Analysis of the human adult urinary metabolome variations with age, body mass index and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses

Supporting Information

Etienne A. Thévenot, Aurélie Roux, Ying Xu, Eric Ezan, and Christophe Junot

Content

Data description and normalization (Figures S1 and S2) 2
Implementation of a comprehensive workflow for univariate hypothesis testing and OPLS modeling (Figures S3 and S4) 3
Statistical selection of metabolites with physiological variations (Figure S5) 4
Comparison between univariate and multivariate selection of metabolites in the case of single-response, single-predictive PLS models with standardized variables (theory; Figure S6) 5
References 7

Data description and normalization (Figures S1 and S2)

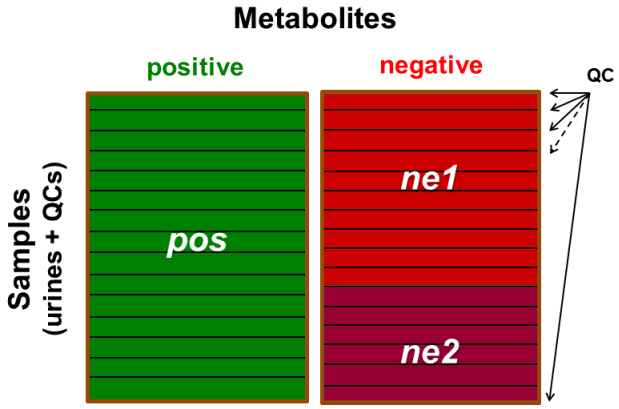

Figure S1: Data description. Initial datasets before normalization consist in two peak tables for the positive and negative mode, respectively, resulting from a three batch LC-MS acquisition (*pos*, *ne1* and *ne2*). Within each batch, pooled samples (QC) were measured regularly for subsequent normalization.

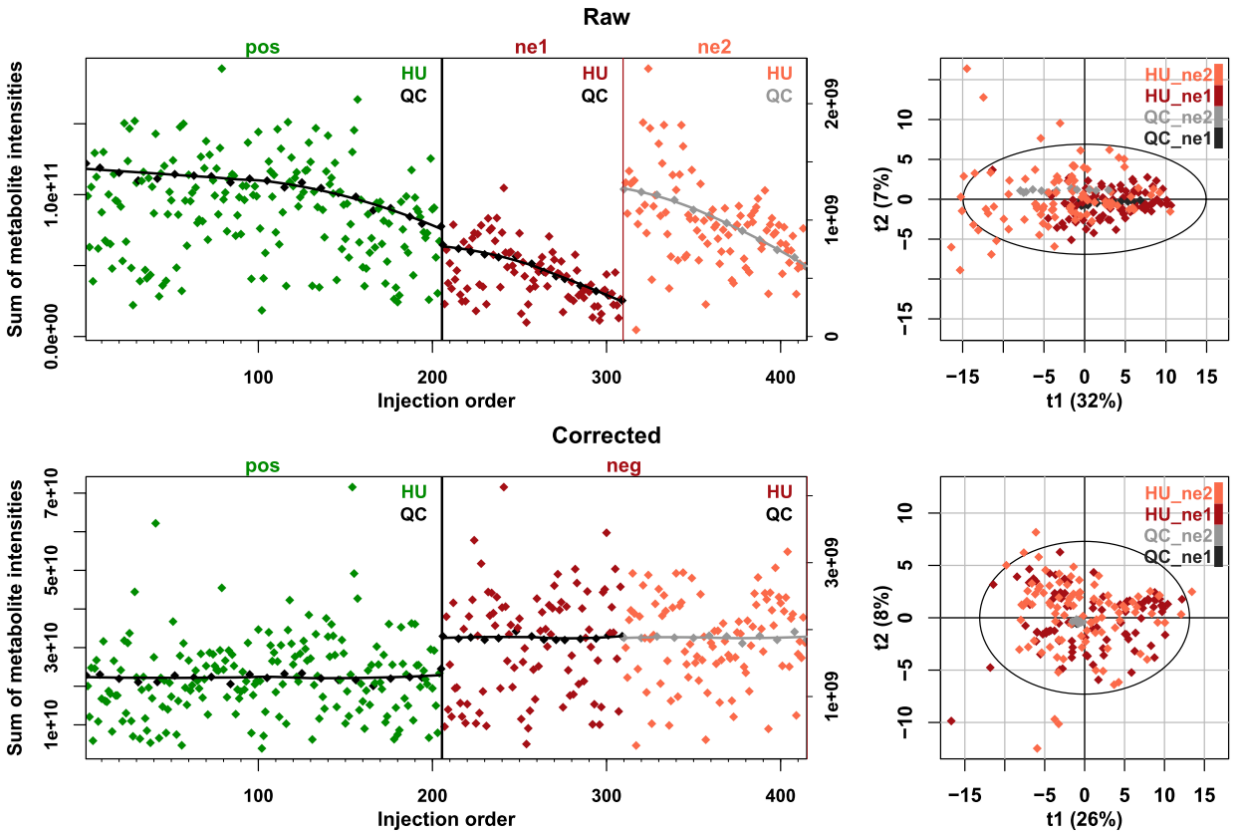

Figure S2: Signal drift correction and batch effect removal. Peak tables are visualized before (top) and after (bottom) batch normalization. Two types of graphics are used to assess the quality of signal drift correction and batch effect removal. **Left:** The sum of all metabolite intensities is plotted for each observation (HU or QC) as a filled diamond with a color corresponding to its batch. QCs are in black (or in grey for *ne2* to facilitate visual comparison with *ne1*). The *loess* curve obtained by QC modeling within each batch is superimposed. **Right:** Score plot resulting from principal component analysis of the negative ionization mode peak table. PCA was performed on standardized data and the percentage of total variance explained by the two first components is given in parenthesis. The Hotellings' T2 ellipse at a significance level of $p = 0.05$ is shown.

Implementation of a comprehensive workflow for univariate hypothesis testing and OPLS modeling (Figures S3 and S4)

The OPLS modeling of the age, BMI, and gender responses with the negative dataset as predictor models (Figure S3) were similar to those obtained with the dataset from the positive ion mode presented in the Figure 2 of the article, except for the BMI modeling which was not found significant by cross validation.

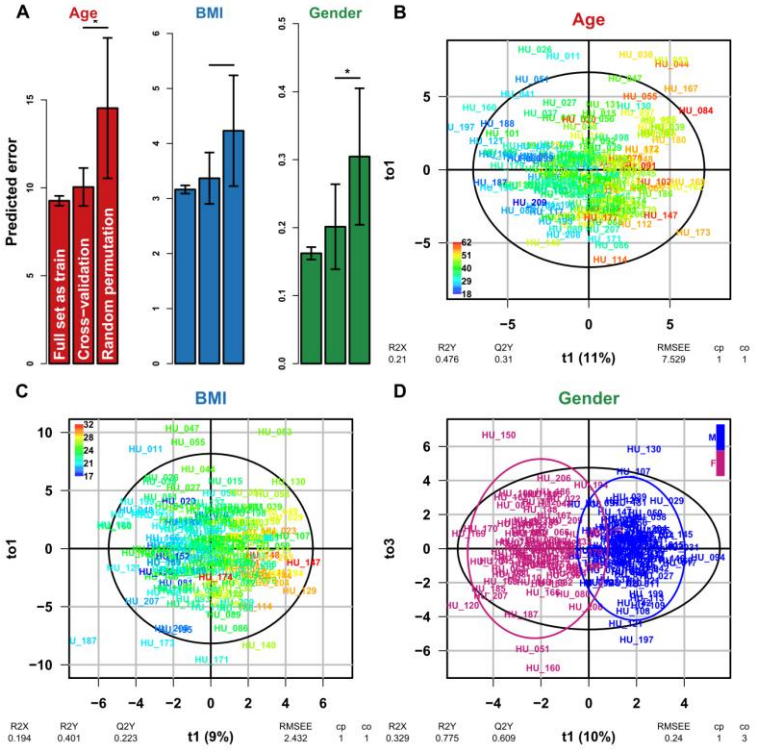

Figure S3:.OPLS models with data obtained in the negative ionization mode.

The response variance explained (R2Y) and the predictive performance of the model (Q2Y) were shown significant by comparing the model built with the true response values and 10^4 models built with random permutations of the response values[1] (Figure S4).

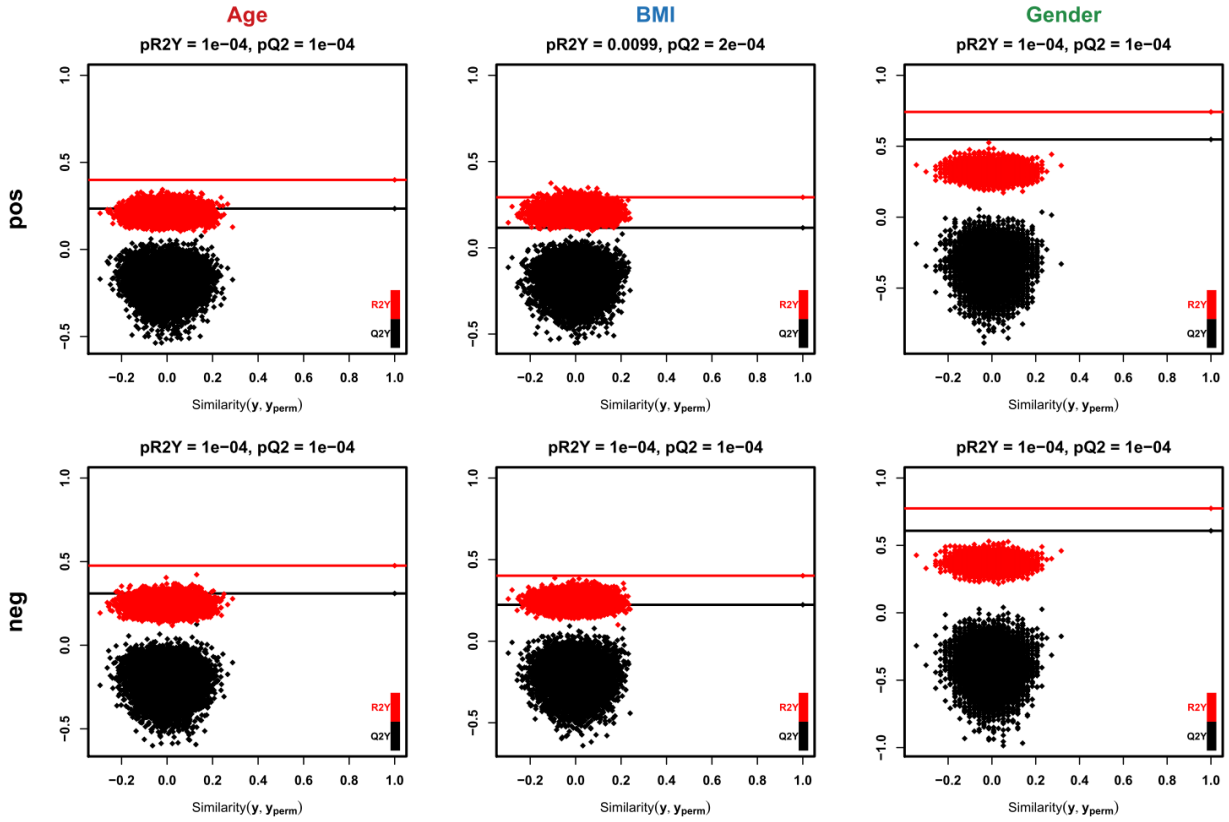

Figure S4. Significance of R2Y and Q2Y values estimated by permutation testing.

Statistical selection of metabolites with physiological variations (Figure S5)

The univariate-multivariate patterns obtained with the negative dataset (Figure S5) are similar to those obtained with the datasets from the positive mode presented in the Figure 3 of the article.

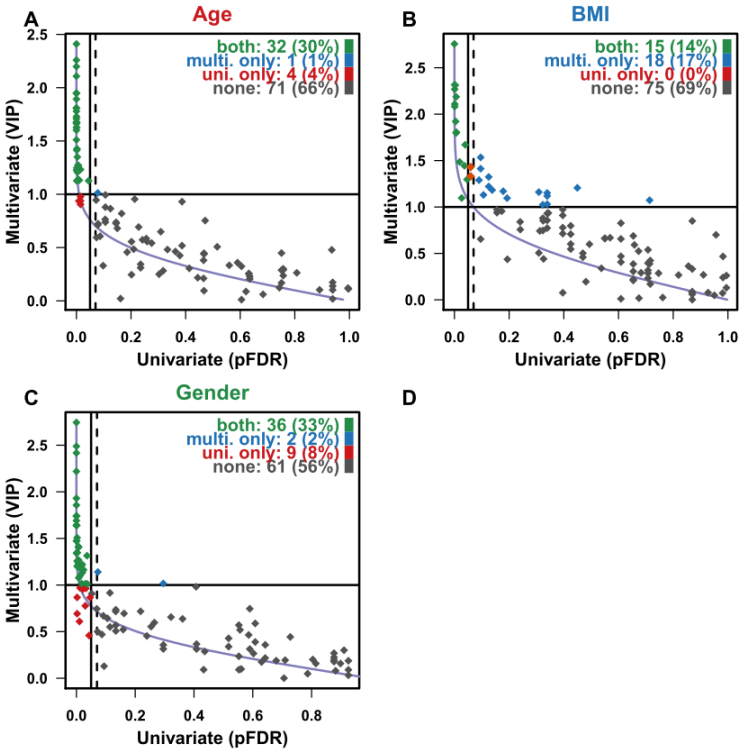

Figure S5. *p*-values vs. VIP comparison for the datasets from the negative ionization mode

Comparison between univariate and multivariate selection of metabolites in the case of single-response, single-predictive PLS models with standardized variables (theory; Figure S6)

In the case one-predictive component PLS or OPLS models of a single response are built from standardized predictors (i.e., mean-centered and unit-variance scaled), we demonstrate the relationship between VIP and *p*-values from the Pearson correlation test (to extend the demonstration to the qualitative Gender variable the two levels, "F" and "M" can be encoded as 1 and 2).

Algorithms will be described hereafter by using standard nomenclature: upper case bold face letters for matrices (e.g., **A**), lower case letters to denote vectors (bold face: e.g., **x**) and scalars (italics: e.g., *x*).

In the particular case of one-predictive component PLS models, the general formula[2] for VIPs can be simplified to:

$$VIP_j = \sqrt{m} \times |w_j| \qquad (1)$$

for each predictor *j*, were *m* is the total number of features and **w** is the vector of loading weights (||**w**|| = 1). For a single-response model, **w** is obtained by computing the vector **X**'**y** / **y**'**y**, where **X** is the matrix of predictors and **y** is the response,[3] and then setting its norm to 1. This is identical to computing **X**'**y** and setting its norm to 1. If the predictors and the response are mean-centered (as is generally the case), **X**'**y** is the vector of the covariances between the predictors and the response, and its *j*th element can thus be written as $r_j \times s_{xj} \times s_y$, where r_j is the correlation coefficient between **y** and the *j*th predictor $\mathbf{x_j}$, and s_y and s_{xj} their standard deviations. As the **X**'**y** will be normalized, only the $r_j \times s_{xj}$ product has to be computed. Thus

$$\mathbf{w} = \mathbf{r} \,.\, \mathbf{sd_x} \,/\, \|\mathbf{r} \,.\, \mathbf{sd_x}\| \qquad (2)$$

where **r** is the vector of correlation coefficients between the predictors and **y**, and $\mathbf{sd_x}$ is the vector of standard deviations of the predictors ("." is the cross product between two vectors).

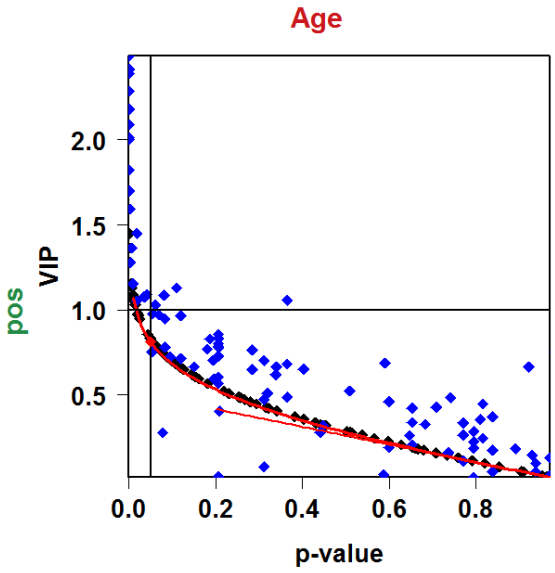

Figure S6. Relationship between VIP from one-predictive PLS or OPLS models with standardized variables, and *p*-values from Pearson correlation test. The (p_j, VIP_j) ordered pairs corresponding respectively to the VIP values from OPLS modelling of the age response with the *pos* dataset (standardized scaling), and the non-corrected *p*-values from the Pearson correlation test are shown as black diamonds. The $y = \Phi^{-1}(1 - x / 2) / z_{rms}$ is shown in red (z_{rms} is the quadratic mean of the z_j quantiles from the standard normal distribution; $z_{rms} = 2.5$ for the *pos* dataset and the age response). The half-tangent at $p = 1$, whose slope is $-1 / z_{rms} \times 1 / 2 \times \sqrt{(2 \times \pi)}$, is displayed in red. The vertical (resp. horizontal) black line corresponds to the univariate (resp. multivariate) thresholds. The point on the curve with the univariate 0.05 threshold value as abscissa and $1.96 / z_{rms}$ as ordinate is colored in red. The ($pFDR_j$,VIP_j) points corresponding to the corrected *p*-values of the non-parametric Spearman test (i.e., identical to Figure 3A) are superimposed in blue.

If we further assume that the predictors are standardized, equation (2) becomes:

$$\mathbf{w} = \mathbf{r} / \|\mathbf{r}\| \qquad (3)$$

Finally, the relationship between the Pearson correlation coefficient r_j and the Student t_j statistic of the parametric test is well known:

$$r_j = t_j / \sqrt{(t_j^2 + n - 2)} \qquad (4)$$

where n is the total number of observations. For $n - 2 > 30$, the Student distribution T can be approximated by the standard normal distribution Z, leading to:

$$t_j \sim z_j = \Phi^{-1}(1 - p_j / 2) \qquad (5)$$

where Φ^{-1} is the inverse of the probability density function of the standard normal distribution and p_j is the *p*-value of the correlation test.

By combining equations (4) and (5) we get:

$$r_j = z_j / \sqrt{(z_j^2 + n - 2)} \qquad (6)$$

Equation (6) can be approximated by its tangent around 0,

$$r_j \sim z_j / \sqrt{(n - 2)} \qquad (7)$$

because, on the one hand, when p_j varies from 0 to 1 ($1 - p_j /2$ varies from 1 to 0.5), z_j tends rapidly towards its vertical asymptote (equation 5), and, on the other hand, since $0 << n - 2$, r_j varies slowly towards its asymptote (equation 4).

Combining (3) and (7) leads to:

$$\mathbf{w} \sim \mathbf{z} / \|\mathbf{z}\| \qquad (8)$$

Finally, by combining equations (1) and (8) and noting z_{rms} the root mean square (or quadratic mean) of z values, $z_{rms} = \sqrt{(1 / m \times (z_1^2 + z_2^2 + \ldots + z_m^2)}$, we obtain:

$$VIP_j \sim 1 / z_{rms} \times z_j \qquad (9)$$

Equation (9) indicates that the shape of the $VIP = f(p)$ curve comes from the Φ^{-1} function and that the curvature increases with the z_{rms} factor (i.e., the quadratic mean of the z_j values). The latter point is illustrated on Figure S6 by:

i) drawing the curve half-tangent at $p = 1$, whose slope is proportional to $-1 / z_{rms}$ (the actual value is $-1 / z_{rms} \times 1 / 2 \times \sqrt{(2 \times \pi)}$),

ii) plotting the point at $p = 0.05$: since, $z_{p = 1 - 0.05 / 2} = 1.96 \sim 2$, the ordinate is $\sim 2 / z_{rms}$

The curve therefore passes below the intersection of the two threshold lines if the z_j values have a (quadratic) mean superior to the $z_{p = 1 - 0.05 / 2} = 1.96$ threshold value.

References

(1) Szymanska, E.; Saccenti, E.; Smilde, A.; Westerhuis, J. Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics* **2012**, *8* (1), 3-16.

(2) Mehmood, T.; Liland, K.H.; Snipen, L.; Saebo, S. A review of variable selection methods in Partial Least Squares Regression. *Chemom. Intell. Lab. Syst.* **2012**, *118* (), 62-69.

(3) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58* (2), 109-130.