# Variance component analysis to assess protein quantification in biomarker discovery. Application to MALDI-TOF mass spectrometry

Catherine Mercier, Amna Klich, Caroline Truntzer, Vincent Picaud, Jean-François Giovannelli, Patrick Ducoroy, Pierre Grangeat, Delphine Maucort-Boulch, Pascal Roy

# Variance component analysis to assess protein quantification in biomarker discovery. Application to MALDI-TOF mass spectrometry

Catherine Mercier[1,2,3,4,5], Amna Klich[1,2,3,4,5], Caroline Truntzer[6], Vincent Picaud[7,8], Jean-François Giovannelli[9,10,11], Patrick Ducoroy[6], Pierre Grangeat[12,13], Delphine Maucort-Boulch[1,2,3,4,5], Pascal Roy[1,2,3,4,5]

1 Service de Biostatistique-Bioinformatique, Hospices Civils de Lyon, Lyon, France
2 Université de Lyon, Lyon, France
3 Département Biostatistiques et Modélisation pour la santé et l'environnement, Université Lyon 1, Villeurbanne 69622, France
4 Pôle Rhône-Alpes de Bioinformatique (PRABI), Villeurbanne, France
5 CNRS UMR 5558, Laboratoire de Biométrie et Biologie Évolutive (LBBE), Équipe Bio-statistique Santé, Villeurbanne, France
6 Clinical and Innovation Proteomic Platform (CLIPP), Pôle de Recherche Université de Bourgogne, Dijon, France
7 Commissariat à l'Énergie Atomique et aux Énergies Alternatives, Gif-sur-Yvette, France
8 Université Paris-Saclay, Saint-Aubin, France
9 CNRS UMR 5218, Laboratoire de l'Intégration du Matériau au Système (IMS), Talence, France
10 Département micro Technologies pour la biologie et la santé, Université de Bordeaux, Talence, France
11 Institut Polytechnique de Bordeaux (Bordeaux INP), Talence, France
12 Innovation en micro et nanotechnologie, Université de Grenoble-Alpes, Grenoble, France
13 Commissariat à l'Énergie Atomique et aux Énergies Alternatives, Laboratoire d'Électronique et de Technologie de l'Information, MINATEC Campus, Grenoble, France

Correspondence : Catherine Mercier, Service de Biostatistique-Bioinformatique, Hospices Civils de Lyon, 162 avenue Lacassagne, Lyon, France. Email: catherine.mercier@chu-lyon

## Abstract

Controlling the technological variability on an analytical chain is critical for biomarker discovery. The sources of technological variability should be modeled, which calls for specific experimental design, signal processing, and statistical analysis. Furthermore, with unbalanced data, the various components of variability cannot be estimated with the sequential or adjusted sums of squares of usual software programs. We propose a novel approach to variance component analysis with application to the matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) technology and use this approach for protein quantification by a classical signal processing algorithm and two more recent ones (BHI-PRO 1 and 2). Given the high technological variability, the quantification failed to restitute the known quantities of five out of nine proteins present in a controlled solution. There was a linear relationship between protein quantities and peak intensities for four out of nine peaks with all algorithms. The biological component of the variance was higher with BHI-PRO than with the classical algorithm (80–95% with BHI-PRO 1, 79–95% with BHI-PRO 2 vs. 56–90%); thus, BHI-PRO were more efficient in protein quantification. The technological component of the variance was higher with the classical algorithm than with

BHI-PRO (6–25% vs. 2.5–9.6% with BHI-PRO 1 and 3.5–11.9% with BHI-PRO 2). The chemical component was also higher with the classical algorithm (3.6–18.7% vs. < 3.5%). Thus, BHI-PRO were better in removing noise from signal when the expected peaks are detected. Overall, either BHI-PRO algorithm may reduce the technological variance from 25 to 10% and thus improve protein quantification and biomarker validation.

## Keywords

biomarker discovery, experimental design, sum of squares type, technological variability, variance components

## 1 INTRODUCTION

A biomarker is "any substance, structure, or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease" (WHO, 2001). In the recent years, there has been a growing interest in using high throughput technologies to discover biomarkers. Biomarker discovery is based on different expressions of biomarkers between groups. Depending on the study design, these groups may be either diagnostic groups (diseased vs. healthy subjects), prognostic groups (relapsing vs. remitting event-free subjects), or theranostic groups (responders vs. nonresponders to a specific treatment). A differential expression is usually defined by the ratio of between-group variance to within-group variance, which are both components of the biological variance. When technological variance is added to biological variance (to within-group variance or to within- and between-group variances), this ratio becomes smaller and group comparisons become less powerful and less reliable. Controlling the technological variability on analytical chains is thus a critical point in biomarker discovery. This is why the sources of technological variability should be modeled and taken into account through experimental design and statistical analysis (Cairns, 2011; Carr, 2014; Käll & Vitek, 2011; Mercier, Truntzer, & Pecqueur, 2009).

Thanks to a reasonable cost and a simplified sample preparation, linear MALDI-TOF spectrometry is a growing technology in clinical microbiology. It allows identification of pathogen biomarkers in body fluids and large-scale detection of proteins present in a complex biological mixture. The measurements reflect the proteomic profiles of the samples under study. However, due to the low resolution of linear MALDI-TOF instruments, a robust and accurate protein quantification remains a challenging task.

We addressed the question of protein quantification within the context of biomarker discovery by linear MALDI-TOF spectrometry. Signal processing is necessary to extract meaningful biological information from the observed signal. This processing may fail in removing noise from signal and introduce artefacts, which increases the technological variability in measuring biomarker abundance. In addition, before signal processing, a purification process is used to simplify the biological mixture by chemical treatments and each step of this process is an additional source of technological variability.

Maximizing the ratio biological variance to total variance to improve quantification requires a complex modeling in case of unbalanced data and presence of interactions. The usual approaches of variance decomposition cannot always estimate the specific contribution of biological versus technological sources of variability and their interactions. A global approach is proposed here to estimate all the components of the total variance using a specific experimental design and a related statistical analysis plan. This approach is applied within the context of biomarker discovery to linear MALDI-TOF analytical chain using two types of algorithms for signal processing: the classical one and a novel one designed to decrease the impact of the technological variability.

In the following section, MALDI-TOF technology will be briefly described to identify the main sources of variability in its measurements. Section 2.1 will present the data acquisition, the experimental plan, and the sources of technological variability considered and will explain to which extent biological variability is controlled. Section 2.2 will briefly describe the differences between the two types of algorithms used for signal processing. Section 2.3 will present the statistical analysis plan explaining the methods used to quantify the variance components related to the main sources of variability.

# 2 METHODS

## 2.1 Data acquisition and experimental plan

Because the true proteomic profiles in biological samples are unknown, a known biological variability of protein quantities was generated here by dilution. These quantities were generated from a standard preparation used for quality control, the ClinProt Standard (CPS, Bruker Daltronics, Bremen, Germany). The CPS is a mixture of commercial calibrants that contains 11 known proteins. A serial dilution of the CPS stock solution was prepared in saline with dilution factors 1, 2/3, 1/2, 1/4, 1/8, 1/16, 1/32, which corresponds to 160, 120, 80, 40, 20, 10, and 5 µL of CPS. The control solution was pure saline. A constant volume (40 µL) of each dilution was added to a constant volume of plasma (160 µL of a single plasma sample). This way, the volume of CPS used reflects the relative abundance of CPS proteins in the total mixture and the relative abundances of native plasma proteins are constant in the serial dilution; only the relative abundances of CPS proteins vary with the dilution factor. Herein, for clarity or practicality, "relative abundance of CPS proteins" is sometimes replaced by "protein quantities."

Before signal acquisition, the samples were submitted to chemical treatments that simplify them by retaining specific proteins on the basis of their chemical properties. All the samples of the serial dilution were submitted to three main successive steps:

(i) Equalization: To reduce the dynamic range of protein abundances, the samples were treated with ProteoMiner technology (BioRad Ref. 136.3012) using direct elution. The samples were mixed with a highly diverse library of hexapeptides bound to beads. Because the bead capacity limits the binding capacity, highly abundant proteins saturate quickly their ligands and excess proteins are washed out during the process. In contrast, medium- and low-abundance proteins are concentrated on their specific affinity ligands.
(ii) Purification: The samples were purified using C8 hydrophobic magnetic beads to retain specific proteins according to their biochemical properties.
(iii) Spotting: The purified samples were spotted on ground steel target plates with an α-cyano-4-hydroxy-cinnamic matrix. This matrix confers some properties to the sample proteins so that they can be ionized. Ions are accelerated into the flight tube and enter a magnetic field-free region where they are separated according to their velocities (and subsequently sizes) before hitting the detector located at the other end of the tube. The signal acquisition is performed spot by spot in a sequential manner by moving the plate in front of a laser beam.

MS spectra were acquired in a linear mode with UltraFlex Extreme MALDI-TOF (Bruker Daltonics). A robot with a Multi Chanel Arm (96 needles) was used for purification and spotting to decrease the technological variability. This technology is able to analyze 384 samples in a single acquisition, which was sufficient to acquire the whole dataset (with technical replicates) in a single run.

The experimental design that explored the technological variability in the MALDI analytical chain included technical replicates at each step of the chain: (i) Equalization in quadruplicates, (ii) Purification in triplicates, (iii) Spotting in quadruplicates. The experimental design for each dilution *i* is summarized in Figure 1. The total number of replicates for each dilution was then 4 × 3 × 4 = 48. The expected number of spectra for all dilutions was then 8 × 48 = 384 spectra. The spectra from dilution 1 being of poor quality, we kept only 7 × 48 = 336 spectra for analysis.

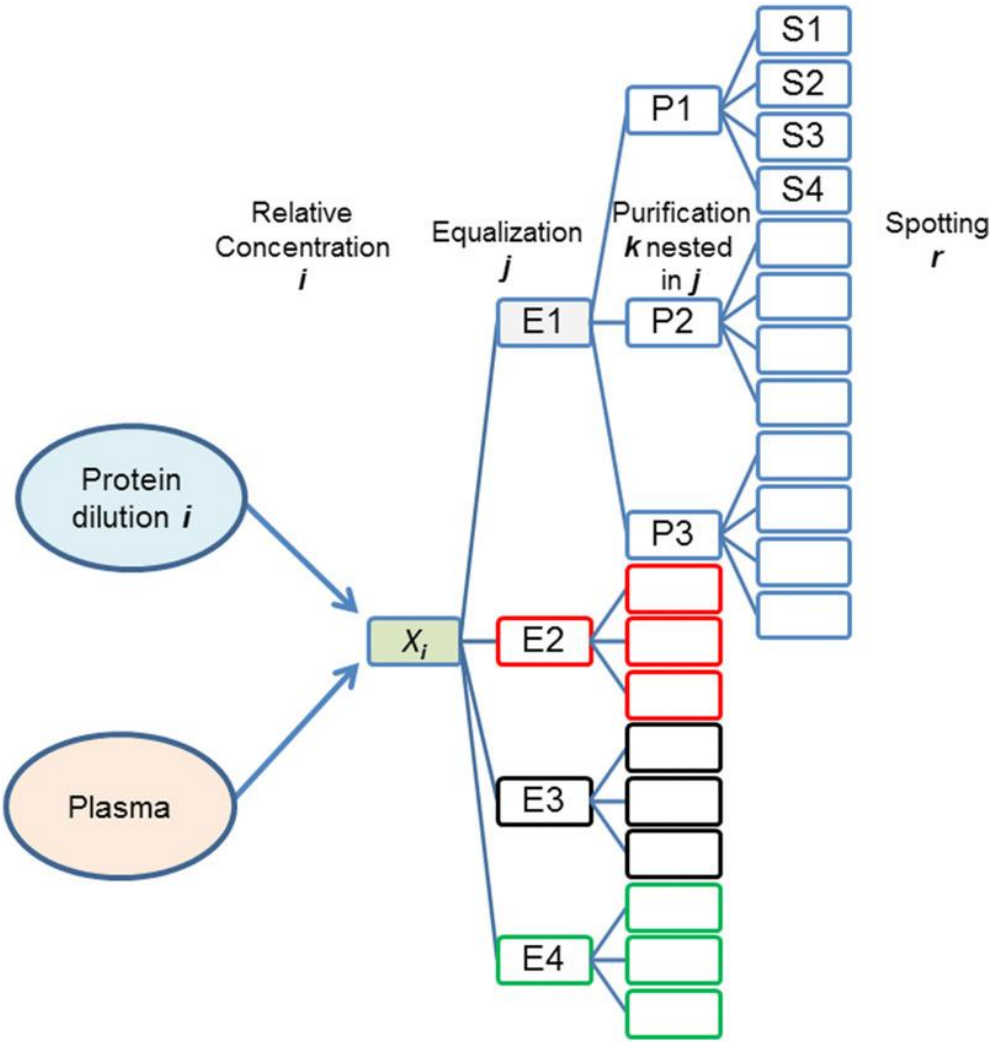


Figure 1. Experimental design. The experimental design generated the biological variance by controlling the relative abundance of proteins via dilution and allowed for the components of the technological variance via replicates at each step of the analytical chain

Moreover, technical controls were added on the acquisition target plates to ensure the quality and reproducibility of the acquisition step. These controls followed the same preanalytical process.

## 2.2 Signal processing

Once acquired in a linear mode, the spectra were processed to extract the meaningful biological information from the observed signal by retrieving the artefacts of technical origin using three algorithms for signal processing: one classical algorithm and two versions of a novel one. The classical signal-processing algorithm was developed by a proteomic platform (CLIPP, Dijon, France). The new signal processing, "BHI-PRO" was developed as part of

BHI-PRO project dedicated to Bayesian hierarchical inversion for mass spectrometry and its application to discovery and validation of new protein biomarkers (Dridi et al., 2014; Gerfault et al., 2014; Gerfault et al., 2013; Grangeat et al., 2013; Szarcherski, 2012; Szarcherski et al., 2013; Szacherski et al., 2014).

After spectra alignment, each spectrum was analyzed according to the classical methods of data processing (Coombes, Baggerly, Morris, Dubitzky, & Berrar, 2007; Roy, Truntzer, Maucourt-Boulch, Jouve, & Molinari, 2011). The last stage was peak list extraction. The extracted peak list is then used as a list of variables in statistical analysis. Usually, peak list extraction is performed in two steps: first, baseline removal (Mazet, 2004; Morháč, 2009; Zhang, Chen, & Liang, 2010), then peak selection using the baseline-free spectrum (Antoniadis, Bigot, & Lambert-Lacroix, 2010; Renard, Kirchner, Steen, Steen, & Hamprecht, 2008; Yang, He, & Yu, 2009). The removal of the baseline is often a critical task because it may introduce irreversible artefacts that corrupt the peak-finding procedure. To alleviate this, the new BHI-PRO algorithms perform the two steps simultaneously.

With the classical algorithm, the first stage of signal processing was denoising; that is, removing the random electronic noise. This was performed with the wavelets methodology. Each spectrum was decomposed into detail and approximation coefficients. Detail coefficients were thresholded so as to retain only the highest coefficients. Once thresholded, the retained coefficients were back-transformed into spectra. Once denoised, the subtraction of the baseline was performed so as to remove the chemical noise. For this, smoothing splines were fitted to local minima of the spectra. This smoothed signal was then subtracted from the original signal. Normalization then aimed at setting all of the spectra on the same scale by dividing each spectrum by the standard deviation of all spectrum intensities (Meuleman et al., 2008). All the spectra were then averaged to calculate a mean spectrum and determine the peak positions. The peak intensities were then measured independently in each spectrum using the peak area (Morris, Coombes, Koomen, Baggerly, & Kobayashi, 2005).

The BHI-PRO algorithm is available in two versions: BHI-PRO 1 processes a single spectrum at a time whereas BHI-PRO 2 is able to process several spectra at the same time. Contrarily to the classical algorithm, BHI-PRO algorithms perform simultaneously the computation of the baseline, the deconvolution, and denoising. For reference, see the open source implementation available on GitHub (https://github.com/vincent-picaud/Joint_Baseline_PeakDeconv).

The input spectra are stacked column by column in a matrix $y$. The BHI-PRO algorithms solve an optimization problem to find two matrices $x_p$ and $x_b$.

The columns of the $x_p$ matrix are sparse vectors in which nonzero components define the position (index of the component) and the height (magnitude of the component) of the deconvolved peaks. The columns of the $x_b$ matrix are vectors that store the computed baseline.

$$(x_b, x_p) = \arg\min_{x_b, x_p \geq 0} \frac{1}{2} \left\| y - (x_b + PSF * x_p) \right\|^2 + \frac{\mu}{2} \|D x_b\|^2 + \frac{\lambda_2}{2} \left\| x_p \right\|^2 + \lambda_1 P_p(x_p)$$

In this equation, the Point Spread Function (*PSF)* is the peak shape, i.e., Gaussian-shaped peak, $D$ is a finite difference matrix used to enforce baseline $x_b$ smoothness, and $P_p$ ($x_p$) is a Lasso-like penalty term (Yuan & Lin, 2006) used to enforce deconvolved peak $x_p$ sparsity.

With BHI-PRO 1, there is only one column, thus y, $x_p$, and $x_b$ are vectors. The penalty term is:

$$P_p(x_p) = \left\| x_p \right\|_1$$

and the solved problem is similar to a Lasso regression. A first run was performed on the mean spectrum of all spectra. The detected peaks were then used to permit only certain

common peak positions. The algorithm was then run on each spectrum with this restricted set of peak positions.

With BHI-PRO2, the penalty term is:

$$P_p(x_p) = \sum_i \left\| x_p[i,:] \right\|_1$$

and the solved problem is similar to a group Lasso regression. This regression groups peaks that share the same positions (the same i-row that represents a common *m/z* value). BHI-PRO 2 was directly applied to the whole set of spectra.

Both versions of BHI-PRO were applied after spectra alignment and normalization of the total ionic current. The number of selected peaks depended essentially on the Lasso regularization parameter $\lambda_1$. This parameter was selected by visual inspection and set to a low value to insure that even small peaks are picked. For total-ion-current-normalized spectra, $\lambda_1 = 2.5 \times 10^{-5}$, $\lambda_2 = 10^{-4}$, and $\mu = 1500$ were used in all computations.

All three quantification approaches (i.e. classical, BHI-PRO 1, and BHI-PRO 2) were applied to the experimental data. The results of the three algorithms provided three estimations of each protein quantity. As the *m/z* ratios of CPS proteins are already known, the positions of the peaks that should be detected in the chosen mass window (1000–10,000 Da) are already known too. Their positions on the *m/z* axis were 1047, 1297, 1349, 1619, 2094, 2467, 3150, 5734, and 8602 Da.

## 2.3 Statistical analysis plan

The performance of the three algorithms was first tested through checking whether the effect of dilution (i.e. division of protein quantities) can be reflected by the peak intensities; then their abilities to minimize the technological part versus the biological part of the variance were compared.

The criteria used to evaluate the performance of the quantification method were: (i) the biological part of the variance (stemming from the serial dilution); (ii) the technological part of the variance (according to the replicates involved in the analytical chain); and, (iii) the modeling error.

The investigation of the technological part of the variance included the parts of the variance due to two chemical steps of the MALDI analytical chain: equalization and C8 purification. The third step (spotting) was allowed for through the residual variance.

Only peaks with a monotone relationship between peak intensities and protein quantities were selected for the statistical analysis. Four peaks were then concerned and were the same with the three algorithms under comparison. Their positions on the *m/z* axis were: 1349, 2094, 3150, and 5734 Da.

A log-transformation of the peak intensities was used to stabilize the variance. This led to exclude from the analysis the control solution (protein quantity = 0) and the lost values generated by null intensities. As twofold dilutions were used (but for dilution factor 2/3), a log2 transformation was applied to the peak intensity *y* and the protein quantity *x*.

The control data (relative to the sample devoid of proteins) were used to compare graphically (boxplots) the peak intensities observed (before log2 transformation) and check whether the relative concentration of a protein is higher than in the control solution. This proved untrue in five peaks out of 9.

For each algorithm and each peak, a linear regression model was built to link the peak intensity to the protein quantity. Applying the log2-log2 transformation, a linear relationship on the transformed scale (i.e. $\log_2 y = \beta_0 + \beta_1 \log_2 x$) corresponds to a polynomial relationship on the original scale (i.e. $y = 2_0^{\beta} \times x_1^{\beta}$). When the slope coefficient $\beta_1$ equals 1 on the log2-log2 scale, the relationship is also linear on the original scale because $y = 2_0^{\beta} \times x$.

The CPS quantity is known but the true abundance of the proteins in the mixture is not exactly known because the chemical steps change the initial abundances. This error on protein abundance is a Berkson error; this means that the estimations of the slope and the intercept of the linear relationship will not be biased but that an error on *x* (the protein abundance) will be transferred to *y* (Berkson, 1950; Carroll, Ruppert, & Stefanski, 1995).

Because the chemical steps may influence the relationship between protein abundance and peak intensity, the slopes, and intercepts were estimated for each equalization replicate and each C8 replicate nested in the equalization factor in a hierarchical model. Because of convergence problems, a fixed-effects model was used instead of a mixed-effects model in the variance component analysis. This is advantageous, especially when the random factor has less than five levels, in which case the estimates of the variance components may be unreliable (Piepho, Büchse, & Emrich, 2003). The R code used for the variance component analysis on simulated data with both models is available as Supplementary Material.

A simple coding was used for the contrast matrix to enable estimating a mean intercept and a mean slope. Model 1 presents these means and deviations from these means in two terms, the first one for the intercept and the second (in brackets) for the slope (See Table 1).

| **Model** | **Formula** |
|---|---|
| Model 1 | $y_{ijkr} = \beta_0 + \beta_{0j}E_j + \beta_{0k(j)}P_k + (\beta_1 + \beta_{1j}E_j + \beta_{1k(j)}P_k)x_i + \varepsilon_{ijkr}$ |
| Model 2 | $y_{ijkr} = \beta_0 + \beta_1 x_1 + \varepsilon_{ijkr}$ |
| Model 3 | $y_{ijkr} = \beta_0 + \beta_{0j}E_j + \beta_{0k(j)}P_k + \varepsilon_{ijkr}$ |
| Model 4 | $y_{ijkr} = \beta_0 + \beta_{0j}E_j + \beta_{0k(j)}P_k + \beta_1 x_i + \varepsilon_{ijkr}$ |
| Model 5 | $y_{ijkr} = \beta_0 + \beta_{0j}E_j + (\beta_1 + \beta_{1j}E_j)x_i + \varepsilon_{ijkr}$ |

Table 1. Models used to link peak intensity to protein abundance

In this model, *i*, *j*, *k*, and *r* are the indices for, respectively, the protein quantity, the equalization replicate, the purification replicate nested in equalization, and the spotting replicate. *E* is the equalization factor, *P* the Purification C8 factor, $\beta_0$ the intercept coefficient, and $\beta_1$ the slope coefficient. Finally, the residual error is $\varepsilon_{ijkr} \sim N(0, \sigma'^2)$

To calculate the slope of each equalization, we added the mean slope coefficient to the slope coefficient of the corresponding equalization. There were three slope coefficients for four equalizations; the fourth slope coefficient was obtained by subtracting the sum of the three slope coefficients from the mean slope coefficient. The coefficients of Model 1 (Table 1) were considered as significantly different from zero when $p < 0.05$ in Wald test.

In the graphical representation of the relationship between the protein quantity and the observed peak intensity on the log2-log2 scale (Fig. 2), the log2 intensities of the peaks were standardized (centered on the mean and divided by the standard deviation) to improve the detection of the outliers seen on the graphs (the distance was expressed in standard deviations). The global linear trend (mean slope plus mean intercept) represents the biological effect (main effect). The points' dispersion on the vertical axis represents the other sources of variability (nonbiological variance and its interaction with the biological variance). The variability of the slopes represents the technological variability (and its interaction with the biological variability) induced by equalization and purification whose effects depend on the properties of each protein. This variability was represented by 12 straight lines (4 equalizations × 3 purifications).

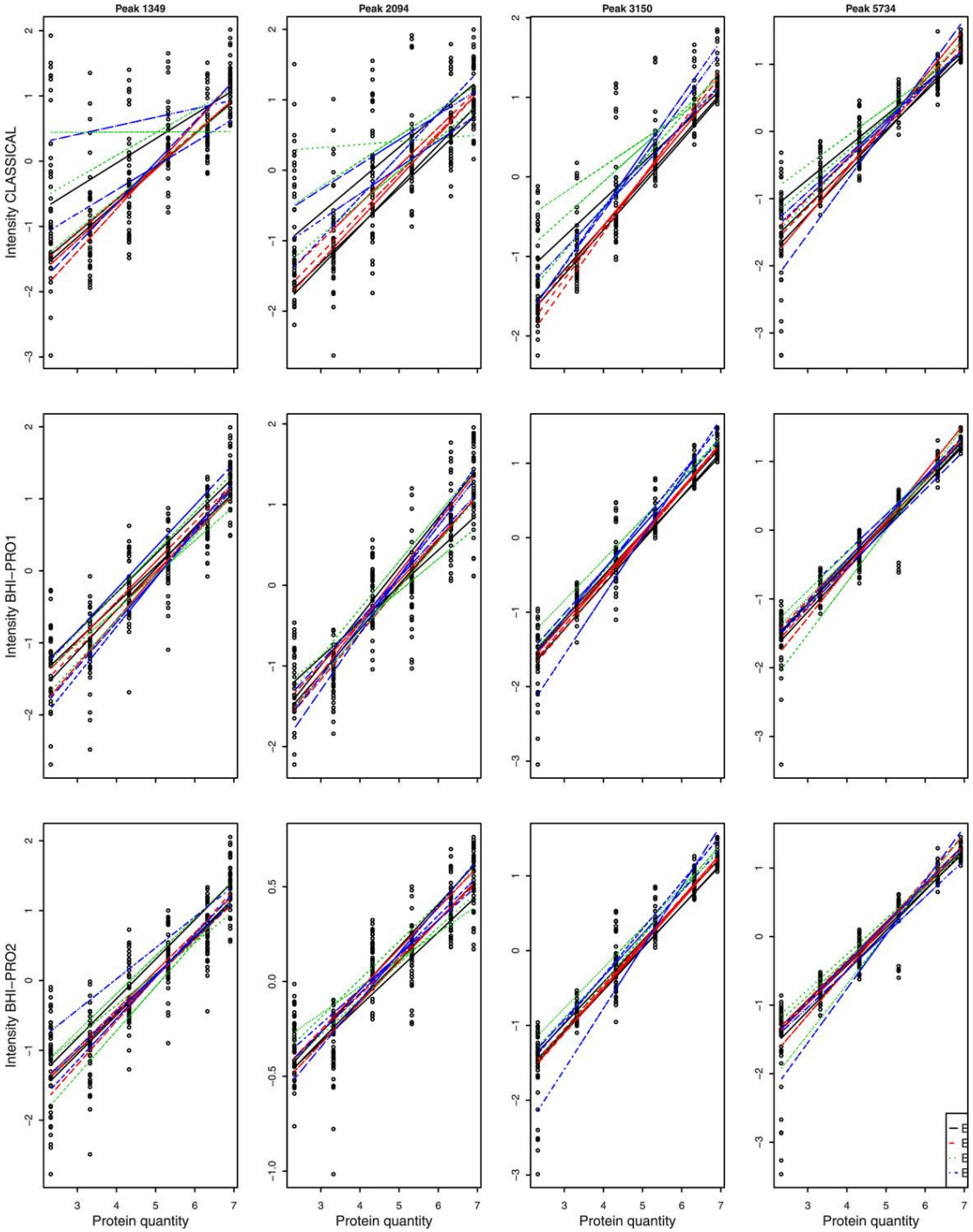

Figure 2. Protein-signal relationship. Linear relationship between the abundance of proteins and peak intensity with the three algorithms (classical, BHI-PRO1, and BHI-PRO2) in four peaks: 1349, 2094, 3150, and 5734 Da (*m/z* position in the spectrum). Both axes are expressed on a $\log_2$ scale. In each panel, 12 lines represent the linear relationships relative to three purifications within each of four equalizations. The legend in the far right panel corresponds to the equalizations

To validate the linear model, we carried out a classical graphical analysis of the residuals (not shown) and checked the independence of the residuals and the homogeneity of the variance across the whole range of observations. Three outliers were excluded (values between 6 and 11 standard deviations under the mean intensity of peak 2094 Da with algorithm BHI-PRO 2).

The final phase was the decomposition of the variance into biological variance and technological variance. The experimental design is unbalanced because of the lost values generated by null intensities or the exclusion of a few outliers. Imbalance introduces a correlation between the main effects and their interactions. The classical approaches that use sequential or adjusted sums of squares (most software programs) cannot estimate the variance components of several sources of variation (called "factors" in the analysis of variance) and their interactions in unbalanced designs. Using sequential sums of squares ("Type I" in most software programs), the first factor in the model is assigned all of the shared variation; the estimation of the main effects depends then on the order of the factors. Using the adjusted sum of squares with higher level terms (interactions) included in the model ("Type III" by default in many software programs) means that each factor is adjusted on the others and on their interactions but then the relationships between higher- and lower order terms (i.e. marginality) is lost. The presence of an interaction between two factors means that both factors are important but that the effect of each depends on the other. The adjusted sum of squares with higher level terms omitted ("Type II") is preferable to associate each factor with the variances of its main effect and the variances of their interactions but the sum of squares will still depend on the factor order.

The analysis of variance (ANOVA) for unbalanced data (Hector, von Felten, & Schmid, 2010; Langsrud, 2003; Nelder & Lane, 1995; Shaw & Mitchell-Olds, 1993) was used to quantify the components of the variance of protein abundance using adjusted sums of squares with higher level terms omitted (Hector et al., 2010) together with nested series of sequential models (Nelder & Lane, 1995). A specific method was applied to include the effects of the interactions on each part of the variance attributable to a given factor. The adjusted main effects were estimated by comparing the complete linear model (Model 1) with the corresponding nested models (Models 2–5). In these nested models, Model 2 included only the effect of the protein quantity, Model 3 only the effect of the chemical steps (equalization and purification), and Model 4 both effects without interaction between them. Model 5 was used to separate the effects of the two chemical steps.

Table 2 presents the components of the analysis of variance. The adjusted sum of squares was calculated: (i) for protein quantity and its interaction with the chemical steps through the difference between Model 3 and Model 1 residual sums of squares (RSS); (ii) for the chemical steps and their interactions with the protein quantity through the difference between Model 2 and Model 1 RSS; (iii) for the interaction between the protein quantity and the chemical steps through the difference between Model 4 and Model 1 RSS. The same process was applied to separate the two components of the chemical source of variance.

| **Source of variation** | **DF** | **Adjusted sum of squares** |
|---|---|---|
| Biology + interaction | JK | *SS(x + x(E + P(E))\|E, P(E))= RSS(Model 3) – RSS(Model 1)* |
| Chemistry + interactions | 2(JK-1) | *SS(E + P(E) + x(E + P(E))\|x)= RSS(Model 2) – RSS(Model 1)* |
| Equalization + interaction | 2(J-1) | *SS(E + xE\|x, P(E), xP) = RSS(Model 2) – RSS(Model 5)* |
| Purification + interaction | 2J(K-1) | *SS(P(E) + xP(E)\|x, E, xE) = RSS(Model 5) – RSS(Model 1)* |
| Interaction | (JK-1) | *SS(x(E + P(E))\|x, E, P(E)) = RSS(Model 4) – RSS(Model 1)* |
| Residual | IJKR-2JK | *RSS(Model 1)* = $\sum_{ijkr}\left(y_{ijkr} - \hat{y}_{ijkr}\right)^2$ |
| Measurement error | (R-1)IJK | $\sum_{ijkr}\left(y_{ijkr} - \bar{y}_{ijk.}\right)^2$ |
| Lack of fit | IJK-2JK | $\sum_{ijkr}\left(\hat{y}_{ijkr} - \bar{y}_{ijk.}\right)^2$ |

DF, degrees of freedom (balanced design); I, number of samples; J, number of equalizations; K, number of purifications per equalization; R, number of replicates; $-\bar{y}_{ijk.}$ , the means of replicate measurements for each sample and each purification; $-\hat{y}_{ijkr}$; the predicted measurements. RSS, residual sums of squares, SS, sums of squares.

Table 2. Components of the analysis of variance

The residual variance was split into two components (Weisberg, 2005): (i) the measurement error, calculated as the sum of the squares of the differences between the spot values and their means. It represents the remaining technological error when the chemical steps are identical and depends on the effect of the spotting step, the instrumental error, and the error of the algorithm in estimating the peak intensity; (ii) the lack of fit to the model.

Each variance component was expressed as percentage of the total variance and used as criterion to compare the algorithms.

The term that includes the mean intercept and the mean slope, $\beta_0 + \beta_1 x$, represents the linear relationship between observed peak intensities and protein quantities (on the log2-log2 scale), which leads to the biological variance generated by dilution. The biological variance of the peak intensities is expected to be proportional to the squared slope coefficient and to vary with the equalization and purification replicates. The biological variance was defined as the variance explained by protein abundance in the dilution series and its interaction with the chemical steps.

Some components of the technological variability are: (i) the variability of the intercept and slopes around their general means due to the chemical steps; that is an equalization effect, $\beta_{0j} E_j + \beta_1 E_j x$ and a purification effect within equalization, $\beta_{0,k(j)} P_k(E_j) + \beta_{1,k(j)} P_k(E_j)x$; this leads to the part of variance attributable to the chemical steps (chemical variance); and, (ii) the measurement error due to the effect of the spotting step, the instrumental error (Mass Spectrometer), and the algorithmic error (peak intensity estimation).

The total technological variance was defined as the variance explained by the chemical replicates, the variance of its interaction with the biological effect, and the measurement error.

The statistical analyses were implemented in R packages *stats*, *alr3*, and *lme4* (freely available from CRAN, https://cran.r-project.org/web/packages/). R scripts of all analyses are given in the Supplementary Material.

# 3 RESULTS AND DISCUSSION

## 3.1 Linearity analysis

Figure 2 shows that there were trends toward a linear relationship between protein quantities and observed peak intensities.

For a given protein quantity, the dispersion of the values along the vertical axis varied between peaks but was greater with the classical algorithm (first row of Fig. 2) than with both BHI-PRO algorithms (rows 2 and 3). The classical algorithm had thus a greater variability of slopes. This variability depended on the equalization replicate (the slopes of equalizations 1 and 2 were more homogeneous than the slopes of other equalizations whatever the peak position). Algorithm BHI-PRO 1 had the lowest variability along the mean regression straight line whatever the peak, thus the smallest dispersion of straight lines.

Table 3 shows slope values close to 1. These slopes varied with the equalization and the mean slope varied between peaks. For peak 3150 Da, the slopes were greater with BHI-PRO algorithms than with the classical algorithm (the 95% confidence intervals did not overlap). The overall slope of the relationship between protein abundances and peak intensities was not equal to 1 on the log2-log2 scale (the 95% CI did not include value 1) in most cases and varied between peaks. This means that there is a polynomial relationship between the peak area and protein abundance and that the value of exponent $\beta_1$ is peak-specific. However, the confidence interval computed for a particular sample does not necessarily include the true value of the parameter because the observed data are random samples from a true population. The intercept and slope variabilities around the mean linear relationship, as evidenced by the dispersion of the intensities around the mean regression line, was interpreted as a technological variability (precisely, chemical variability plus its interaction with biological variability). The two chemical steps (equalization and purification) affected the relative protein abundance in the samples.

| **Algorithm and peak (*m/z*)** | **Mean slope (95% CI)** | **Slopes by equalization (E1–E4)** | | | |
|---|---|---|---|---|---|
| | | **E1** | **E2** | **E3** | **E4** |
| Classical | | | | | |
| 1349 | 0.74 (0.67–0.82) | 0.82 | 1.00 | 0.49 | 0.66 |
| 2094 | 0.69 (0.62–0.75) | 0.79 | 0.86 | 0.44 | 0.67 |
| 3150 | 1.01 (0.96–1.05) | 0.94 | 1.15 | 0.80 | 1.13 |
| 5734 | 1.49 (1.43–1.65) | 1.42 | 1.61 | 1.31 | 1.61 |
| BHI-PRO 1 | | | | | |
| 1349 | 0.84 (0.78–0.89) | 0.79 | 0.84 | 0.79 | 0.92 |
| 2094 | 0.75 (0.70–0.79) | 0.71 | 0.77 | 0.69 | 0.82 |
| 3150 | 1.19 (1.16–1.23) | 1.12 | 1.20 | 1.13 | 1.32 |
| 5734 | 1.44 (1.40–1.47) | 1.41 | 1.51 | 1.48 | 1.35 |

| | | | | | |
|---|---|---|---|---|---|
| BHI-PRO 2 | | | | | |
| 1349 | 0.80 (0.76–0.85) | 0.81 | 0.84 | 0.79 | 0.77 |
| 2094 | 1.20 (0.93–1.48) | 1.74 | 1.35 | 0.79 | 0.94 |
| 3150 | 1.17 (1.14–1.20) | 1.09 | 1.16 | 1.12 | 1.31 |
| 5734 | 1.54 (1.49–1.58) | 1.45 | 1.54 | 1.52 | 1.62 |

Table 3. Estimations of the slope coefficient in Model 1 with the three algorithms (classical, BHI-PRO 1 and BHI-PRO 2) and the four peaks (label = *m/z* position); general mean and equalization means (E1–E4)

## 3.2 Variance decomposition

Table 4 shows that the biological part of the variance varied between peaks; it was higher for peaks 3150 and 5734 Da than for other peaks whatever the algorithm. The biological part was also higher with BHI-PRO algorithms than with the classical algorithm (80% to 95% with BHI-PRO 1, 79% to 95% with BHI-PRO 2 vs. 56% to 90% with the classical algorithm). These results are coherent with the results seen with the mean slopes because the mean squared error of an estimator consists of bias squared plus variance. However, in the variance decomposition method, both sources of error are considered. This means that BHI-PRO algorithms were more efficient than the classical algorithm in protein quantification when peaks were detected. This result was expected because these algorithms were specifically developed to increase the biological part of the variance.

| **Algorithm and peak (*m/z*)** | **Theoretical abundance + interaction** | **Interaction** | **Equalization + interaction** | **Purification + interaction** | **Total** | **Measurement error*** | **Total** |
|---|---|---|---|---|---|---|---|
| Classical | | | | | | | |
| 1349 | 55.58 | 8.97 | 5.77 | 12.91 | 18.68 | 6.65 | 25.34 |
| 2094 | 59.99 | 5.92 | 6.10 | 7.23 | 13.33 | 6.65 | 19.98 |
| 3150 | 84.93 | 2.94 | 4.32 | 2.25 | 6.57 | 1.65 | 8.22 |
| 5734 | 90.00 | 2.35 | 1.11 | 2.47 | 3.58 | 2.55 | 6.13 |
| BHI PRO 1 | | | | | | | |
| 1349 | 82.29 | 0.68 | 0.43 | 2.55 | 2.98 | 6.64 | 9.61 |
| 2094 | 79.68 | 1.53 | 0.36 | 2.18 | 2.53 | 4.44 | 6.98 |
| 3150 | 94.19 | 0.91 | 0.84 | 0.87 | 1.71 | 1.24 | 2.95 |
| 5734 | 94.83 | 0.88 | 0.17 | 1.25 | 1.42 | 1.06 | 2.48 |
| 1349 | 82.29 | 0.68 | 0.43 | 2.55 | 2.98 | 6.64 | 9.61 |
| BHI PRO 2 | | | | | | | |
| 1349 | 78.86 | 0.83 | 0.31 | 3.04 | 3.35 | 8.55 | 11.9 |
| 2094 # | 79.42 | 1.14 | 0.48 | 1.71 | 2.20 | 5.09 | 7.29 |
| 3150 | 94.68 | 1.28 | 0.93 | 1.35 | 2.28 | 0.87 | 3.15 |
| 5734 | 93.54 | 1.66 | 0.26 | 2.12 | 2.38 | 1.08 | 3.46 |

Variance decomposition with the three algorithms (classical, BHI-PRO 1, and BHI-PRO 2) and the four peaks (label = *m/z* position). Variance parts are expressed in percentages of the

total variance of peak intensity. The wording "+interaction" refers to the interaction of the considered factor with biological variance. In the second column, the interaction is with chemical variance. [*]Instrument and algorithm, [†]Deviation from linearity (lack of fit), [#]Three outliers excluded.

The total is not strictly equal to 100% because of the double counting of the interactions. The total is > 100% when the effects of the factors were positively correlated and < 100% when the effects were negatively correlated.

Table 4. Variance decomposition into biological and technological according to the steps of the MALDI-TOF analytical chain

The technological part of the variance varied between peaks; it was higher for peaks 1349 and 2094 Da than for other peaks whatever the algorithm. The technological part was systematically higher with the classical algorithm than with BHI-PRO algorithms (6% to 25% with the classical algorithm vs. 2.5% to 9.6% with BHI-PRO 1 and 3.5% to 11.9% with BHI-PRO 2). The chemical part of the variance was always higher with the classical algorithm than with BHI-PRO algorithms (3.6% to 18.7% vs. less than 3.5%). The measurement error (due to the spotting step, the mass spectrometer error, and the algorithmic error in estimating the peak area) was also higher with the classical algorithm than with both BHI-PRO algorithms in three out of four peaks. This means that BHI-PRO algorithms were better than the classical algorithm in removing noise from the original signal.

BHI-PRO 1 algorithm was more efficient than the classical algorithm because it revealed a higher biological variance (up to 95% of the total variance) and a lower technological variance (< 10% of the total variance). The latter variance included both components of the chemical variance (the equalization variance was divided by 4 and the chemical variance divided by 2) as well as the measurement error. The part of modeling error (lack of fit or deviation from linearity) was divided by 2; thus, the linear estimator was less biased with BHI-PRO 1 than the classical algorithm.

In absence of outliers, BHI-PRO 1 and BHI-PRO 2 were close in performance regarding all variance components. However, we had to exclude three outliers of BHI-PRO 2 measures of one peak; otherwise, the technical part of the variance would have increased to 71% instead of 7.3%.

Replacing the null intensities by the minimum of the observed intensity values showed that BHI-PRO 1 was less performant than BHI-PRO 2 but still more performant than the classical algorithm in three peaks out of four. With BHI-PRO 1, the biological part of the variance was 58.44%, 73.48%, 81.39%, and 92.07% for peaks 1349, 2094, 3150, and 5734 Da, respectively. BHI-PRO 2 was then the most performant algorithm because null intensities were less frequent (even absent) than with BHI-PRO 1.

In comparison with the classical algorithm, the two BHI-PRO algorithms were better in reducing the technological part of the variance of the peak intensities and the modeling error when peaks were detected. When peaks are not detected, caution is required in interpreting variance components because the slope estimation may be biased.

Regarding the four peaks kept in the study, the linear model was able to explain a major part of the variance of protein relative abundance.

One limitation of the present work was the exclusion of null-value peak areas in the modeling on the log scale. As no null intensities were observed with the classical algorithm, this may be seen as favoring BHI-PRO algorithms because the latter puts undetected peaks to 0, especially BHI-PRO 1, while the former puts a nonnull value. Replacing the null intensities by the

minimum nonnull values decreased the performance of BHI-PRO algorithms when null intensities were frequent. This replacement distorts the normal distribution of the residuals and calls for a more complex model able to take into account the resulting mixed distribution (a solution would be the use of a method of normalization for censored data). However, the replacement of null intensities with BHI-PRO would force the quantification despite the assignment of value zero to noise by the LASSO penalization and would reintroduce noise.

Another limitation was the use of fixed effects models rather than mixed effects models because random effects are generalizable to other similar settings. When only a small number of replicates per factor are available, the mixed models may not converge. In the present study simulations, increasing the number of replicates to 5 per factor was not sufficient to overcome this problem. Nearly 10 replicates per factor were necessary to achieve convergence. However, when the models converged, the results of the variance components analysis were very similar to the results obtained with fixed effects models.

The impact of the signal processing steps was not studied here. A new experimental plan with replicates for each step of signal processing would allow integrating these sources of variability in the set of nested models. The set of models would necessarily change with the algorithm (or pipeline) used for signal processing.

Another interesting extension of this experiment would be to investigate the effect of introducing another type of biological variability (e.g. patient variability) on the relationship between protein relative abundance and signal intensity. This would render the experiment closer to the clinical needs.

# 4 CONCLUSIONS

For five proteins out of nine, the MALDI-TOF technology failed to restitute the biological effect given the high technological variability whatever the algorithm used for signal processing. For the four other proteins, the total technological variance reached up to 25% of the total variance of the measurements. When a biological effect is present, protein quantification can be improved at the signal processing step. Choosing the most performant signal-processing algorithm can reduce the technological variance down to about 10% of the total variance. In other words, in calculating the power of a biomarker discovery study using MALDI-TOF, the within-group variance would better be increased by about 10%, which would necessary require a larger sample size.

The present approach that includes the simultaneous development of a specific experimental design and a related model-based variance decomposition is able to improve the powers of studies that use other technologies for biomarker validation such as selected reaction monitoring (SRM).

## ACKNOWLEDGMENTS

This work was supported by the “Agence Nationale pour la Recherche” (Grant ANR 2010 BLAN 0313). The authors thank Aline Jeannin for sample processing, Pauline Salloignon for signal processing with the classical algorithm, Laurent Gerfault and Jean-Philippe Charrier for fruitful discussions regarding the BHI-PRO project, and Jean Iwaz (Hospices Civils de Lyon) for the revision of the final version of the manuscript.

## CONFLICT OF INTEREST

The authors have declared no conflict of interest.

# REFERENCES

Antoniadis, A., Bigot, J., & Lambert-Lacroix, S. (2010). Peaks detection and alignment for mass spectrometry data. Journal de la Société Françaisede Statistique,151, 17–37.

Berkson, J. (1950). Are there two regressions? Journal of the American Statistical Association, 45, 164–180.

Cairns, D. A. (2011). Statistical issues in quality control of proteomic analyses: Good experimental design and planning. Proteomics,11, 1037–1048.

Carr, S. A. (2014). Targeted peptide measurement in biology and medicine: Best practices for mass spectrometry-based assay development using afit-for-purpose approach. Molecular and Cellular Proteomics,13, 907–917.

Carroll, R. J., Ruppert, D., & Stefanski, L. A. (1995). Measurement error in nonlinear models. London: Chapman & Hall/CRC Press.

Coombes, K., Baggerly, K., Morris, J. M., Dubitzky, M. G., & Berrar, D. (Eds.) (2007). Pre-processing mass spectrometry data. Fundamentals ofdata mining in genomics and proteomics. Boston: Kluwer.

Dridi, N., Giremus, A., Giovannelli, J. F., Truntzer, C., Roy, P., Gerfault, L., ... Grangeat, P. (2014). Variable selection for noisy data applied inproteomics. Florence, Italy: ICASSP.

Gerfault, L., Klich, A., Mercier, C., Roy, P., Giovannelli, J. F., Giremus, A., ... Grangeat, P. (2014). Statistical analysis of Bayesian hierarchicalinversion for MRM protein quantification and QDA serum sample classification. Baltimore, USA: 62nd ASMS Conference on Mass Spectrometry and Allied Topics.

Gerfault, L., Szacherski, P., Giovannelli, J. F., Giremus, A., Mahé, P., Fortin, T., ... Grangeat, P. (2013). Assessing MRM protein quantification andserum sample classification performances of a Bayesian hierarchical inversion method on a colorectal cancer cohort. Saint-Malo, France: EuPA.

Grangeat, P., Giovannelli, J. F., Roy, P., Picaud, V., Truntzer, C., Lemoine, J., ... Lacroix, B. (2013). Convergence entre l'analyse biostatistique etles méthodes d'inversion hiérarchique bayésienne pour la recherche et la validation de biomarqueurs par spectrométrie de masse. Brest, France: XXIVème Colloque GRETSI.

Hector, A., von Felten, S., & Schmid, B. (2010). Analysis of variance with unbalanced data: An update for ecology and evolution. Journal of Animal Ecology,79, 308–316.

Käll, L., & Vitek, O. (2011). Computational mass spectrometry-based proteomics. PLoS Computational Biology,7, e1002277.

Langsrud, Ø. (2003). ANOVA for unbalanced data: Use type II instead of type III sums of squares. Statistics and Computing,13, 163–167.

Mazet, V., Brie, D., & Idier, J. (2004). Baseline spectrum estimation using half-quadratic minimization. In Proceedings of the European Signal Pro-cessing Conference, Vienna, Autriche, September 2004.

Mercier, C., Truntzer, C., & Pecqueur, D. (2009). Mixed-model of ANOVA for measurement reproducibility in proteomics. Journal of Proteomics,72, 974–981.

Meuleman, W., Engwegen, J. Y., Gast, M. C., Beijnen, J. H., Reinders, M. J., & Wessels, L. F. (2008). Comparison of normalisation methods for surface-enhanced laser desorption and ionisation (SELDI) time-of-flight (TOF) mass spectrometry data.BMC Bioinformatics, 9, 88.

Morháč, M. (2009). An algorithm for determination of peak regions and baseline elimination in spectroscopic data. Nuclear Instruments and Methodsin Physics Research Section A, 600, 478–487.

Morris, J. S., Coombes, K. R., Koomen, J., Baggerly, K. A., & Kobayashi, R. (2005). Feature extraction and quantification for mass spectrometry inbiomedical applications using the mean spectrum. Bioinformatics, 21, 1764–1775.

Nelder, J. A., & Lane, P. W. (1995). The computer analysis of factorial experiments: In Memoriam—Frank Yates. The American Statistician, 49, 382–385.

Piepho, H. P., Büchse, A., & Emrich, K. (2003). A Hitchhiker's guide to mixed models for randomized experiments. Journal of Agronomy and Crop Science, 189, 310–322.

Renard, B. Y., Kirchner, M., Steen, H., Steen, J. A., & Hamprecht, F. A. (2008). NITPICK: Peak identification for mass spectrometry data. BMC Bioinformatics, 9, 355.

Roy, P., Truntzer, C., Maucourt-Boulch, D., Jouve, T., & Molinari, N. (2011). Protein mass spectra data analysis for clinical biomarker discovery: A global review. Briefings in Bioinformatics, 12, 176–186.

Shaw, R. G., & Mitchell-Olds, T. (1993). ANOVA for unbalanced data: An overview. Ecology, 74, 1638–1645.

Szacherski, P., Gerfault, L., Giovannelli, J. F., Giremus, A., Mahé, P., Fortin, T., ... Grangeat, P. (2013). MRM protein quantification and serum sample classification. 61st ASMS Conference on Mass Spectrometry and Allied Topics; Minneapolis, USA.

Szacherski, P., Giovannelli, J. F., Gerfault, L., Mahé, P., Charrier, J. P., Giremus, A., ... Grangeat, P. (2014). Classification of proteomic MS data as Bayesian solution of an inverse problem. IEEE Access, 2, 1248–1262.

Szacherski, P., Giovannelli, J. F., Giremus, A., & Grangeat, P. (2012). Robust MS serum sample classification in proteomics by the use of inverse problems. IEEE International Workshop on Genomic Signal Processing and Statistics; Washington, USA.

Weisberg, S. (2005). Weights, lack of fit, and more. In: Applied Linear Regression (3rd ed., pp. 96–114). New York: Wiley.

WHO. (2001). International programme on chemical safety biomarkers in risk assessment: Validity and validation. Retrieved from https://www.inchem.org/documents/ehc/ehc/ehc222.htm

Yang, C., He, Z., & Yu, W. (2009). Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. BMC Bioinformatics, 10, 4.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society. Series B Statistical Methodology, 68, 49–67.

Zhang, Z. M., Chen, S., & Liang, Y. Z. (2010). Baseline correction using adaptive iteratively reweighted penalized least squares. Analyst, 135, 1138–1146.