Identification of Analytical Factors Affecting Complex Proteomics Profiles Acquired in a Factorial Design Study with Analysis of Variance: Simultaneous Component Analysis

Vikram Mitra, Natalia Govorukhina, Goitzen Zwanenburg, Huub Hoefsloot, Inge Westra, Age Smilde, Theo Reijmers, Ate G. J. van der Zee, Frank Suits, Rainer Bischoff, and Péter Horvátovics

Analytical Biochemistry, Department of Pharmacy, University of Groningen, A. Deusinglaan 1, 9713 AV Groningen, The Netherlands
Swammerdam Institute for Life Science, University of Amsterdam, The Netherlands, Science Park 904, 1098 XH Amsterdam, The Netherlands
Analytical BioSciences, Leiden University, Einsteinweg 55, 2333 CC Leiden, The Netherlands
Department of Gynecology, University Medical Centre Groningen, Hanzeplein 1, 9713 GZ Groningen, The Netherlands
IBM T.J. Watson Research Centre, 1101 Kitchawan Road, Yorktown Heights, New York 10598, United States
Netherlands Bioinformatics Centre, Geert Grooteplein 28, 6525 GA Nijmegen, The Netherlands
Netherlands Proteomics Centre, Padualaan 8, 3584 CH Utrecht, The Netherlands

ABSTRACT: Complex shotgun proteomics peptide profiles obtained in quantitative differential protein expression studies, such as in biomarker discovery, may be affected by multiple experimental factors. These preanalytical factors may affect the measured protein abundances which in turn influence the outcome of the associated statistical analysis and validation. It is therefore important to determine which factors influence the abundance of peptides in a complex proteomics experiment and to identify those peptides that are most influenced by these factors. In the current study we analyzed depleted human serum samples to evaluate experimental factors that may influence the resulting peptide profile such as the residence time in the autosampler at 4 °C, stopping or not stopping the trypsin digestion with acid, the type of blood collection tube, hemolysis levels, differing the results of the ASCA model. Permutation tests were used to identify which of the preanalytical factors in determining the results of the ASCA model. The design required analysis of 16 samples in which the main effects were not confounded by two-factor interactions. Data preprocessing using the Threshold Avoiding Proteomics Pipeline (Suits, F.; Hoekman, B.; Rosenling, T.; Bischoff, R.; Horvátovics, P. Anal. Chem. 2011, 83, 7786–7794, ref 1) produced a data-matrix containing quantitative information on 2559 peaks. The intensity of the peaks was log-transformed, and peaks having intensities of a low t-test significance (p-value > 0.05) and a low absolute fold ratio (<2) between the two levels of each factor were removed. The remaining peaks were subjected to analysis of variance (ANOVA)-simultaneous component analysis (ASCA).2 Permutation tests were used to identify which of the preanalytical factors influenced the abundance of the measured peptides most significantly. The most important preanalytical factors affecting peptide intensity were (1) the hemolysis level, (2) stopping trypsin digestion with acid, and (3) the trypsin/protein ratio. This provides guidelines for the experimentalist to keep the ratio of trypsin/protein constant and to control the trypsin reaction by stopping it with acid at an accurately set pH. The hemolysis level cannot be controlled tightly as it depends on the status of a patient’s blood (e.g., red blood cells are more fragile in patients undergoing chemotherapy) and the care with which blood was sampled (e.g., by avoiding shear stress). However, its level can be determined with a simple UV spectrophotometric measurement and samples with extreme levels or the peaks affected by hemolysis can be discarded from further analysis. The loadings of the ASCA model led to peptide peaks that were most affected by a given factor, for example, to hemoglobin-derived peptides in the case of the hemolysis level. Peak intensity differences for these peptides were assessed by means of extracted ion chromatograms confirming the results of the ASCA model.

Differential protein expression analysis using high-throughput liquid chromatography–mass spectrometry (LC–MS(/MS)) profiling platforms, such as in biomarker

**Supporting Information**

© 2016 American Chemical Society
discovery and validation, is a major task in proteomics. Human blood is an easily accessible body fluid that transports biomolecules related to the immune defense, signaling, and cell and tissue leakage and is therefore an important biological matrix for biomarker discovery and validation. Serum obtained from blood after clotting and centrifugation or filtration is a highly complex fluid in terms of protein composition. A total of 1924 proteins were identified in an analysis combining multiple data sets in PeptideAtlas and it is estimated that more than 10,000 proteins may be present in serum with a dynamic concentration range spanning 11 orders of magnitude. Analytical chemists responsible for sample preparation and analysis, e.g., by LC−MS/(MS) of serum, must strive to minimize variation introduced by preanalytical steps. Variation introduced by differences in analytical parameters may influence the quality of the data and thereby create a biased statistical outcome. The complexity of sample preparation methods in proteomics, the large number of detected analytes and the varying experimental parameters increase the risk of identifying false positive biomarker candidates or, alternatively, missing potentially relevant ones. It is therefore important to determine the effect of preanalytical, experimental factors on the measured peptide profiles. Identifying the sources of experimental variation allows better control of sample quality and minimizes false discoveries by ensuring minimal technical variation within a proteomics study. Also, samples that do not meet the quality criteria for vital noncontrollable experimental factors can be excluded from further analysis.

Experimental design is a well-established method that is used in many research fields and industrial processes to reveal the effects of experimental factors on measured response variables. Currently there is no study design to assess the effect of preanalytical factors on complex proteomics profiles in a comparable and objective manner, while this is commonplace in single-molecule, bioanalytical studies and even part of guidelines for regulated bioanalysis. We address this issue using a multivariate statistical approach to evaluate the effect of individual experimental factors and the interactions between two or multiple factors on complex peptide profiles. Since proteomics analyses are time-consuming, we opted for a fractional factorial design to reduce the number of analyses per study, at the expense of confounding potential interactions between factors thus focusing on the main effects only.

LC−MS/(MS) data from digested, complex proteomics samples contain quantitative information on thousands of peptides requiring multivariate statistical approaches. Additionally, to interpret the data matrix that arises from an experimental design study, a statistical method that incorporates the structure of the data set is required. Multivariate extensions of analysis of variance (ANOVA) (MANOVA) break down for large numbers of variables. ANOVA−simultaneous component analysis (ASCA) is a technique that can handle large numbers of variables and includes all aspects of an experimental design into the statistical analysis of the data.

Collecting multivariate data from an experimental design is not common in proteomics studies but is more frequently used in metabolomics and genetics studies. For example, Scholten et al. describe a statistical linear model of microarray data measured with an experimental design. Johnson et al. used principle component analysis (PCA) followed by multivariate analysis of variance of the resulting principal components for a multivariate evaluation of complex metabolomics data sets. In their study, the direct injection mass spectrometry and Fourier transform-infrared data from metabolite profiling experiments of wild type and an ethylene signaling mutant of Arabidopsis was assessed with the PCA-MANOVA approach. This approach identified multiple significant main effects and interactions. Canonical variate analysis was used to investigate further the biological meaning of the findings. In a proteomics study by Szalowska et al., a fractional factorial design including 32 experiments was applied to reveal the effect of different preanalytical factors on surface enhanced laser desorption ionization (SELDI) spectrum quality of the secretome of human visceral adipose tissue. The spectrum quality was assessed using univariate ANOVA by four scores, the total number of detected peaks, the average signal-to-noise ratio, peak shape, and peak resolution. Andrews et al. used full and fractional factorial designs to optimize parameters of LC−tandem mass spectrometry (MS/MS) analyses with the goal to improve sequence coverage of Saccharomyces cerevisiae. This study evaluated the number of identified protein groups, unique peptides and spectral counts applying univariate ANOVA on half normal quantile probability plots to identify factors that have a significant influence on these quality measures. Harrington et al. in two studies applied ANOVA principal component analysis to correct for the variance of confounding factors on the factors of interest. In their MALDI-MS discovery study, protein biomarkers from amniotic fluid were identified that can be used to detect intra-amniotic infections, which is one of the main causes of premature delivery. The main difference between ANOVA-PCA and ASCA, as used in this study, is that ANOVA-PCA includes the residual variance in the PCA step whereas ASCA looks solely at the variance due to factors and interaction terms. However, until now no study reported a multivariate analysis of complex proteomics data obtained in an experimental design with the goal to identify the most influential factors and affected peaks on the entire peptide profiles. The multivariate approach is essential since many compound-related signals correlate, such as different charge states of the same peptide, different uniquely mapping peptides from the same protein, and proteins whose expression is coregulated.

This study presents statistical methodology that has the following three aims: (1) assess the performance of ASCA analysis with respect to the number of compounds versus sample size using peaks obtained after filtering with Volcano plot parameters (t-test significance and fold ratio change), (2) to determine factors that affect the measured peptide profile in an experimental design study for complex proteomics samples, and (3) to identify the peaks that are most affected by these factors with their effect size. First, a simulated data set is used to assess the performance of ASCA analysis where significant and nonsignificant factors as well the peaks affected and nonaffected by factors are known. This setup allows assessment of the effect of peak selection based on Volcano plot parameters on ASCA performance. Finally ASCA analysis of an experimental design data set is shown for a data set that was obtained with a human serum sample and included seven factors at two levels with a 27 design at resolution IV.

## METHODS

### Experimental Design

In this study of human serum samples that were depleted of the six most abundant proteins, we identified seven factors that could have an important effect on the peptide profiles: (1) type of blood collection tube, (2) clotting time, (3) hemolysis level, (4) trypsin/protein ratio, (5)
stopping or not stopping trypsin digestion with acid, (6) number of freeze–thaw cycles, and (7) residence time of the digest in the autosampler at 4 °C. Each of these factors was analyzed at two levels covering the expected range of conditions.

A full factorial design with seven factors varied at two levels would require at least $2^7 = 128$ analyses and would give information on the main effect of each factor and on all interactions between the different factors from the second until the seventh order. In order to limit sample analysis time and cost, a fractional factorial design with resolution IV (2IV design) was performed. This design requires 16 LC–MS analyses and gives sufficient resolution to screen for the main effects, since these effects are not confounded by any other main effects or by any two-factor interactions. The design follows the general concept of the experimental design strategy, and we have used the software MODDE (version 7.0.0.1) to determine the level distribution for all 7 factors for the minimal set of 16 samples. Table 1 lists the distribution of the levels for each of the 7 factors for the 16 experiments used in this study. The exact definitions of the two levels, low and high are indicated for each factor.

Figure 1 shows the workflow that illustrates the main steps in the analysis of the experimental design single stage-LC–MS data. The work presented in this article includes wet lab and bioinformatics data analysis steps. Page 3 of the Supporting Information contains a detailed description of the three main steps that need bioinformatics intervention.

**Blood Sample Collection.** Blood was collected from a healthy volunteer (male) and serum obtained from the University Medical Center Groningen (UMCG, The Netherlands). Serum was stored at −80 °C in aliquots until analysis. At the UMCG, all blood donors are routinely asked to give their blood samples in a biobank for future research. Relevant data are retrieved and transferred into an anonymous, password-protected database. The personal identity is protected by a study specific, unique code, and the true identity is only known to two dedicated data managers. According to Dutch regulations, these precautions mean that no further institutional review board approval is needed (http://www.federa.org).

**Sample Preparation and Description of Experimental Factors. Blood Collection Tubes (Factor 1 "Blood Collection Tube").** Two kinds of tubes, which were both in use at the UMCG to establish the serum biobank, were evaluated for blood collection: BD368430 (low level, a "red stopper clotting tube", which is a glass tube with a siliconized inner wall to avoid retention of red blood cells on the walls of the tube) and BD367784 (high level, a "gel tube", which is a glass tube with a separation gel and micronized silica to accelerate clotting). During centrifugation, the polymer gel moves up the inner wall of this tube forming a barrier between the supernatant (serum) and sediment (blood clot and cells).

**Clotting Time (Factor 2 "Clotting Time").** Blood samples were allowed to clot for 2 (low level) or 6 (high level) hours at room temperature prior to centrifugation to obtain serum.

**Level of Hemolysis (Factor 3 "Hemolysis").** To simulate an elevated (high) level of hemolysis, a lysate of red blood cells was added to the serum prior to depletion. Hemoglobin is not removed by the Multiple Affinity Removal column (Agilent) used to deplete the six most abundant proteins. Red blood cells were collected according to the following protocol: 0.5 mL of lysis buffer (NH₄Cl 155 mmol/L, EDTA 0.1 mmol/L) was added to 0.5 mL of fresh blood and centrifuged for 20 min at 2 000 rpm. A volume of 4 mL of lysis buffer was added to the pellet and incubated overnight at 4 °C. The next day the lysate was filtered through spin filters (0.22 μm; no. 5185-5990, Agilent) at 13 000 rpm. Aliquots of the filtrate were stored at −80 °C.

The amount of lysed red blood cells that should be added to serum to mimic an increased (high) level of hemolysis was determined by the addition of different volumes (1, 3, 5, 7, and 10 μL) of red blood cell lysate to 20 μL of serum that was immediately diluted with ice-cold water to a total volume of 60 μL followed by centrifugation at 13 000 rpm for 30 min at 4 °C. Another 15 μL of ice-cold water was added to the supernatant and the absorbance was measured at 340, 380, 415, and 450 nm (BioWave S2100 UV–vis diode array spectrophotometer (Biochrom Ltd., Cambridge, U.K.)). A calibration line with respect to hemoglobin (Hb, Hemoglobin human, Sigma, no. 9008-02–0) was obtained using the following formula: Hb [g/L] = (167.2 × A415 − 83.6 × A340/380 − 83.6 × Hb [g/L] = (167.2 × A415 − 83.6 × A340/380 − 83.6 × A340/380 − 83.6 × A380)

### Table 1. Main Parameters of the Levels Used in the Fractional Factorial Design

<table>
<thead>
<tr>
<th>sample name</th>
<th>blood collection tube</th>
<th>hemolysis</th>
<th>clotting time</th>
<th>freeze–thaw cycles</th>
<th>trypsin digestion</th>
<th>stopping trypsin</th>
<th>sample stability</th>
</tr>
</thead>
<tbody>
<tr>
<td>16090534</td>
<td>BD368430</td>
<td>low</td>
<td>2 h</td>
<td>1 cycle</td>
<td>6%</td>
<td>yes</td>
<td>0 days</td>
</tr>
<tr>
<td>16090526</td>
<td>BD367784</td>
<td>low</td>
<td>2 h</td>
<td>1 cycle</td>
<td>11%</td>
<td>yes</td>
<td>30 days</td>
</tr>
<tr>
<td>16090532</td>
<td>BD368430</td>
<td>high</td>
<td>2 h</td>
<td>1 cycle</td>
<td>11%</td>
<td>no</td>
<td>0 days</td>
</tr>
<tr>
<td>16090540</td>
<td>BD367784</td>
<td>high</td>
<td>2 h</td>
<td>1 cycle</td>
<td>6%</td>
<td>no</td>
<td>30 days</td>
</tr>
<tr>
<td>16090541</td>
<td>BD368430</td>
<td>low</td>
<td>6 h</td>
<td>1 cycle</td>
<td>11%</td>
<td>no</td>
<td>30 days</td>
</tr>
<tr>
<td>16090525</td>
<td>BD367784</td>
<td>low</td>
<td>6 h</td>
<td>1 cycle</td>
<td>6%</td>
<td>no</td>
<td>0 days</td>
</tr>
<tr>
<td>16090533</td>
<td>BD368430</td>
<td>high</td>
<td>6 h</td>
<td>1 cycle</td>
<td>6%</td>
<td>yes</td>
<td>30 days</td>
</tr>
<tr>
<td>16090539</td>
<td>BD367784</td>
<td>high</td>
<td>6 h</td>
<td>1 cycle</td>
<td>11%</td>
<td>yes</td>
<td>0 days</td>
</tr>
<tr>
<td>16090531</td>
<td>BD368430</td>
<td>low</td>
<td>2 h</td>
<td>3 cycles</td>
<td>6%</td>
<td>no</td>
<td>30 days</td>
</tr>
<tr>
<td>16090538</td>
<td>BD367784</td>
<td>low</td>
<td>2 h</td>
<td>3 cycles</td>
<td>11%</td>
<td>no</td>
<td>0 days</td>
</tr>
<tr>
<td>16090536</td>
<td>BD368430</td>
<td>high</td>
<td>2 h</td>
<td>3 cycles</td>
<td>11%</td>
<td>yes</td>
<td>30 days</td>
</tr>
<tr>
<td>16090530</td>
<td>BD367784</td>
<td>high</td>
<td>2 h</td>
<td>3 cycles</td>
<td>6%</td>
<td>yes</td>
<td>0 days</td>
</tr>
<tr>
<td>16090535</td>
<td>BD368430</td>
<td>low</td>
<td>6 h</td>
<td>3 cycles</td>
<td>6%</td>
<td>yes</td>
<td>0 days</td>
</tr>
<tr>
<td>16090529</td>
<td>BD367784</td>
<td>low</td>
<td>6 h</td>
<td>3 cycles</td>
<td>6%</td>
<td>yes</td>
<td>0 days</td>
</tr>
<tr>
<td>16090528</td>
<td>BD368430</td>
<td>high</td>
<td>6 h</td>
<td>3 cycles</td>
<td>11%</td>
<td>no</td>
<td>30 days</td>
</tr>
<tr>
<td>16090542</td>
<td>BD367784</td>
<td>high</td>
<td>6 h</td>
<td>3 cycles</td>
<td>11%</td>
<td>no</td>
<td>30 days</td>
</tr>
</tbody>
</table>

*Each of the seven factors (shown in columns 2–8) was varied at two levels. The table gives the names of the 16 samples (first column) and the level name for each of the factors at which the sample was measured.
The absorbance of the used serum sample from the biobank was measured and the amount (in μL) of red blood cells lysate necessary to be added to reach a high level of 89 g/L Hb was calculated from the calibration curve (high level). This level of hemolysis is observed in cancer patients receiving chemotherapy and was obtained with addition of 4 μL of red blood cell lysate (containing 6.68 μg of Hb) to 75 μL of serum. The original serum sample was used as the low level of hemolysis.

Depletion of High-Abundance Serum Proteins. A volume of 80 μL (80% of the total amount (20 μL of crude serum mixed with 80 μL of buffer A (Agilent))) of diluted crude serum was injected on a Multiple Affinity Removal column (Agilent, 4.6 mm × 50 mm, no. S185-5984) after filtration through 0.22 μm spin filters (no. S185-5990) at 13 000g and 4 °C for 10 min to remove particulates. The removal of the six most abundant proteins was performed on a LaChrom HPLC System (Merck Hitachi) with detection at 280 nm using the following timetable: 0−9 min, 100% buffer A (0.25 mL/min); 9.0−9.1 min, linear gradient 0−100 B % (1 mL/min), 9.1−12.5 min, 100% buffer B (1 mL/min); 12.5−12.6 min, linear gradient 100−0% buffer B (1 mL/min); 12.6−20 min, 100% buffer A (1 mL/min). The flow-through fraction (depleted serum collected between 2 and 6 min, total volume of ∼1 mL) was collected.19

Protein concentrations were determined with the Micro BCA Protein assay reagent kit (Pierce) and calculated for an average protein molecular weight of 50 kDa. Bovine serum albumin was used as the calibration standard.

Figure 1. General scheme of the experimental analysis and statistical evaluation including sample preparation, LC−MS analysis, data preprocessing, and ASCA analysis. Proteins in biological samples were digested to peptides, which are in this example influenced by seven factors each at two levels. This affects the peptide profile obtained with LC−MS as shown by the extracted ion chromatogram (EIC) of one sample as an example (upper right panel). LC−MS raw data were preprocessed with the TAPP pipeline.1 The natural logarithm intensity values from the data matrix were then filtered using a fold change of 2 and a t-test significance of 0.05 as shown in a Volcano plot. These thresholds were determined based on simulation analysis. The resulting data matrix was then submitted to ASCA analysis, which identified the significant factors and provided a ranked list of discriminatory peaks as shown by the extracted ion chromatogram and box plots (green traces correspond to high, blue traces to low level of a factor). Peptide identity was annotated based on the analysis on a QTOF LC−MS/MS platform.
Digestion of Serum Samples (Factor 4 “Trypsin Digestion”). Trypsin (sequencing grade modified trypsin, Promega, no. VS111 in ratios 1:20 (high level) or 1:100 (low level) wt/wt (enzyme to total protein in depleted serum) was used for digestion at 37 °C at 450 rpm overnight (Eppendorf thermomixer).

Stopping Trypsin Digestion (Factor 5 “Stopping Trypsin”). To stop the reaction with trypsin for the high factor level, formic acid was added after overnight digestion to reach a final concentration of 0.5% (v/v). For samples having the low level of this factor, this step was left out.

Freeze–Thaw Cycles (Factor 6 “Freeze–Thaw Cycle”). The two levels of this factor consisted of one (low level) or three (high level) freeze–thaw cycles (80 °C/room temperature), respectively, in which one freeze–thaw cycle was indispensable during sample collection, storage, and the analysis procedure. Aliquots until the analysis were stored at −80 °C.

Sample Stability in the Autosampler (Factor 7 “Sample Stability”). Stability of the trypsin-digested serum samples was evaluated by keeping them for 30 days at 4 °C in the autosampler (high level) or by injecting them directly (low level) after thawing.

LC–MS Analysis. LC–MS Iontrap. All LC–MS analyses were performed on an Agilent 1100 capillary HPLC system coupled online to an SL iontrap mass spectrometer (Santa Clara, CA) equipped with an Atlantis dC 18 column (1.0 mm × 150 mm, 3 μm, Agilent Technology) that was protected by an Atlantis dC 18 in-line trap column (3 μm, 2.1 mm × 20 mm guard column, Agilent Technology). A volume of 40 μL of the pretreated (depleted and digested) fractions corresponding to guard column, Agilent Technology). A volume of 40 μL of the pretreated (depleted and digested) fractions corresponding to ~8 μg. The autosampler (catalog no. G1367A) was equipped with a 100 μL injection loop and a temperature-controlled cooler (catalog no. G1330A) maintaining the samples at 4 °C. The HPLC system had the following additional components: capillary pump (catalog no. G1376A), solvent degasser (catalog no. G1379A), UV detector (catalog no. G1314A), and column holder (catalog no. G1316A). The sample was injected and washed in the back-flush mode for 30 min (0.1% aqueous formic acid (FA, 98–100%, pro analysis, Merck, Darmstadt, Germany) and 3% acetonitrile (ACN, HPLC-S grade, Biosolve, Valkenswaard, The Netherlands) in ultrapure water (resistance 18.2 Ω·cm obtained with a Sartorius Stedim purification system, Nieuwegein, The Netherlands) at a flow rate of 50 μL/min). Peptides were eluted in a linear gradient from 0 to 70% (0.5%/min) ACN containing 0.1% FA at a flow rate of 20 μL/min. After each injection, the in-line trap and the analytical column were washed with eluent A (H2O/ACN/FA; % of 49:50:1) for 20 min prior to the next injection. The injection order of the samples listed in Table 1 was randomized. The following settings were used for mass spectrometry during LC–MS. Nebulizer gas, 16.0 psi N2; drying gas, 6.0 L/min N2; skimmer, 40.0 V; ionization voltage, 3500 V; cap. exit, 158.5 V; Oct. 1, 12.0 V; Oct. 2, 2.48 V; Oct. rf, 150 Vpp (Voltage, Peak Power Point); Lens 1, −5.0 V; Lens 2, −60.0 V; Trap drive, 52.9; T, 274 °C; multiplier voltage, 1938 V; ionization mode, positive; scan resolution, enhanced (5 500 m/z per second scan speed) and scanning mode was single stage MS acquisition only except for two samples, which were analyzed in data dependent LC–MS/MS mod. Target mass, 600; scan range, 100–1500 m/z. Average number of spectra, 2. Spectra were saved in centroid mode. LC–MS chromatographic data were acquired with Bruker Data Analysis software, version 2.1 (Build 37). In MS/MS mode, the following parameters were used: fragmentation time, 40 000 μs; fragmentation width, 10.00 m/z. MS/MS acquisition parameters were Isol Coars High Mass < 2 090: 150; Isol Fine High. Mass, 70; Isol Fine Low. Mass < 2 090: 200; Isol Coarse High, Ext mass > 2 090: 200; Isol Fine High. Ext mass < 2 090: 200; Isol Fine Low. Ext mass < 2 090: 200; Isolation delay, 0 μs.

Description of data preprocessing and LC–MS/MS analysis using Agilent QTOF instrument of human serum sample used for peak annotation and details of annotation transfer can be found in pages 1–3 in the Supporting Information.

ASCA Analysis. ASCA is a multivariate technique to analyze a data matrix, X, from an experimental design. Each row in the data matrix represents a measurement of multiple variables that form the columns of the data matrix. When using ASCA, the data matrix is decomposed using ANOVA style in a matrix, M, containing the column averages of X, main effect matrices Aij for each experimental factor ki, two factor interaction matrices that describe Aij and describe the interaction between factors k and j, higher interaction matrices if appropriate, and a residual matrix, E, that contains the variation not explained by the main effects and interactions. In this paper we only consider the main effects, hence, the data matrix is written as the sum of the overall means, the main effect matrices Aij and the matrix E with residuals:

\[ X = M + \sum_k A_{ij} + E \]

With ASCA, a principal component analysis (PCA) is performed on each of the effect matrices Aij. When a factor m has two factor levels, PCA on the effect matrix Aij will give a single principal component indicating the difference between the two factor levels. The size of the principal component of factor k, measured by the sum of squares (SSQ) of the elements of the effect matrix Aij gives the variance that is explained by factor k. By comparing the SSQ of the different factors we can order the experimental factors according to the variance they explain. The loadings of the principal component give the contribution of each of the variables to the factor effect; the larger the loading of a variable, the larger the difference of its measured values between the two factor levels.

To assess the significance of the difference in factor level of factor k, we use a permutation test in which the SSQ of Aij is used as a test statistic. For each factor, k, the class labels of the levels of the remaining factors were randomly permuted 10 000 times. The p-value for factor k was determined as the fraction of the permutations with a SSQ larger than the SSQ of the effect matrix Aij.

Structure of the Data Set. The peptide data matrix, X has n = 16 rows (the number of measured samples, see Table 1) and V = 2 559 columns (the single stage matched peaks). The matched peaks show a large variation in intensity; therefore, the peak intensities were natural log-scaled before the ASCA analysis. Missing values in X contain 0 values, which were replaced by noise sampled from a normal distribution of \( N(\mu = 6.042, \sigma = 0.5391) \) (see section 1.2 “Data pre-processing
and quantification" in the Supporting Information for origin of the distribution). To avoid bias due to one particular noise realization, all ASCA analysis was repeated for 100 different noise realizations. In this study we assessed the performance of ASCA using data set X obtained with statistical simulation where all significant factors and affected peaks are known followed by analysis of data set obtained from an experimental fractional factorial design described above (see the detailed description in the Materials and Methods section 1.4 "Parameters of the simulated dataset" in the Supporting Information).

The source code of the ASCA analysis and of the simulation including raw LC−MS files and TAPP1 preprocessed data is available at https://github.com/vikrammitra/ASCA.

### RESULTS AND DISCUSSION

**Method Validation Using Statistical Simulation. ASCA Analysis of Simulated Data.** Proteomics molecular profiling data contains quantitative information on several thousands of compounds measured in low number of samples. ASCA is a multivariate statistical approach that provides the possibility to evaluate if factor is significant taking all compounds into account. However, we expect that the high ratio of number of measured compounds versus sample size hamper the performance of ASCA analysis. In order to study the relation of ASCA performance with the respect to the number of variables versus sample size, we performed simulation study where all significant factors and affected peaks are known using the same setup that was obtained for the experimental design, i.e., using the same number of variables and factors, the same noise level and the number of significant factors, and affected peaks that we estimate to be reasonable in the analyzed data set.

Figure 2. Bar plots illustrating the SSQ (blue plots a, b, c, and g) and the significance of SSQs obtained with 10 000 permutations (red plots d, e, f, and h) of main effects using ASCA analysis in simulated data sets (plots a−f) and in the data set obtained with the experimental design (plots g and h). Bar plots a and d were obtained with ASCA analysis using all peaks, bar plots b, e, g, and h were obtained after union of selected peaks having a fold change of 2 and a t-test significance of 0.05 for each of the factors, while bar plots c and f were obtained with the union of selected peaks with fold change of 4 and a t-test significance of 0.01. The blue dashed line shows the threshold of 0.05 for SSQ’s significance (red plots d−f and h). The plots obtained with simulation show that peak selection with moderate threshold (plots b and e) improve slightly the performance of ASCA compared to no peak selection (left panels, plots a and d), while the peak list obtained with more extreme thresholds has a decreased performance (plots c and f).
included in the simulation our assumption that not all factors are significant and factors may influence the peptide profile differently with respect of effect size and number of affected peptides. Therefore, for the simulation study we used seven preanalytical factors from which three (factors 1, 3 and 5) were constructed to have a significant effect on 5%, 3.5%, and 5% of randomly chosen peaks with a mean difference in peak intensity between the two factor levels of 3, 4 and 6, respectively. The matrix was filled with random variables from the noise distribution found in the experimental fractional factorial design data set of \( N(\mu = 6.042, \sigma = 0.539) \) for peaks not affected by factors. These parameters are much similar to the real experimental design data set. We did permutation tests on 100 different realizations of the simulated data matrix and found that on average at the \( \alpha = 0.05 \) significance level factors 3 and 5 were identified as significant; however, factor 1 was not found to have a significant effect at the \( \alpha = 0.05 \) level (Figure 2d) but had a \( p \)-value close to the significance threshold.

The failure to identify factor 1 as a significant factor may be due to the fact that the mean difference in peak intensity between the two levels of factor 1 is too small and that the number of peaks that are significantly affected by the factor is too low.

To further study the effect size (i.e., mean differences between factor levels) and the number of the significant factor affected peaks and the ratio of number of variable versus sample size on the ability of ASCA to identify significant factors, we filtered the peaks based on their position in the volcano plot considering all factors (Figure S1a). The selection of differentially expressed features using fold change values and their significance in discriminating groups has been used in genomics and proteomics experiments.30,21 For each factor, we selected the peaks in the data matrix \( X \) based on their \( p \)-value a fold change. The fold change values were calculated for each factor as the log ratio of the average of the peak intensities for the two levels. The \( p \)-values were obtained from a two samples \( t \)-test where the samples comprised the peak intensities of the eight peaks for each of the two levels of the factor under consideration (see Table 1). The final peak list was the union of peaks selected after Volcano filtering of all factors. Figure S1a shows an example of a volcano plot with selected peaks for a simulated and the experimental data set described above.

We studied the variance explained by the ASCA model (SSQ) and the significance obtained from permutation tests. These two parameters can be used to assess how peak selection affect ASCA performance. We expect a factor more frequently affected peaks showed a mean rank (\( \pm \)standard deviation) for factors 1, 3, and 5 were 117.05 ± 12.61, 117.05 ± 12.71, 127.23 ± 0.87, respectively, while the number of selected peaks was on average 462.99 ± 12.68. For factors 1, 3, and 5 the affected peaks showed a mean rank (\( \pm \)standard deviation) of the absolute loadings of 60.28 ± 1.69, 44.80 ± 0.57, and 64.12 ± 0.44 compared to the mean rank of the nonaffected peaks corresponding to 290.10 ± 6.92, 276.21 ± 6.37, and 295.61 ± 6.40, respectively. The mean ranks of the
nonaffected peaks of the nonsignificant factors 2, 4, 6, and 7 were 232.00 ± 6.34. This shows that the significant factor-affected peaks had lower loading ranks than nonaffected peaks and that those loadings can be used to identify peaks that are affected by significant factors.

Experimental Design Data Analysis. In experimental design, the data set is not known a priori, in which factors are significant and which peaks are affected by significant factors. Therefore, it is possible to assess factor significance using a permutation test of a given data set (with or without Volcano filtering), but it is not possible to select a Volcano based filtering threshold based on the experimental data because there is no possibility to validate the peak selection performance using g and f-scores. In this situation, optimization of the ASCA performance using Volcano peak filtering parameter may lead to overfitting. In this situation, the best that can be done is to have an assumption on the number of significant factors and number of affected peaks with a fold change and try to assess this scenario using simulation as shown in the previous section. The last step is to perform ASCA analysis with the threshold that is reasonable based on the outcome of the simulation study.

We have therefore performed the ASCA analysis of the experimental design data set with the threshold 0.05 for t-test significance and 2 for the fold ratio, which resulted in 918.4 ± 8.4 selected peaks upon 100 repeated ASCA analyses. Plots in Figure 2g show the SSQ, while the plots in Figure 2h
show the SSQ’s significance using peaks selected with \( t \)-test significance and fold ratio thresholds of 0.05 and 2.

A Volcano plot of peaks extracted with the applied threshold is presented in Figure S1b. ASCA was then performed on the filtered list of peaks to identify the main effects that affect the peptide profile of depleted human serum significantly. Figure 2h shows that two factors, hemolysis and stopping trypsin, are below the threshold of 0.05 for SSQ’s significance, of which hemolysis has the strongest influence. Trypsin digestion (trypsin/protein ratio) almost reached the threshold of 0.05. The other factors such as blood collection tube, clotting time, freeze–thaw cycle, and sample stability in the autosampler show an SSQ’s significance close to 1.

The “high level” of the most significant hemolysis factor was obtained by spiking a lysate of red blood cells into the serum samples corresponding real hemolysis levels observed in serum samples from cancer patients that were treated with chemotherapy (see description of factors in section 1.3 in the Material and Methods in the Supporting Information). In this study, the hemolysis factor can be considered as positive control for the identification of significant factors, because due to the spiking of a red blood cell lysate we expect to find peptides that show clear differences in concentration between the factor levels. The “stopping trypsin” factor has an effect on the trypsin digestion step for different amounts of trypsin/protein ratios, while the trypsin digestion factor with levels of different pH may influence the stability of certain peptides in the complex mixture. The lack of a significant effect of the clotting time is in agreement with the results of our previous study showing that there is no effect of clotting time on the digested serum peptide profile, except for fibrinopeptides when studied as single factor between 1 and 8 h.22

Figure 3 shows the loadings of the 10 most affected peaks for the significant factors of hemolysis, stopping trypsin and trypsin digestion and presents the quantification data obtained after preprocessing and raw data by means of extracted ion chromatograms (EIC) after retention time alignment and by using box plots of the preprocessed LC–MS data for two representative peaks for each of these three factors. Table S3 contains the peptide sequences and protein names of the annotated peaks. The EICs for the raw data and the box-plots for the preprocessed data show that the selected peaks are highly discriminative between the two factor levels that were identified to influence the depleted human serum profile significantly.

Annotation accuracy of the matched single stage peaks should be taken with care due to the resolution differences between the QTOF (~10 000) and the ion trap data (~1 000) and the fact that slight orthogonality between the separation in the two LC–MS platforms was observed. Annotation transfer of peaks revealed that the hemolysis factor mainly affected peptides of two hemoglobin isoforms (5 peptides), and peptide from alpha-2-macroglobulin (2 peptides), complement factor B precursor (1 peptide) and plasminogen (1 peptide), from which hemoglobin related peptides are originating from red blood cells. This is not surprising since addition of a red blood cell lysate to the sample was used to mimic low and high hemolysis levels. The other two proteins (alpha-2-macroglobulin, complement factor B precursor and plasminogen) were also reported to be present in the red blood cell proteome23 or the identified peptides may be a result of activity of proteases released from lysed red blood cells.24

CONCLUSION

Complex proteomics samples are routinely analyzed for biomarker discovery and validation. These samples are often acquired, stored, and processed under different conditions before they are analyzed by LC–MS/(MS) or other methodologies. Experimental design is the most efficient way to identify which preanalytical factors have an effect on a particular peptide in the peptide profile of complex biological samples.25

We applied ASCA\(^2\) to identify the most significant factors in an experimental design data set prepared for depleted human serum using seven factors that we considered to have an influence on the peptide profile. We simulated a data set using the same number of peaks and factors as for the experimental data, defining three factors to have an effect on a small fraction of the peaks. This simulated data set revealed that ASCA is able to identify the factors with a significant effect and that the loadings of ASCA can be applied to obtain a ranked list of affected peaks. We subsequently applied ASCA to a matched LC–MS peak data matrix X to identify preanalytical factors that have an influence on the peptide profile of depleted human serum after filtering out peaks based on fold change and \( t \)-test significance with thresholds that were obtained at maximal variance significance. The analysis revealed that hemolysis and stopping the trypsin reaction with acid have a significant influence, while the trypsin/protein ratio as a factor almost reached the significance level. EICs and quantification of peaks using the TAPP\(^1\) pipeline confirmed these differences for selected peptides in the raw LC–MS data. We demonstrated that simulation using design similar to the real data is an useful tool to assess the performance of ASCA with respect to peak quality and the simulation can guide to obtain a reasonable Volcano filtering threshold to analyze real data. The small difference of ASCA performance between the optimally selected peak set and using all observed peaks in this study could be more pronounced in data obtained from cell lysates or tissue samples with modern high-resolution LC–MS instrumentation, where the number of quantified peptides is several tens or even hundreds of thousands leading to more extreme ratios between the number of detected peaks and sample size. Importance of the peak selection effect on ASCA performance can be more pronounced in data sets where, e.g., the noise level is larger than in the studied data set. Therefore, the effect of peak selection on ASCA performance should be evaluated using simulation.

From the two significant factors (hemolysis, stopping trypsin) and one factor almost reaching significance (trypsin digestion) that have been identified in this study to affect the tryptic peptide profile of depleted human serum, hemolysis has the most pronounced effect, since hemolysis is related to addition of the lysed red blood cell content to the analyzed sample, which alters the proteome composition of human serum. In real samples it is not possible to keep the hemolysis level constant, but by measuring the light adsorption at a given set of wavelengths, it is possible to measure the level of hemolysis and discard samples, which are exceeding a certain threshold. This threshold may have to be decided on a case-by-case basis with respect to the study design. The two other factors having an influence on the peptide profile of human serum are the trypsin digestion step and stopping trypsin reaction through the addition of formic acid. The level of these two factors is set by the experimentalist and can be controlled during sample preparation. The other factors appear to be less
critical but again this depends on the study design and the depth to which the serum proteome will be measured.

Our manuscript describes the application of ASCA to analyze multivariate data obtained in a fractional factorial design study with seven factors each varied at two levels. Under the applied experimental design 16 samples were analyzed, which produces a resolution of IV and allows one to study the main effects only confounded by three-way interactions. Our generic statistical method is applicable to other situations where "omics" experiments generate multivariate, highly complex data sets from which it is hard to assess the influence of preanalytical factors on sample quality and which allows one to identify the analytes that are affected by a given factor.

### ASSOCIATED CONTENT

#### Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.5b03483.

Additional materials and methods and figures (PDF)

Matlab plots obtained with simulation evaluating random matrix and includes number of selected variables, SSQ, and significance of p-value for 7 nonsignificant factors repeated 15 times (ZIP)

Boxplot.pdf shows boxplots, while the EICPlots.pdf contains extracted ions chromatograms of the 10 most significant peaks for hemolysis, trypsin digestion, and stopping trypsin factors (ZIP)

### AUTHOR INFORMATION

#### Corresponding Author

*E-mail: p.l.horvatovich@rug.nl. Phone: +31-50-363-3341. Fax: +31-50-363-7582.*

#### Notes

The authors declare no competing financial interest.

#### ACKNOWLEDGMENTS

We thank Lorenza Franciosi for her support in sample preparation.

#### REFERENCES