarker discovery [16–19] and diagnosis of different diseases such as diabetes and cardiovascular, neurodegenerative or liver-related diseases [20–24].

The success of this discipline has been paralleled by the outstanding improvements in high-throughput technologies. Nuclear magnetic resonance (NMR) and mass spectrometry (MS) combined with chromatography are the most frequent analytical platforms for the determination of metabolites in biological samples [25]. Due to its significantly higher throughput and sensitivity, MS plays a strategic role in the metabolomics field [26–28]. Consequently, metabolomics has progressed from qualitative and untargeted methods to a variety of targeted quantitative and semiquantitative approaches, now used as routine methods in many laboratories.

Despite the potential of these high-throughput technologies, the success of a metabolomics study relies on careful sample selection and some basic considerations about the experimental design [29,30]. As in other 'omics' fields, a typical metabolomic experiment is likely to generate an enormous amount of data. The task of extracting the hidden meaningful information buried in these data is essential but remains challenging. Indeed, the Metabolomics Standards Initiative (MSI), in an effort to make this process more accessible, has proposed general guidelines for various stages of metabolomics experiment, such as sample preparation, experimental analysis, quality control, metabolite identification and notation and data analysis [31–33].

We designed this work as a methodical data handling guideline, focused on obtaining a metabolic signature of aging in a healthy population. We described the steps for a correct application of multivariate analysis. 1) In data pre-processing, instrumental raw data is refined for posterior data analysis. This step includes intra- and inter-batch normalization, a critical procedure for analyses of large cohorts of samples. We introduced here the proposal to use robust regression methods to estimate the intra-batch drift function, as an improvement of previously described intra-batch normalization [34,35]. 2) Data pre-treatment step checks whether the variables and observations meet certain statistical assumptions. Statistical assumption assessment is always understood but often not mentioned in metabolomics publications. In this work, we meticulously described this process to avoid misunderstanding this situation as signifying a lack of importance. 3) Data analysis is the step where questions are answered, by finding for example the appropriate algorithm or the key biomarker. 4) Interpretation step finds the real-life significance of the results. Several multivariate methods can be applied to obtain a signature of the aging process. We believed it would be interesting to handle age as a categorical variable and a continuous variable to use two different strategies to achieve the same objective. This approach defines the type of statistical tests to be used and, therefore, affects the visualization of the results. A multivariate analysis of variance (MANOVA) and a linear regression model were tested as multivariate models to decipher the metabolic signature of aging.

2. Material and methods

2.1. Sample details

Serum samples and anthropometric data of healthy male and female volunteers included in this study were provided by the Basque Biobank for Research-OEHUN (http://www.biobancovasco.org/) and were processed with appropriate approval of the Ethics Committee. Inclusion criteria were normal blood pressure, normal urine and serum biochemistry, moderate alcohol intake (lower than 30 g/day) and body mass index under 30 kg/m^2. Individuals taking medication for diabetes, hypertension or hyperlipidemia were excluded. The study enrolled 263 volunteers, all of Caucasian origin.

Blood specimens were collected under fasting conditions, from May 2012 to May 2013. The samples from the donors were collected by venipuncture in Becton Dickinson yellow-top tubes. Blood samples were allowed to clot and then centrifuged at 1800 x *g*. Sera were analyzed in a COBAS 6000 (Roche Diagnostics GmbH, Germany) and hematological parameters in a GEN-S (Beckman COULTER Inc., USA) at OSARTEN K.E. laboratory. Serum aliquots were transferred to cryovials and stored at −80 °C until metabolomics analysis. Table 1 summarizes the analytical and hematological characteristics of the volunteers, classified according to their age and gender.

2.2. Metabolic profiling

Targeted serum metabolic profiles were semiquantified as previously described [34]. Briefly, ultra performance liquid chromatography (UPLC)-single quadrupole–MS amino acid analysis system was combined with two separate UPLC-time-of-flight–MS based platforms analyzing methanol and chloroform/methanol serum extracts. Identified ion features in the methanol extract platform included non-esterified fatty acids, acyl carnitines, N-acyl ethanolamines, bile acids, steroids, oxidized fatty acids, monoacylglycerophospholipids, and monoetherglycerophospholipids. The chloroform/methanol extract platform covered glycerolipids, cholesteryl esters, sphingolipids, diacylglycerophospholipids, acyl-ether-glycerophospholipids and primary fatty acid amides. Each extract was spiked with metabolites, not detected in unspiked human serum extracts: tryptophan-d5(indole-d5), PC(13:0/0:0), NEFA(19:0) and dehydrocholic acid in methanol extract; SM(d18:1/6:0), PE(17:0/17:0), PC(19:0/19:0), TAG(13:0/13:0/13:0), TAG(17:0/17:0/17:0), Cer(d18:1/17:0) and ChoE(12:0) in chloroform/methanol extract. These compounds were considered as internal standards and used for internal response correction, as detailed in the Pre-treatment section. Lipid nomenclature and