Analysis and Quantification of Diagnostic Serum Markers and Protein Signatures for Gaucher Disease*[S]

Johannes P. C. Vissers‡§, James I. Langridge‡, and Johannes M. F. G. Aerts¶

Novel approaches for the qualitative and quantitative proteomics analysis by nanoscale LC-MS applied to the study of protein expression response in depleted and undepleted serum of Gaucher patients undergoing enzyme replacement therapy are presented. Particular emphasis is given to the method reproducibility of these LC-MS experiments without the use of isotopic labels. The level of chitotriosidase, an established Gaucher biomarker, was assessed by means of an absolute concentration determination technique for alternate scanning LC-MS generated data. Disease associated proteins, including fibrinogens, complement cascade proteins, and members of the high density lipoprotein serum content, were recognized by various clustering methods and sorting and intensity profile grouping of identified peptides. Condition-unique LC-MS protein signatures could be generated utilizing the measured serum protein concentrations and are presented for all investigated conditions. The clustering results of the study were also used as input for gene ontology searches to determine the correlation between the molecular functions of the identified peptides and proteins. ***Molecular & Cellular Proteomics 6:755–766, 2007.***

The most frequently encountered inherent lysosomal storage disorder in man is glycosylceramidosis, better known as Gaucher disease. The disorder is caused by an inherited deficiency in glucocerebrosidase, the enzyme catalyzing the degradation of glucosylceramide to ceramide and glucose. Lipid accumulation is restricted to tissue macrophages, so-called Gaucher cells, that act as the starting point of pathophysiological processes resulting in clinical symptoms. The clinical presentation of Gaucher disease is heterogeneous with respect to age, nature, and progression of symptoms (1). Clinical manifestation is accompanied by abnormalities in serum composition. The most striking abnormality is a thousandfold elevated serum level of chitotriosidase, a protein massively expressed and secreted by the pathological Gaucher cells (2). Although chitotriosidase is an excellent biomarker, a major drawback is the frequent genetic deficiency in this enzyme among Caucasians with approximately one in every 20 individuals not expressing any chitotriosidase (3). This limitation has prompted a search for further Gaucher cell biomarkers. The availability of biomarkers for Gaucher disease is of particular importance because the availability of effective therapeutic interventions based on supplementation with recombinant glucocerebrosidase (4) or pharmacological reduction of glycosphingolipid biosynthesis (5) is costly. The monitoring of chitotriosidase as a biomarker of Gaucher disease is generally applied in a clinical setting for both therapy initiation and optimization of individual dose regimes (6). Given the limitations concerning chitotriosidase, identification and quantification of additional biomarkers for Gaucher disease is therefore of great value (7).

2D[1] gel-based separation methods combined with mass spectrometry have been the standard for the separation, identification, and quantification of proteins. The method has to date the greatest potential to separate complex protein mixtures comprising up to thousands of components. It also has limitations with regard to the separation of certain protein classes and quantification in general. The quantitative limitations have been detailed elsewhere (8, 9), but they primarily arise from ambiguity in the identification of multiple proteins present in a single spot, identification of proteins at both extremes of the pI range, small proteins, variants and modifications, in-gel degradation, and variation in extraction efficiency. As a complementary alternative, LC-MS-based relative quantification methods have emerged to identify and quantify peptides and proteins in mixtures of various complexities. The majority of these relative quantification techniques use the introduction of stable isotopes into the samples including ICAT (10), isobaric tag for relative and absolute quantification (iTRAQ) (11), *in vivo* stable isotope labeling by amino acids in cell culture (SILAC) (12), and ^{18}O labeling (13, 14). They typically require multiple sample preparation steps that could result in an increase in experiment variability and a decrease in accuracy. Recent articles have reviewed stable isotope labeling approaches and contrasted their advantages and limitations with quantitative differential in-gel electrophoresis methods (15–17).

From the ‡Waters Corporation, MS Technologies Center, Manchester M22 5PP, United Kingdom and ¶Department of Biochemistry, Academic Medical Center, University of Amsterdam, Amsterdam, 1105 AZ, The Netherlands

Received, August 9, 2006, and in revised form, December 8, 2006
Published, MCP Papers in Press, February 9, 2007, DOI 10.1074/mcp.M600303-MCP200

[1] The abbreviations used are: 2D, two-dimensional; PCA, principal component analysis.

More recently, label-free LC-MS quantification methods have been described to determine relative abundances of proteins between multiple conditions (18–24). These methods are typically based on determining peak area ratios of the same peptides between different conditions. The quantitative reproducibility of these methods depends upon the peptide cluster efficiency, which is determined by the mass measurement accuracy and precision and the chromatographic retention time reproducibility obtained during the experiment. A recent independent study from the Association of Biomolecular Resource Facilities evaluated quantitative proteomics approaches, and it was concluded that label-free methods did at least as well as stable isotope labeling methods.[2] Moreover Silva *et al.* (25) discovered that a label-free approach allows for the estimation of absolute protein concentrations, which were subsequently used for stoichiometry studies.

In this study, a gel-free and label-free LC-MS approach is presented to conduct qualitative and quantitative serum analysis. The Gaucher disease protein serum profile was examined as it is biochemically and quantitatively well defined. The identification and enzyme activity determination of a known Gaucher disease biomarker will be demonstrated and cross-validated with its biochemical known activity. Furthermore clustering methods are described to evaluate the data quality of quantitative label-free LC-MS data sets. Clustering was also used for trend identifications based on absolute determined concentrations. Intensity profiling by *K*-means clustering of identified peptides was used to identify interrelating proteins, for example proteins that are components of the same biochemical pathway.

EXPERIMENTAL PROCEDURES

Sample Group—The control and patient samples studied were known either by contact with the Netherlands Gaucher Society or by referral to the Academic Medical Center. The diagnosis of Gaucher disease was based on deficient glucocerebrosidase activity in leukocytes and/or urine samples. EDTA plasma samples were obtained from freshly drawn blood and immediately stored at −20 °C. Serum samples from the patients were obtained prior to treatment by means of enzyme replacement therapy and after 6.5 years of treatment. Chitotriosidase activity was measured as described previously (2).

Sample Preparation/Protein Depletion—Serum samples from the control, the patient pretreatment, and the patient post-treatment were either digested as received or passed through a 10-cm × 4.6-mm multiaffinity removal system column (Agilent Technologies, Palo Alto, CA) to deplete the samples. Hence targeted high abundance proteins, including albumin, IgG, antitrypsin, IgA, transferrin, and haptoglobin, were removed.

10 μl of the undepleted serum samples was diluted with 50 mM ammonium bicarbonate (Sigma-Aldrich) prior to enzymatic digestion. A 20-μl aliquot of the serum samples was used for depletion with the multiaffinity removal system according to the manufacturer's protocol. The mobile phase buffers were provided with the system and used as received. Briefly 20 μl of serum were diluted 5-fold with 80 μl of buffer A, and particulates were removed by centrifugation through a 0.22-μm spin filter (Millipore, Billerica, MA) at 13,000 rpm for 3 min. The proteins were separated with a step gradient; the first 10 min of the gradient were maintained at 100% mobile phase A at 0.5 ml/min followed by a step to 100% mobile phase B with a flow rate of 1.0 ml/min in 0.1 min where the composition was maintained for 7 min. Reconditioning of the column was conducted with mobile phase A buffer at 1.0 ml/min for 11 min. The depletion efficiency was estimated to be 50% based on UV absorption peak area ratio of the break-through and bound fraction. The flow-through fractions were collected and buffer-exchanged with 50 mM ammonium bicarbonate, and the volume was reduced to 80 μl.

Protein Digestion Protocols—10 μl of undepleted serum was diluted with 65 μl of 50 mM ammonium bicarbonate solution and denatured in the presence of 10 μl of 1% RapiGest detergent solution (Waters Corp., Milford, MA) at 80 °C for 15 min (26). The serum samples were reduced in the presence of 5 μl of 100 mM dithiothreitol (Sigma-Aldrich) at 60 °C for 30 min. The proteins were alkylated in the dark in the presence of 5 μl of 200 mM iodoacetamide (Sigma-Aldrich) at ambient temperature for 30 min. Proteolytic digestion was initiated by adding 15 μl of 0.5 μg/μl sequencing grade, modified trypsin (Promega, Madison WI) and incubated overnight at 37 °C. Breakdown of the acid-labile detergent was achieved in the presence of 4 μl of an aqueous 12 M HCl solution at 37 °C for 15 min. The tryptic peptide solutions were centrifuged at 13,000 rpm for 10 min, and the supernatant was collected. The enzymatic digestion and treatment of the depleted serum solutions was as described above with the exception of the addition of 20 μl of 0.5 μg/μl trypsin solution.

Prior to analyses, the tryptic peptide solutions were 10-fold diluted with an aqueous 0.1% formic acid (Sigma-Aldrich) solution. A protein digest internal standard was added (1:1 dilution with 100 fmol/μl enolase from *Saccharomyces cerevisiae*) to perform absolute quantification. The LC-MS analyses were performed using 2 μl of the final serum protein digest mixtures.

Recombinant chitotriosidase (Genzyme, Cambridge, MA) was digested as described above with minor modification. 87 μl of a 50 mM ammonium bicarbonate solution was added to 5 μl of 1 mg/ml chitotriosidase stock solution. The recombinant chitotriosidase was reduced in the presence of 1 μl of 100 mM dithiothreitol at 60 °C for 30 min. Alkylation was conducted in the dark for 30 min by adding 2 μl of 100 mM iodoacetamide. Digestion was initiated by adding 5 μl of 0.5 μg/μl modified sequencing grade trypsin and incubated overnight at 37 °C.

LC-MS Configuration—Nanoscale LC separation of tryptic peptides was performed with a NanoAcquity system (Waters Corp., Milford, MA) equipped with a Symmetry C_{18} 5 μm, 5-mm × 300-μm precolumn and an Atlantis C_{18} 3 μm, 15-cm × 75-μm analytical reversed phase column (Waters Corp.). The samples were initially transferred with an aqueous 0.1% formic acid solution to the precolumn with a flow rate of 4 μl/min for 3 min. Mobile phase A was water with 0.1% formic acid, and mobile phase B was 0.1% formic acid in acetonitrile. The peptides were separated with a gradient of 3–40% mobile phase B over 90 min at a flow rate of 300 nl/min followed by a 10-min rinse with 90% of mobile phase B. The column was re-equilibrated at initial conditions for 20 min. The column temperature was maintained at 35 °C. The lock mass was delivered from the auxiliary pump of the NanoAcquity pump with a constant flow rate of 200 nl/min at a concentration of 100 fmol of [Glu^1]fibrinopeptide B/μl to the reference sprayer of the NanoLockSpray source of the mass spectrometer. All samples were analyzed in triplicate.

Analysis of tryptic peptides was performed using a Q-Tof Premier mass spectrometer (Waters Corp., Manchester, UK). For all measurements, the mass spectrometer was operated in the v-mode of anal-

[2] A. M. Falick, J. A. Kowalak, W. Lane, K. Lilley, B. Phinney, C. Turck, S. Weintraub, E. Witkowska, and N. Yates, The Proteomics Research Group 2006 Quantitative Proteomics Study, Association of Biomolecular Resource Facilities, unpublished data.

ysis with a typical resolving power of at least 10,000 full-width half-maximum. All analyses were performed using positive nanoelectrospray ion mode. The time-of-flight analyzer of the mass spectrometer was externally calibrated with NaI from *m/z* 50 to 1990 with the data post acquisition lock mass corrected using the monoisotopic mass of the doubly charged precursor of [Glu[1]]fibrinopeptide B. The reference sprayer was sampled with a frequency of 30 s. Accurate mass LC-MS data were collected in an alternating low energy and elevated energy mode of acquisition (27, 28). The spectral acquisition time in each mode was 1.5 s with a 0.1-s interscan delay. In low energy MS mode, data were collected at a constant collision energy of 4 eV. In elevated energy MS mode, the collision energy was ramped from 15 to 40 eV during each 1.5-s data collection cycle with one complete cycle of low and elevated energy data acquired every 3.2 s. The radio frequency applied to the quadrupole mass analyzer was adjusted such that ions from *m/z* 300 to 2000 were efficiently transmitted, ensuring that any ions less than *m/z* 300 observed in the LC-MS data only arose from dissociations in the collision cell.

Data Processing and Protein Identification—Continuum LC-MS data were processed and searched using ProteinLynx GlobalServer version 2.2.5 (Waters Corp.). Protein identifications were obtained with the embedded ion accounting algorithm of the software and searching a human database to which data from *S. cerevisiae* enolase were appended. The ion detection, clustering, and normalization were performed using ProteinLynx GlobalServer. The principles of the applied data clustering and normalization have been explained in great detail in previous publications (18, 20). Intensity measurements are typically adjusted on those components, *i.e.* deisotoped and charge state-reduced accurate mass retention time pairs, that replicate throughout the complete experiment for analysis at the accurate mass/retention cluster level. Components are typically clustered together with a $<$10 ppm mass precision and a $<$0.25-min time tolerance. Alignment of elevated energy ions with low energy precursor peptide ions is conducted with an approximate precision of $\pm$0.05 min. For analysis on the protein identification and quantification level the observed intensity measurements are normalized on the intensity measurement of the identified peptides of the digested internal standard.

The underlying principles of the ion accounting search algorithm have been recently described by Li *et al*.[3] In brief, all fragment ions within a retention time window associated to 1/10 of the chromatographic peak width of a precursor ion are time-aligned or assigned to the precursor. The resulting precursor-product ion list is then queried against a database utilizing an iterative three-step process whereby the culmination of each loop increases the selectivity and sensitivity of the next. In addition, the method utilizes limited database queries whereby each query accesses different sets and subsets of peptides from the proteins present in the database.

During the first step, the data are matched to only correctly cleaved proteolytic peptides whose precursor and product ion mass tolerances are within the specified tolerances, typically 10 ppm for precursor ions and 20 ppm for product ions. As a consequence of these database search tolerances, each submitted precursor provides multiple tentative peptide identifications. However, the overall strategy of the search algorithm requires that only one peptide identification is provided for each detected precursor. As a result, all other low ranking tentative peptide identifications to each securely identified precursor are not considered. In addition, the product ions used for the validation of each high ranking precursor are removed from the precursor-product list of other co-eluting precursors, thereby eliminating them for consideration when identifying coincidentally detected precursors. During the second step, precursor and product ions that have not yet been assigned are queried against a subset database of the identified proteins from the first step. This includes missed cleavages, in-source fragments, neutral losses, and variable modifications. During the last step, the remaining unidentified ions are considered against the complete database for additional protein identifications, including peptide mass fingerprint identifications.

The protein identifications were based on the detection of more than two fragment ions per peptide, more then two peptides measured per protein, and identification of the protein in at least two of three injections. The false positive rate of the ion accounting identification algorithm is typically 3–4% with a randomized database 5 times the size of the original utilized database. However, by using replication as a filter, the false positive rate is minimized as false positive identifications have a random nature and as such do not tend to replicate across injections. Additional data analysis was performed with Decisionsite (Spotfire, Somerville, MA), Excel (Microsoft Corp., Redmond, WA), and Simca-P+ (Umetrics, Umeå, Sweden).

RESULTS

Data Quality Evaluation—The observed intensity measurements were normalized for injection volume and protein load variability before conducting quantitative comparisons between conditions by applying scaling as outlined under "Experimental Procedures" and in previously published studies (18, 20). A binary comparison of the peptide precursor intensity measurements of two injections of one of the investigated conditions is discussed. A 45° diagonal line is obtained (Fig. 1) with almost no variation throughout the detected range. Intersection through the origin would have been obtained if not for the scatter on measurements for low intensity ions, *i.e.* no or minimal deviation between matched components. This example demonstrates the expected distribution in the instance of no obvious change between the investigated injections or conditions. The number of detected accurate mass/retention time pairs identified in both injections was 9292 and 9145 of which 8364 were found to be common to both injections. The number of non-redundant identified peptides from these two particular injections, utilizing the high energy fragmentation spectra and the search criteria described under "Experimental Procedures," were 1705 (18.3%) and 1725 (18.9%), respectively. This search considered normal tryptic cleavage rules with only one missed tryptic cleavage site allowed and was limited to consider only a single modification, carbamidomethylation of cysteine residues. The summed precursor intensities of these non-redundant identifications for the two injections mentioned correspond to 45.3 and 50.7% of the total ion intensity (amount) that can be detected. These fraction numbers can be considered adequate for depleted sera and are comparable with those reported previously for microbial systems (19).

These types of quality control measurements were performed on all injections and conditions. For the depleted samples, comprising three conditions/nine injections, the

[3] Li, G.-Z., Golick, D., Gorenstein, M. V., Silva, J. C., Vissers, J. P. C., and Geromanos, S. J. (2006) A novel ion accounting algorithm for protein database searches, Poster W079 presented at the Human Proteome Organisation (HUPO) 5th Annual World Congress, Long Beach, CA (October 28–November 1, 2006).

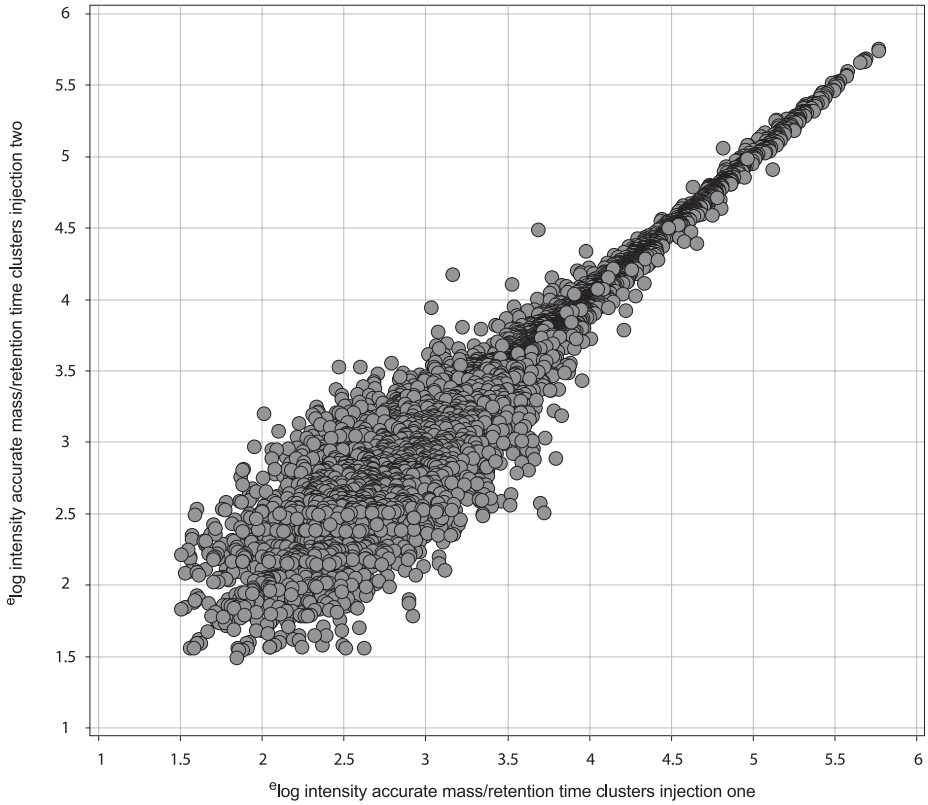

FIG. 1. **e log intensity accurate mass/retention time clusters for injection 1 *versus* e log intensity accurate mass/retention time clusters for injection 2 of one of the investigated conditions (depleted pretreatment serum).**

measured median and average mass precision were 1.90 and 2.52 ppm, respectively. The median and average retention time errors were 0.80 and 0.91%. This emphasizes the required stability of intensity, mass measurement, and retention time for label-free quantitative LC-MS measurements. These observations are within the typical error measurements reported in a previous study (18) where also more detail is given on accurate mass and retention time clustering, data normalization/scaling, and quantification.

Relative Quantification—Prior to conducting quantitative comparisons between conditions, the observed intensity measurements were normalized on the intensity measurement of the internal standard peptides. In contrast to the normalization method mentioned above, this method utilizes the three most abundant peptides identified to a protein for normalization (25). In those instances where the protein identification was based on two peptides, normalization was conducted with the two best ionizing peptides. Details on protein identification are described under "Experimental Procedures."

The relative standard deviation on the summed intensity measurement of the three most abundant peptides identified to a protein for all identified proteins for the six investigated conditions was found to be equal to 13.6% (see Supplemental Table 1; statistical outliners not excluded), which agrees well with earlier reported values using label-free quantification techniques (18, 20, 24, 29, 30). The significance of regulation level was specified at 30%. Hence 1.3-fold ($\pm$0.30 natural log scale) was used as a threshold to identify significant up- or down-regulation, which is typically 2–3 times the estimated error on the intensity measurement. The provision for a precursor ion to be included for a qualitative measurement was identification based on the search criteria described under "Experimental Procedures." Hence an assured precursor intensity threshold, typically $>$250 counts per acquisition scan, had to be reached to generate fragment ions of sufficient intensity for identification. In total, 108 non-redundant proteins were identified in the complete sample set of which 46 proteins were common to depleted and undepleted serum. 20 proteins were uniquely identified in the undepleted samples. A further 42 unique proteins were identified in the depleted serum samples.

The relative ratios and variation were individually calculated for each protein from the absolute quantification results calculated within the undepleted and depleted data sets (see Supplemental Table 2, a and b). These were calculated using the normalized summed ion intensity as described above and expressed as relative values. Of the 66 proteins identified in the undepleted sera, 35 were found to be common across all conditions, control, pretreatment, and post-treatment. For the 88 proteins identified in the depleted sera the cross-section of the three conditions equaled 56 proteins. Both cross-sections were analyzed independently, and the relative summed intensity ratio of the pre- and post-treatment samples *versus* the control samples was expressed. The majority (50 of 56) of the commonly identified proteins in the depleted identification cross-section show a clear trend to normalize as a result of

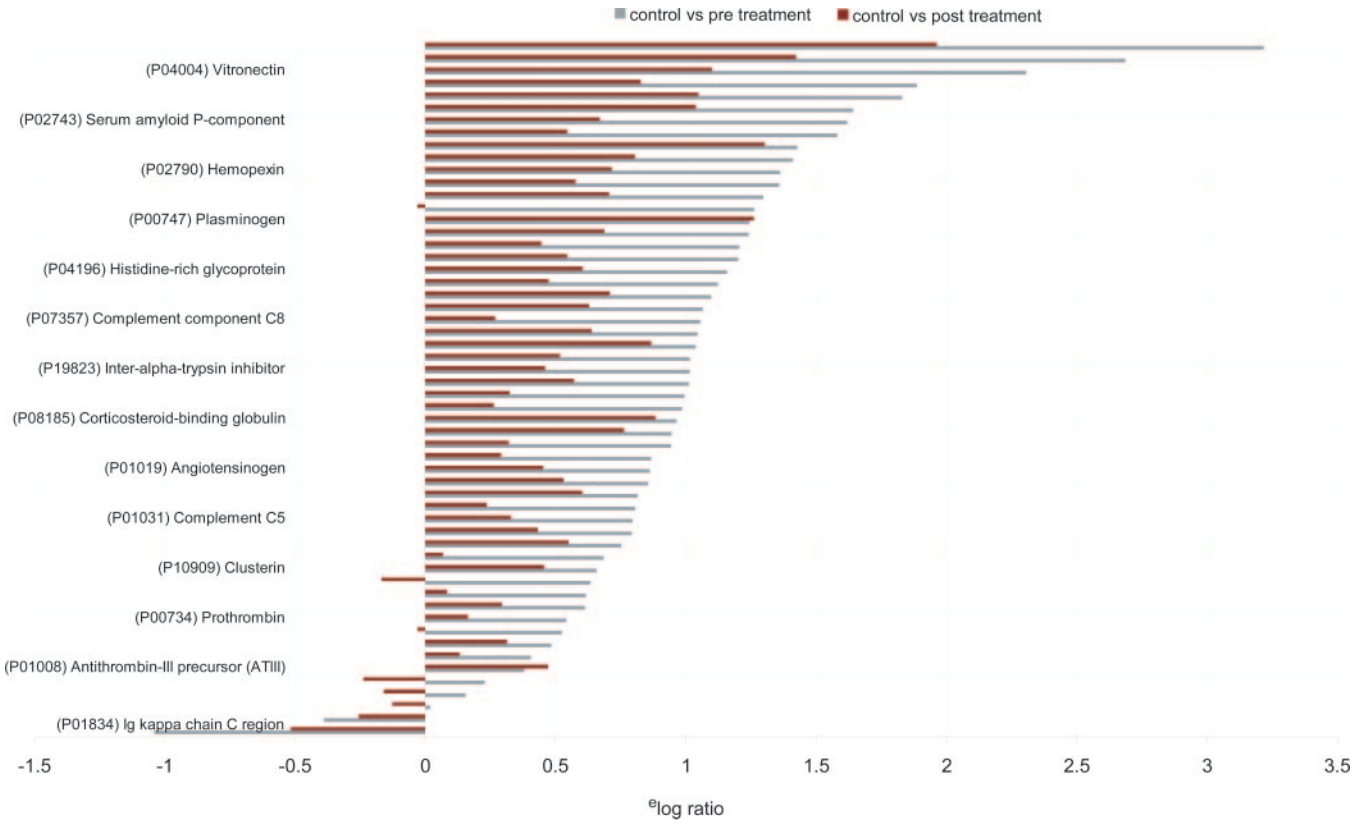

FIG. 2. **Relative protein quantification results for the identification intersection of the depleted serum samples, *i.e.* pretreatment *versus* control and post-treatment *versus* control (see Supplemental Table 2b for a complete overview of the identified and quantified proteins).** Relative pretreatment/control quantities (*gray bars*) and relative post-treatment/control quantities (*red bars*) are shown.

treatment (Fig. 2). A few proteins even show some overshoot. Similar results were obtained for the undepleted cross-section data set. In this instance, 29 of 35 proteins exhibited a similar trend upon treatment.

In addition to the relative amounts, estimated absolute serum protein concentrations are provided for the investigated undepleted control sample (see supplementary Table 2c) as these results are not affected/biased due to any sample handling or affinity depletion. The results of a more comprehensive evaluation with regard to estimating absolute protein serum concentrations by means of LC-MS are presented elsewhere (25). The lower limit of quantification, based upon protein identification, was found to be ~2.7 orders lower compared with the most abundant protein present in serum. This again required that the peptide had to be identified as described in the previous paragraphs; this is independently applicable to both depleted and undepleted serum. The estimated limit of detection at the peptide level, assuming that the precursor retention time and accurate mass are known, is ~3.5 orders lower compared with the highest abundant identified peptides; this approaches the linear dynamic range of the analytical technique used in this work.

Chitotriosidase Enzyme Activity—Recently published quantification rules (25) were used to calculate the amount and enzyme activity of chitotriosidase, a known biomarker for symptomatic Gaucher disease patients. Monitoring the concentration of chitotriosidase and other regulated proteins during treatment by means of LC-MS could be a measure for treatment efficacy. The absolute quantification method relies on the fact that the average MS signal response for the three most intense tryptic peptides per mole of protein is a constant. Given a reference, a digest of enolase from *S. cerevisiae* in this case, this relationship is used to calculate an instrument response factor for each analysis.

With this method, the average concentration of the three injections of chitotriosidase in the pretreatment sample was equal to 1.59 ± 0.31 fmol/μl, which can be calculated back to an actual enzyme activity of 39,500 ± 7860 nmol/ml·h. The determined amino acid sequence coverage for chitotriosidase was 29.2%. The enzyme activity for chitotriosidase was also determined with 4-methylumbelliferyl β-D-*N*,*N*′,*N*″-triacetylchitotriose substrate assay (31) and found to be equal to 31,800 nmol/ml·h ± 5%. The chitotriosidase level measured with both methods is in the same order of magnitude and varies by only 20%. The advantage of the LC-MS approach is the ability to calculate absolute concentrations of multiple proteins simultaneously without the requirement for isotope-labeled internal standards. It was necessary to deplete the

TABLE I
Measured concentration and activity of chitotriosidase of four pretreatment type 1 Gaucher disease patients
Three technical replicates were performed per sample.

Patient	Chitotriosidase concentration[a]	Chitotriosidase activity[b]	Chitotriosidase activity[c]
	fmol/μl	*nmol/ml·h*	*nmol/ml·h*
A[d]	1.59 ± 0.31	39,500 ± 7,860	31,800 ± 1,590
B[e]	0.99 ± 0.16	27,600 ± 4,370	15,900 ± 800
C[e]	1.59 ± 0.18	44,600 ± 4,950	62,100 ± 3,100
D[e]	1.01 ± 0.05	28,400 ± 1,400	20,400 ± 1,020

[a] Chitotriosidase concentration determined by means of LC-MS (see "Experimental Procedures" for details).
[b] Chitotriosidase activity derived from LC-MS concentration measurements.
[c] Chitotriosidase activity accessed by means of 4-methylumbelliferyl β-D-*N*,*N*′,*N*″-triacetylchitotriose substrate assay (see Ref. 31 for method details).
[d] Sample preparation and depletion as described under "Experimental Procedures."
[e] As in Table I, Footnote d but with minor modification in terms of the affinity column batch material, dilutions, and injection volumes.

patient serum to address the serum sample dynamic range to identify and quantify chitotriosidase in the pretreatment sample. The applied methods described in this and the following sections, however, can be equally successfully applied to undepleted samples.

Three additional samples obtained from other type 1 Gaucher disease patients were analyzed to statistically validate the levels and determined concentration from the LC-MS data discussed in the previous paragraph. The results of these experiments are summarized in Table I and are in agreement with the above mentioned observations that chitotriosidase is significantly elevated in the serum of patients suffering from type I Gaucher disease.

LC-MS Protein Signatures—Most protein signatures are based on comparative studies involving 2D gel separation of the proteins in their intact form and subsequent identification by means of MALDI-TOF-MS or LC-MS/MS (15, 16). The signatures with the 2D gel approach are derived from gel image analyses, identifying protein groups that share a motif or exhibit common change, and are typically presented in table format. Further analysis is sometimes performed by means of principal component analysis (PCA), or hierarchical clustering of the data, based on protein identification and their associated regulation. Gene and protein microarrays are more common tools to visualize condition-specific signatures, which allow subsequent searching of the profile against assembled collections of microarrays to identify, or classify, disease.

The absolute protein concentration values allow for the generation of so-called condition-specific signatures based on label-free quantitative LC-MS data sets, which were recently introduced for a breast cancer study.[4] Briefly a reference protein is identified within the condition that is present at a given level. Alternatively the exogenous protein spike can be used. Note that this does not necessarily have to be the same protein in every condition as the protein concentration signature will not be relative to another condition but simply relative to a constant amount. Hence it is not important that the protein identity is identical as long as the amount is. The absolute protein amounts for all other identified proteins in a condition are expressed *versus* the absolute protein amount of the reference (Fig. 3). A number of proteins, undergoing significant change, are color-annotated to illustrate the signature usefulness. For instance, apolipoprotein A-I and complement C3 are at a relatively high concentration in the control, at a relatively low concentration when treatment is started, and subsequently close to the control concentration level again in the post-treatment sample. By annotating selected proteins, disease types can be characterized in a global manner by looking at a specific panel of proteins within the plasma proteome as a whole. In this example, apolipoprotein A-I, apolipoprotein C-II, complement C3, and chitotriosidase are proteins that have been shown previously to be regulated in Gaucher patients (4, 6, 32). To date, signatures have been determined for the pre- and post-treatment patient samples and a single control. Midtreatment signatures are currently considered as a treatment monitor tool. A more extensive study with a larger patient group is currently ongoing to clinically validate the identified serum signatures.

Data Clustering: Multivariate Analysis—The data can be clustered in various ways to identify the quality of the data set, to identify differences between conditions, and to generate profiles. PCA can be conducted on the peptide mass/retention time cluster or the protein identification/protein concentration (amount) level. Prior to PCA on the peptide mass/retention time cluster level, normalization of the data set was conducted by normalizing on the total ion intensity for every injection as described under "Experimental Procedures." From the results (Fig. 4*a*) it can be concluded that the replicate injections for each condition are consistent as they cluster closely together. In other words, PCA allows for the rapid verification of the quality of the conducted experiment. Furthermore it can be seen that the PCA experiment separated the three investigated conditions, control, patient pre-, and post-treatment. Hence the enzyme replacement therapy had a clear effect shown by the separation of the pre- and post-treatment injections. However, PCA by itself at the peptide mass/retention time cluster level is not conclusive in deter-

[4] Vissers, J. P. C., Kipping, M., Reimer, T., Kasten, A., Koy, C., Langridge, J. I., and Glocker, M. O. (2006) Quantification of diagnostic protein signatures of polygenic diseases characterized by mass spectrometric proteome analysis: a study on mamma carcinoma, Poster 168 presented at the 2006 Meeting of the Association of Biomolecular Resource Facilities, Long Beach, CA (February 11–14, 2006).

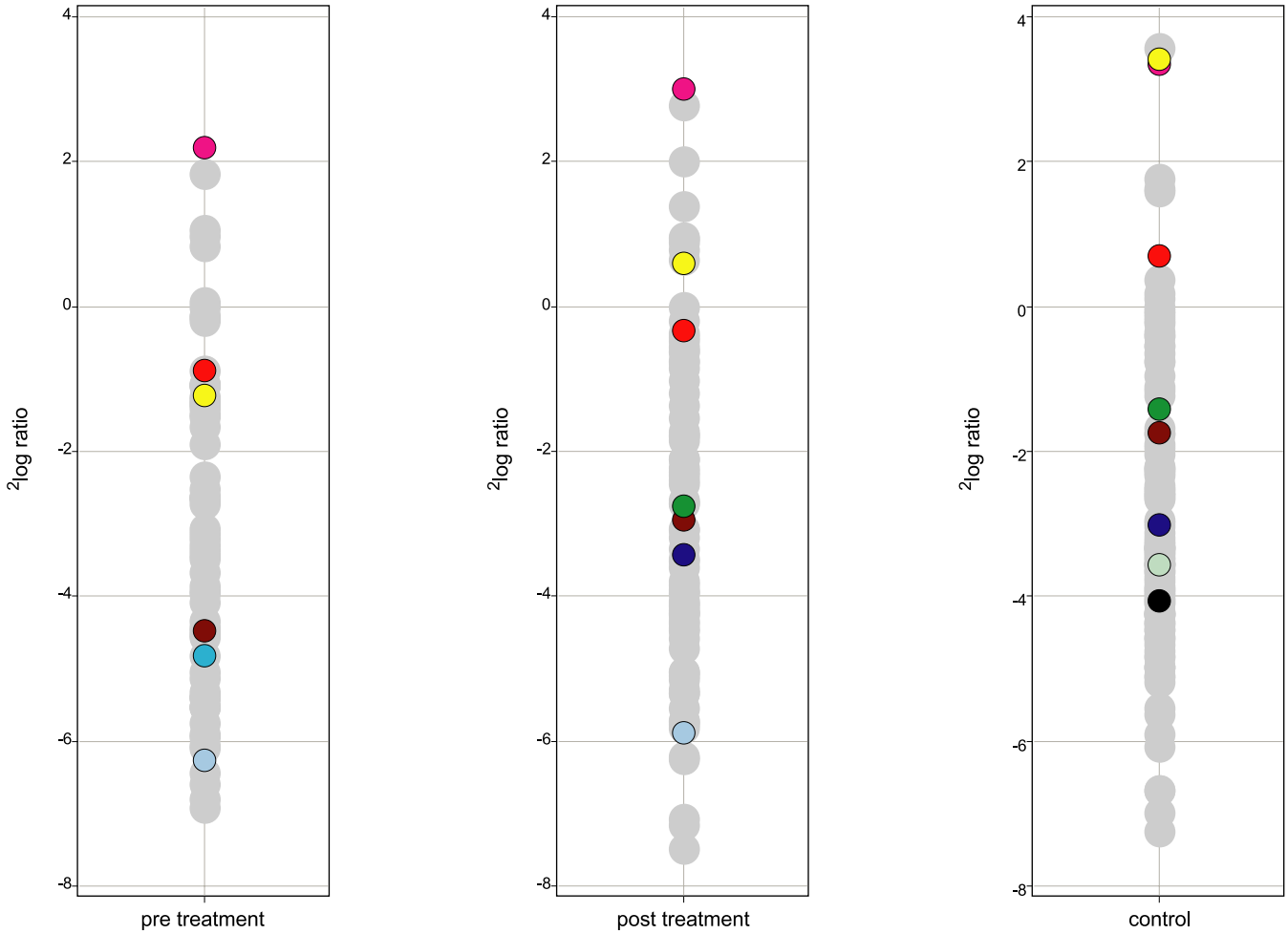

FIG. 3. **Condition-independent absolute quantification depleted serum protein signatures: ^{2}log ratio of the absolute protein concentration *versus* absolute internal standard concentration for the pretreatment serum sample (*left*), the post-treatment serum sample (*middle*), and the control serum sample (*right*).** Annotated proteins are: α_1-acid glycoprotein (*red*), α_2-antiplasmin (*dark blue*), apolipoprotein A-I (*yellow*), apolipoprotein C-II (*black*), C4b-binding protein α chain (*bright green*), chitotriosidase (*turquoise*), complement C1q subcomponent (*light blue*), complement C3 (*pink*), serum amyloid A-4 (*gray-green*), and vitronectin (*brown-red*).

mining treatment efficiency as the number of dimensions was not significantly reduced. Therefore this type of analysis can be regarded as a first pass control for an experiment as outlier injections are easily identified and condition similarities/differences can easily be detected.

The protein identification results for the depleted and undepleted serum samples were annotated with absolute determined concentrations for PCA at the protein identification level (Fig. 4*b*). However, PCA can be easily skewed by the presence of unique proteins such as the targeted depleted set of proteins in the case of a comparison between the depleted and undepleted serum samples. Therefore, only the proteins that were identified in all injections within all conditions were taken into consideration for PCA. The two main principle components can be easily identified as treatment and depletion (Fig. 4*b*). The latter is a result of unspecific partial binding of non-targeted proteins or additional sample handling losses. For instance, the absolute estimated concentrations of apolipoprotein C-III and vitronectin for the non-depleted control sample, based upon three consecutive injections and identifications, were $9.85{\cdot}10^7$ and $2.24{\cdot}10^8$ pg/ml; the reported literature values are, respectively, $10.00{\cdot}10^7$ and $2.60{\cdot}10^8$ pg/ml (33). This agreement between LC-MS and biochemical quantification methods has been studied across a larger patient population (25), and good correlation was observed between techniques for 11 well characterized serum proteins. The absolute concentration values of apolipoprotein C-III and vitronectin in the depleted control sample were $1.55{\cdot}10^7$ and $0.86{\cdot}10^8$ pg/ml indicating a non-specific loss of protein as a result of sample handling. The average loss of protein as a result of the applied depletion technique and additional sample handling was found to be ~50%. However, no generic depletion fractionation factor can be derived as this is dependent on the interaction of these proteins with either the affinity column or their interaction with the targeted proteins. This is despite the fact that the applied depletion technique is

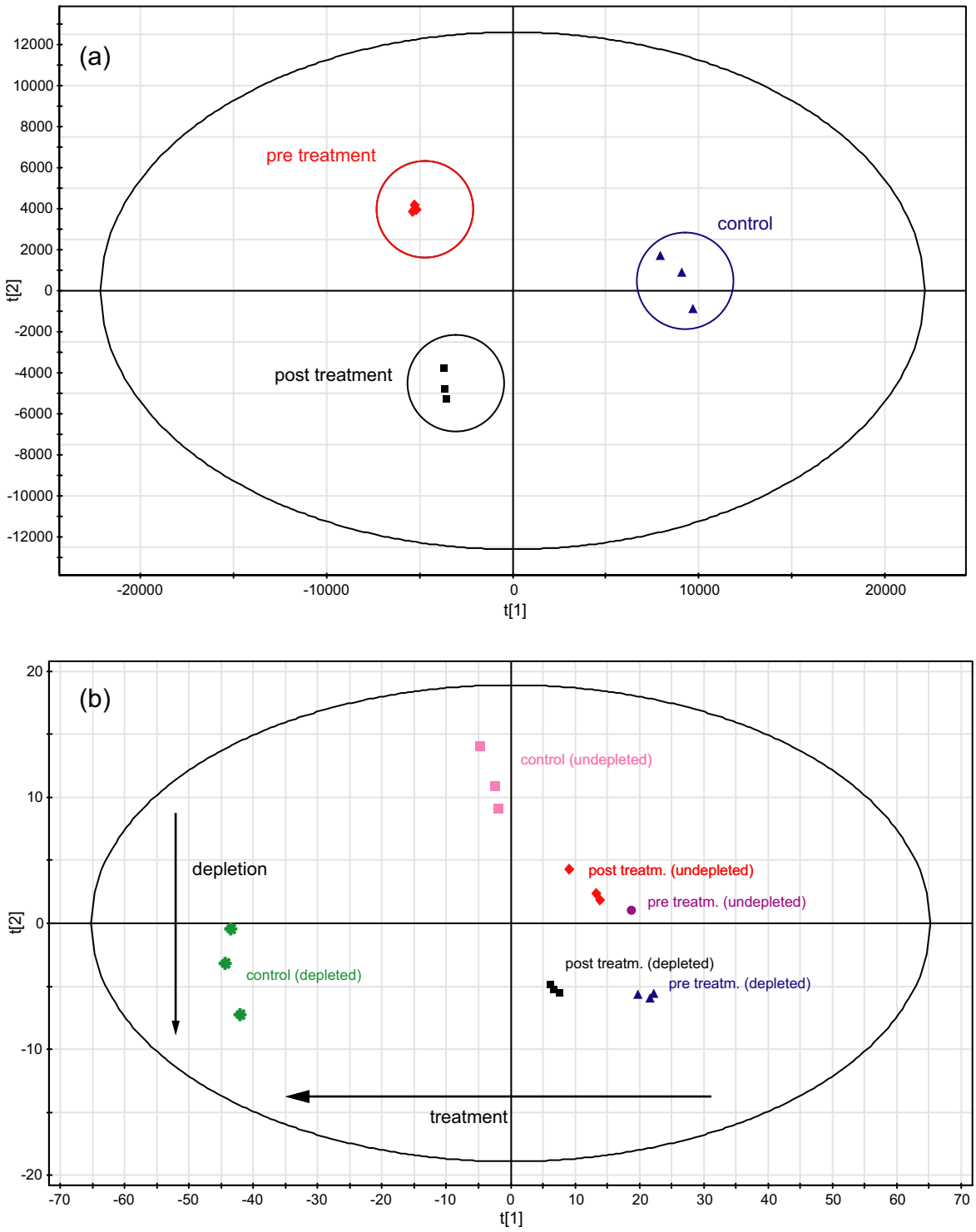

FIG. 4. *a*, principal component analysis, utilizing Pareto scaling, of the accurate mass/retention time pairs of the depleted serum samples, pretreatment (*red diamonds*), post-treatment (*black squares*), and control serum (*blue triangles*). *b*, partial least squares projection to latent structures analysis, supervised clustering utilizing Pareto scaling, of the absolute amounts (pg/ml) of the serum proteins common to all conditions. Conditions are: undepleted pretreatment (*purple circles*), undepleted post-treatment (*red diamonds*), undepleted control serum (*pink squares*), depleted pretreatment (*blue triangles*), depleted post-treatment (*black squares*), and depleted control serum (*green stars*). *treatm.*, treatment.

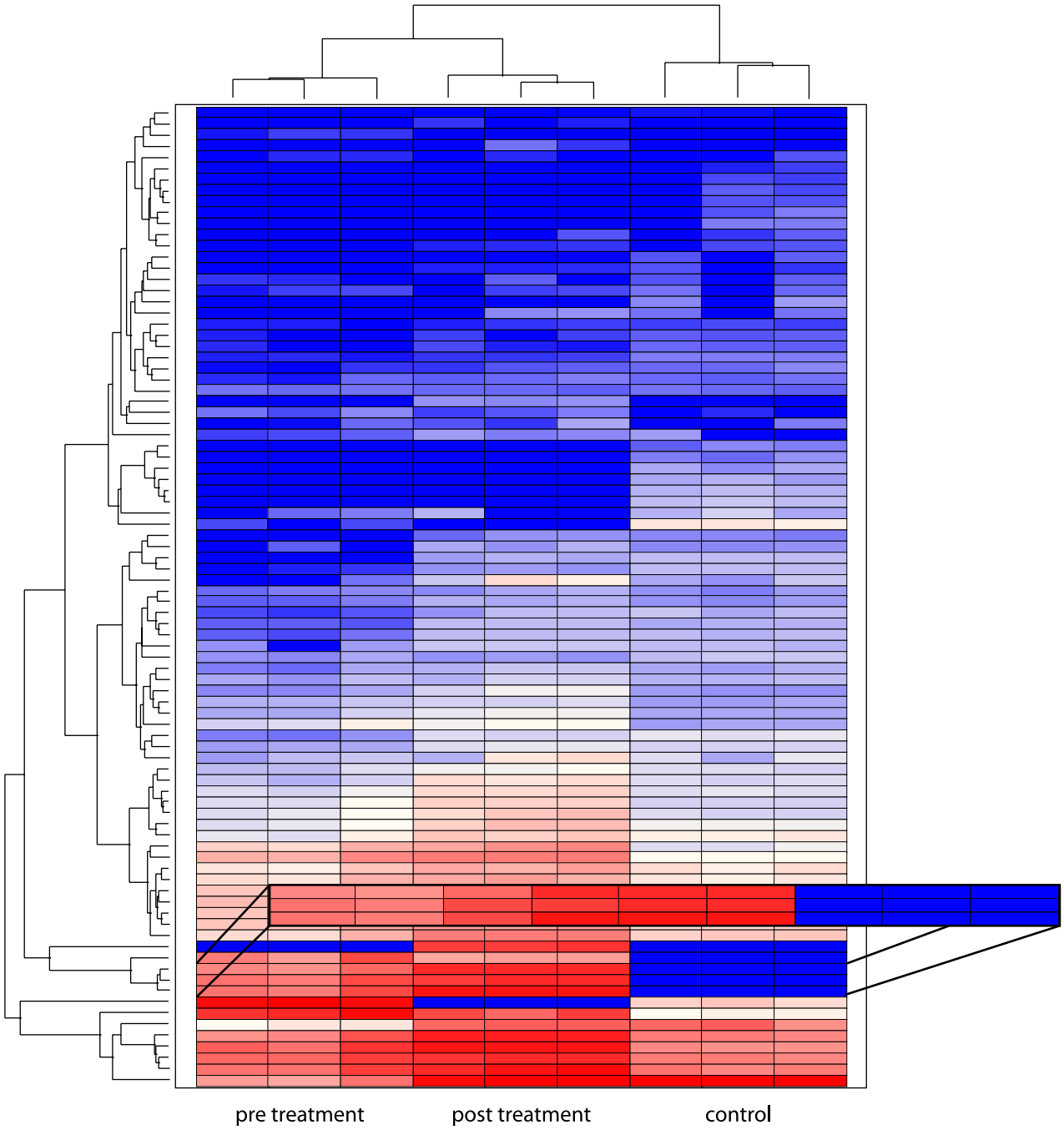

FIG. 5. **Unweighted average hierarchical clustering with euclidean distance similarity measurement of the ^{2}log absolute concentrations (fmol/μl) of the triplicate injections of the depleted serum samples by means of euclidean (nearest) distance similarity measurements.** The color legend for the heat map is as follows: lowest abundance, *blue*; medium abundance, *white*; highest abundance, *red*.

reported to be reproducible and robust.[5] Extreme instances were observed where >80% of non-targeted protein was removed, *e.g.* clusterin and complement C1r, or the protein concentration was not affected, *e.g.* α_{1B}-glycoprotein, α_1-microglycoprotein, and kininogen. These findings are intriguing but will vary dependent on the applied depletion technology (affinity efficiency and kinetics) and protein-protein interactions and how well the latter can be minimized.

PCA at the protein level using absolute concentrations also illustrates the effect of treatment as the cluster for the post-treatment sample migrates closer to the control, agreeing with

[5] Chakraborty, A. B., Berger, S. J., Dorschel, C., Geromanos, S. J., Li, G.-Z., and Gebler, J. C. (2006) Is subtractive affinity depletion of abundant serum proteins useful and reproducible?, Poster 547 presented at the 54th ASMS Conference on Mass Spectrometry, Seattle, WA (May 28–June 1, 2006).

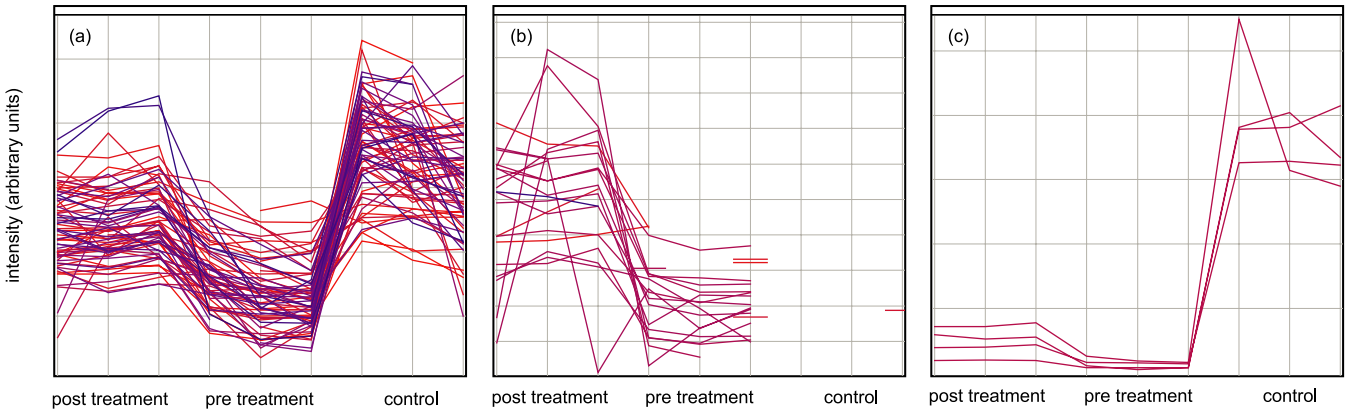

FIG. 6. ***K*-means clustering examples, Euclidean distance similarity measurement and data centroid-based search cluster initialization, of the intensities of peptides positively identified to a protein for the depleted serum samples, post-treatment (*left three columns*), pretreatment (*middle three columns*) and control serum (*right three columns*).** The maximum number of clusters (*K*) was defined as 50. Detailed information in terms of number of identified peptide and proteins for profiles *a*, *b*, and *c* is provided in supplemental Table 3.

the relative quantification experiment results (Fig. 2). Also here it can be observed that triplicate injections cluster together and that PCA is well suited for multivariate analysis of multicondition experiments.

A common method for the multivariate analysis of gene expression or 2D gel sample set visualization is hierarchal clustering. The determined absolute protein concentrations by LC-MS can also be applied to this type of analysis. The ^{2}log values of the absolute quantities were used to cluster injections of the depleted samples. As illustrated by the condition dendrograms displayed on the top of the heat map visualization (Fig. 5), the triplicate injections for each condition are closest in origin similarity with the pre- and post-treatment sample next closest. This is in agreement with the PCA results at the identification/protein concentration level (Fig. 4). Differences in expression and concentration can be easily distinguished in heat map visualizations. For example, in the selection (Fig. 5), three proteins are highly abundant in the pre- and post-treatment sample and not identified in the control, corresponding to fibrinogen α, β, and γ chains.

K-means Clustering—The intensities of peptides identified to a protein can also be clustered based upon their profile by *K*-means clustering. This method is used for grouping data points (peptide precursor intensities) identified to a protein into a predetermined number of clusters based on their similarity. Three example clusters for the depleted samples will be discussed. The first cluster (Fig. 6*a*) comprises 60 specific peptide profiles. This particular cluster highlights the response of a group of peptides to treatment. The *three left columns* correspond to the triplicate injections of post-treatment, the *middle three columns* correspond to pretreatment, and the *right three columns* correspond to control serum samples. The profiled peptides can be mapped to their parent proteins and summarized in detail (see Supplemental Table 3). 10 proteins contributed at least three peptides to the profile, accounting for 74% of the identified peptides within the cluster. For the second example cluster (Fig. 6*b*), three proteins contributed with at least three peptides, accounting for 82% of the identified profiles. In the last example (Fig. 6*c*), one protein contributed to all peptide profiles (100%).

These results demonstrate the possibility of grouping peptides based on their profile change across conditions leading to the identification of proteins from multiple peptide profiles. Possibly this can also be used to identify proteins that are involved in the same biological process or pathway. For instance, of the 10 proteins that contributed to the first profile cluster, six are part of the complement and coagulation cascades (34), namely complement C3 and C4, complement factors B and H, kininogen, and plasminogen. The lower intensity peptides of these proteins clustered together into two other clusters, which have a profile shape almost identical to this cluster. Additional proteins were identified within these clusters that are also part of the complement cascade pathway, namely complement C1R, C1S, C2, C5, C6, C6, and C8 and complement factor 1. Hence all proteins identified in the complement cascade pathway show the same regulation trends (see Supplemental Fig. 1). It has been reported that Gaucher patients show a low level of coagulation activation. In a study with 30 patients, parameters of coagulation and fibrinolysis were analyzed pre- and post-treatment with enzyme replacement therapy (32). Severe abnormalities in the coagulation system were noted, contributing to the bleeding tendency of Gaucher patients. The reduction in serum content of the proteins in the coagulation pathway in this study is consistent with the earlier investigation (32). The same holds for the observed increase in fibrinogen α, β, and γ chain described in the previous hierarchical clustering section.

The second cluster (Fig. 6*b*) represents proteins that are up-regulated in patient pre- and particularly in post-treatment samples; these are the previously mentioned fibrinolysis proteins. The third cluster (Fig. 6*c*) represents apolipoprotein A-I,

a protein down-regulated in pre-treatment and only marginally normalizing upon therapy. Apolipoprotein A-I is a component of the high density lipoprotein serum content. The high density lipoprotein in Gaucher patients is extremely low and known to be poorly normalized following enzyme replacement therapy (6). The results obtained with the presented label-free quantitative MS methods are, therefore, again consistent with biochemical findings. Further investigation of the cluster profiles is ongoing to identify potentially new markers of interest and relationships between the identified proteins.

DISCUSSION

A label-free LC-MS method has been described for the absolute and relative quantification of Gaucher disease-related biomarkers and indicators in undepleted and affinity-depleted serum. Several novel concepts have been presented including absolute protein concentration measurement in serum, PCA using absolute serum protein concentrations, clustering of peptide profiles to elucidate protein families, and condition-specific signatures based upon LC-MS data. To our knowledge these techniques have not been described previously and are presented here for the first time. The level of chitotriosidase was estimated by the absolutely determined amounts identified in depleted patient serum and found to be in good agreement with the level based on the activity measured by means of a biochemical assay. This was compared with serum chitotriosidase levels determined from three further type I Gaucher patients of which all showed a significant overexpression of this protein compared with control serum. Furthermore clustering approaches have been presented that allow for data quality assessment when the data are analyzed at the accurate mass/retention time level. Cluster analyses at the absolute protein concentration level revealed protein relationships and treatment effects and were confirmed with biochemical assay data as well. Condition-unique LC-MS protein signatures were established that allow for the analysis of a single condition. The absolute concentration LC-MS signatures do not require comparative analysis and are therefore easily extended to larger scale studies. Lastly peptide intensity clustering was shown for the identification of proteins involved in either the same complex or biochemical pathway. 12 proteins were identified that are all involved in the complement cascade using this approach, illustrating that they share similar expected stoichiometry.

Acknowledgments—Chris Hughes is kindly acknowledged for assistance with the depletion of the serum samples. Keith Richardson, Guo-Zhong Li, and Scott Geromanos are thanked for help with the analysis of the data and the involved statistics.

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

S The on-line version of this article (available at http://www.mcponline.org) contains supplemental material.

§ To whom correspondence should be addressed: Waters Corp., Market Development Proteomics, Atlas Park, Simonsway, Manchester M22 5PP, UK. Tel.: 44-161-435-4100; Fax: 44-161-435-4444; E-mail: hans_vissers@waters.com.

REFERENCES

1. Beutler E., and Grabowski, G. (1995) Gaucher disease, in The Metabolic Basis of Inherited Disease (Scriver, C. R., Beaudet, A. L., Sly, W. S., and Valle, D., eds) pp. 2641–2670, McGraw-Hill Publishing Co., New York
2. Hollak, C. E. M., van Weely, S., van Oers, M. H. J., and Aerts, J. M. F. G. (1994) Marked elevation of plasma chitotriosidase activity; a novel hallmark of Gaucher disease. *J. Clin. Investig.* **93,** 1288–1292
3. Boot, R. G., Renkema, G. H., Verhoek, M., Strijland, A., Bliek, J., de Meulemeester, T. M., Mannens, M. M., and Aerts, J. M. F. G. (1998) The human chitotriosidase gene. Nature of inherited enzyme deficiency. *J. Biol. Chem.* **273,** 25680–25685
4. Barton, N. W., Furbish, F. S., Murray, G. J., Garfield, M., and Brady, R. O. (1990) Therapeutic response to intravenous infusions of glucocerebrosidase in a patient with Gaucher disease. *Proc. Natl. Acad. Sci. U. S. A.* **87,** 1913–1916
5. Cox, T., Lachmann, R., Hollak, C. E., Aerts, J. M. F. G., van Weely, S., Hrebicek, M., Platt, F., Butters, T., Dwek, R., Moyses, C., Gow, I., Elstein, D., and Zimran, A. (2000) Novel oral treatment of Gaucher's disease with N-butyldeoxynojirimycin (OGT918) to decrease substrate biosynthesis. *Lancet* **355,** 1481–1485
6. Aerts, J. M. F. G., Hollak, C. E. M., van Breemen, M., Maas, M., Groener, J. E., and Boot, R. G. (2005) Identification and use of biomarkers in Gaucher disease and other lysosomal storage diseases. *Acta Paediatr. Suppl.* **94,** 43–46, 37–38
7. Boot, R. G., Verhoek, M., de Fost, M., Hollak, C. E. M., Maas, M., Bleijlevens, B., van Breemen, M. J., van Meurs, M., Boven, L. A., Laman, J. D., Moran, M. T., Cox, T. M., and Aerts, J. M. F. G. (2004) Marked elevation of the chemokine CCL18/PARC in Gaucher disease: a novel surrogate marker for assessing therapeutic intervention. *Blood* **103,** 33–39
8. Rabilloud, T. (2002) Two-dimensional gel electrophoresis in proteomics: old, old fashioned, but it still climbs up the mountains. *Proteomics* **2,** 3–10
9. Corthals, G. L., Wasinger, V. C., Hochstrasser, D. F., and Sanchez, J. C. (2000) The dynamic range of protein expression: a challenge for proteomic research. *Electrophoresis* **21,** 1104–1115
10. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biothechnol.* **17,** 994–999
11. Ross, R. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A., and Pappin, D. J. (2004) Multiplex protein quantification in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **3,** 1154–1169
12. Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1,** 376–386
13. Shevchenko, A., Chernushevich, I., Ens, W., Standing, K. G., Thomas, B., Wilm, M., and Mann, M. (1997) Rapid 'de novo' peptide sequencing by a combination of nanoelectrospray, isotopic labeling and a quadrupole/time-of-flight mass spectrometer. *Rapid Commun. Mass Spectrom.* **11,** 1015–1024
14. Yao, X., Freas, A., Ramirez, J., Demirev, P. A., and Fensalau, C. (2001) ^{18}O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Anal. Chem.* **73,** 2836–2842
15. Lill, J. (2003) Proteomic tools for quantification by mass spectrometry. *Mass Spectrom. Rev.* **22,** 182–194
16. Hamdan, M., and Righetti, P. G. (2002) Modern strategies of protein quantification in proteome analysis: advantages and limitations. *Mass Spectrom. Rev.* **21,** 287–302
17. Ong, S.-E., and Mann, M. (2005) Mass spectrometry-based proteomics turn quantitative. *Nat. Chem. Biol.* **5,** 252–262
18. Silva, J. C., Denny, R., Dorschel, C. A., Gorenstein, M., Kass, I. J., Li, G.-Z., McKenna, T., Nold, M. J., Richardson, K., Young, P., and Geromanos, S. J. (2005) Quantitative proteomic analysis by accurate mass retention

time pairs. *Anal. Chem.* **77,** 2187–2200
19. Silva, J. C., Denny, R., Dorschel, C. A., Gorenstein, M. V., Li, G.-Z., Richardson, K., Wall, D., and Geromanos, S. J. (2006) Simultaneous qualitative and quantitative analysis of the Escherichia coli proteome: a sweet tale. *Mol. Cell. Proteomics* **5,** 589–607
20. Hughes, M. A., Silva, J. C., Geromanos, S. J., and Townsed, C. A. (2006) Quantitative proteomic analysis of drug-induced changes in mycobacteria. *J. Proteome Res.* **5,** 54–63
21. Wang, W., Zhou, H., Lin, H., Roy, S., Shaler, T. A., Hill, L. R., Norton, S., Kumar, P., Anderle, M., and Becker, C. (2003) Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal. Chem.* **75,** 4818–4826
22. Radulovic, D., Jelveh, S., Ryu, S., Hamilton, T. G., Foss, E., Mao, Y., and Emili, A. (2004) Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **3,** 984–997
23. Wiener, M. C., Sachs, J. R., Deyanova, E. G., and Yates, N. A. (2004) Differential mass spectrometry: a label-free LC-MS method for finding differences in complex peptide and protein mixtures. *Anal. Chem.* **76,** 6085–6096
24. America, A. H. P., Cordewener, J. H. G., van Geffen, M. H. A., Lommen, A., Vissers, J. P. C., Bino, R. J., and Hall, R. D. (2006) Alignment and statistical difference analysis of complex peptide data sets generated by multidimensional LC-MS. *Proteomics* **6,** 641–653
25. Silva, J. C., Gorenstein, M. V., Li, G.-Z, Vissers, J. P. C., and Geromanos, S. J. (2006) Absolute quantification of proteins by LCMS^{E}. A virtue of parallel MS acquisition. *Mol. Cell. Proteomics* **5,** 144–156
26. Yu, Y. Q., Gilar, M., Lee, P. J., Bouvier, E. S. P., and Gebler, J. C. (2003) Enzyme-friendly, mass spectrometry-compatible surfactant for in-solution enzymatic digestion of proteins. *Anal. Chem.* **75,** 6023–6028
27. Bateman, R. H., Langridge, J. I., McKenna, T., and Richardson, K. *Methods and Apparatus for Mass Spectrometry*. U. S. Patent 2,385,918A, September 26, 2006
28. Bateman, R. H., Carruthers, R., Hoyes, J. B., Jones C., Langridge, J. I., Millar, A., and Vissers, J. P. C. (2002) A novel precursor ion discovery method on a hybrid quadrupole orthogonal acceleration time-of-flight mass spectrometer for studying protein phosphorylation. *J. Am. Soc. Mass Spectrom.* **13,** 792–803
29. Fortier, M.-H., Bonneil, E., Goodley, P., and Thibault, P. (2005) Integrated microfluidic device for mass spectrometry-based proteomics and its application to biomarker discovery programs. *Anal. Chem.* **77,** 1631–1640
30. Ghitun, M., Bonneil, E., Fortier, M.-H., Yin, H., Killeen, K., and Thibault, P. (2006) Integrated microfluidic devices with enhanced separation performance: application to phosphoproteome analyses of differentiated cell model systems. *J. Sep. Sci.* **29,** 1539–1549
31. Aerts, J. M. F. G., Donker-Koopman, W. E., Koot, M., Barranger, J. A., Tager, J. M., and Schram, A. W. (1986) Deficient activity of glucocerebrosidase in urine from patients with type 1 Gaucher disease. *Clin. Chim. Acta* **158,** 155–164
32. Hollak, C. E. M., Levi, M., Barends, F., Aerts, J. M. F. G., and van Oers, H. J. (1997) Coagulation abnormalities in type 1 Gaucher disease are due to low-grade activation and can be partly restored by enzyme supplementation therapy. *Br. J. Haematolog.* **96,** 470–476
33. Anderson, L. (2005) Candidate-based proteomics in the search for biomarkers of cardiovascular disease. *J. Physiol.* **563,** 23–60
34. Janeway, C. A., Jr., Travers, P., Walport, M., Shlomchik, M. J. (2001) in *Immunobiology, Part I, an Introduction to Immunobiology and Innate Immunity*, 5th Ed., pp. 1–92, Garland Publishing, New York