Available online at www.sciencedirect.com
SCIENCE DIRECT®
International Journal of Mass Spectrometry 238 (2004) 163–171
International Journal of Mass Spectrometry
www.elsevier.com/locate/ijms

Differential expression profiling of serum proteins and metabolites for biomarker discovery

Sushmita Mimi Roy*, Markus Anderle, Hua Lin, Christopher H. Becker

SurroMed, Inc., 1430 O'Brien Drive, Menlo Park, CA 94025, USA

Received 1 October 2003; accepted 1 March 2004
Available online 30 September 2004

Abstract

A liquid chromatography-mass spectrometry (LC-MS) proteomics and metabolomics platform is presented for quantitative differential expression analysis. Proteome profiles obtained from 1.5 μL of human serum show ∼5000 de-isotoped and quantifiable molecular ions. Approximately 1500 metabolites are observed from 100 μL of serum. Quantification is based on reproducible sample preparation and linear signal intensity as a function of concentration. The platform is validated using human serum, but is generally applicable to all biological fluids and tissues. The median coefficient of variation (CV) for ∼5000 proteomic and ∼1500 metabolomic molecular ions is approximately 25%. For the case of C-reactive protein, results agree with quantification by immunoassay. The independent contributions of two sources of variance, namely sample preparation and LC-MS analysis, are respectively quantified as 20.4 and 15.1% for the proteome, and 19.5 and 13.5% for the metabolome, for median CV values. Furthermore, biological diversity for ∼20 healthy individuals is estimated by measuring the variance of ∼6500 proteomic and metabolomic molecular ions in sera for each sample; the median CV is 22.3% for the proteome and 16.7% for the metabolome. Finally, quantitative differential expression profiling is applied to a clinical study comparing healthy individuals and rheumatoid arthritis (RA) patients.

Keywords: Differential proteomics; Biomarkers; Liquid chromatography-mass spectrometry; Quantification

1. Introduction

The discovery of new diagnostic and prognostic markers of therapeutic response holds tremendous promise for drug discovery and development. Several useful biomarkers of disease have been discovered by hypothesis-driven research using directed technologies such as conventional biochemical methods and immunological assays [1,2]. However, a large number of serious diseases and conditions warrant non-hypothesis driven research because reliable hypotheses are weak or lacking. We present such a platform for differential expression analysis of proteins and metabolites in human serum. Not only is it undirected and comprehensive, this platform is also applicable to all body fluids or tissues. This new platform is described and the ability to differentially quantify proteins and metabolites is demonstrated in a study of rheumatoid arthritis (RA).

Several methods have been published for quantifying biomolecules for differential profiling. Quantification of specific analytes using spiking with isotopically labeled compounds has contributed significantly to hypothesis-driven discoveries involving metabolites and proteins and biochemical pathway analysis. One- and two-dimensional (2D) gels have been used to separate proteins and quantify gel spots by silver staining, fluorescence or radioactive labeling [3]. These differentially stained spots have been identified by MS/MS, leading to interesting putative biomarkers [4]. However, 2D-gel methods have inherent shortcomings such as loss of low and high molecular weight proteins, limited dynamic range, low reproducibility and a requirement for large amounts of material. These are well known, and have

* Corresponding author. Tel.: +1 650 470 2373; fax: +1 650 470 2400.
E-mail address: mroy@surromed.com (S.M. Roy).

1387-3806/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijms.2004.03.017

spurred many to look at on-line chromatography coupled to mass spectrometry [5,6].

The coupling of a separation method such as reverse-phase liquid-chromatography to mass spectrometry allows the measurement of a large number of biomolecules, each with a characteristic *m*/*z* and retention time, from a relatively small amount of a complex biological material without sacrificing sensitivity or throughput. Typically, the proteins must be enzymatically digested with trypsin to yield fragments that are small enough for sensitive and identifiable mass spectrometric analysis. Using such liquid chromatography-mass spectrometry (LC-MS) approaches, several groups have demonstrated quantification.

This LC-MS (also MALDI-MS, i.e., matrix-assisted laser desorption ionization mass spectrometry) quantification can be based on isotopic labeling, called isotope coded affinity tags (ICAT) and related methods. In this approach, a specific amino acid in two samples is differentially and isotopically labeled and subsequently separated from peptide background by solid-phase capture, wash and release [7–9]. The ratio of intensities of the molecule from the two sources with different isotopic labels can then be determined. However, drawbacks include sample preparation complexity, reagent expense, material losses, non-specifically captured peptide background, required presence of a specific amino acid and elimination of peptides without this amino acid, a challenge to compare a large number of samples, and frequent difficulty in obtaining useful tandem mass spectrometry (MS/MS) fragmentation patterns.

Both 2D-gel and ICAT related technologies fail to address a very important part of the biological sample, namely, the metabolites. Since differential profiling for biomarker discovery would benefit from a technology that can compare metabolites of various chemical structure as well as proteins, and compare these molecules over relatively large numbers of subjects, there is good reason to explore other approaches such as the one presented here.

Our approach to quantify the relative concentrations of analytes by LC-MS is based on separation of the biological fluid into proteins and metabolites followed by direct spectral intensity measurement of all molecular ions present in the metabolome and in the trypsin-digested proteome, for each sample independently, and then comparing those intensities for such molecules. Historically, concerns have been expressed about non-linearities and ion suppression effects in the circumstance of complex biological matrices [10]. However, for our relatively long chromatographic gradients (∼1 h), we have found that the molecular ion intensities of spiked analytes increase in a near linear fashion with concentration, even in a complex biological matrix such as serum [11]. Differential quantification is further optimized by a simple global normalization of data between samples [12,13]. The analysis is made possible by developing and employing a computer application, MassView^{TM} software, which de-isotopes and tracks the molecular ions and performs normalization by employing signals of molecules that do not change concentration from sample to sample. The linear signal response coupled to our ability to reproducibly prepare samples and normalize, correlate and quantify molecular ion intensities over several samples forms the basis of this new quantitative proteomics and metabolomics platform.

In this paper, using this approach, we demonstrate quantification of more than 5000 proteome and ∼1500 metabolome de-isotoped molecular ions per sample, for significant numbers of serum samples. We measure the coefficients of variation related with making these measurements over 20 human serum samples from a common pooled source. We investigate the contribution to this overall variation from sample preparation prior to LC-MS and from the LC-MS analysis itself. Also, in this paper, we demonstrate an application of this quantitative differential proteomics with comparison of serum from 19 normal individuals and 19 RA patients. Biological variance in a healthy population is estimated by quantifying the variance observed in intensities of ∼6500 molecular ions from the sera of these 19 normal individuals.

2. Materials and methods

2.1. Sample preparation

For studies with pooled serum, human serum for proteomics was purchased from Sigma-Aldrich (St. Louis, MO), and human serum for metabolomics was a mixture from four anonymous healthy donors collected from Stanford Blood Center (Palo Alto, CA). For the rheumatoid arthritis study, serum was collected from patients diagnosed with rheumatoid arthritis as well as individuals with no severe symptoms of RA. The handling of these biological materials must be performed in accordance with U.S. Department of Health and Human Services guidelines for Level 2 laboratory biosafety as found in *Biosafety in Microbiological and Biomedical Laboratories*, 4th Edition, HHS Publication No. (CDC) 93-8395. Affinity beads for albumin and IgG removal were from ProMetic Biosciences (Cambridge, UK). All other general reagents were purchased either from Fisher or VWR Scientific.

Serum (1 mL) was fractioned into serum proteome and serum metabolome using a 5-kDa molecular weight cut-off spin filter (Millipore Corp., Bedford, MA). Twenty-five microliters of the high molecular weight fraction (serum proteome) were diluted with 25 mM PBS buffer (pH 6.0) before it was applied to affinity beads (Prometic Biosciences, Cambridge, UK) for human serum albumin and immunoglobulin G (IgG) removal. The albumin- and IgG-depleted serum proteome was denatured by 6 M guanidine hydrochloride, reduced by 10 mM dithioreitol and alkylated with 25 mM iodoacetic acid/NaOH at room temperature. The denaturant and reduction–alkylation reagents were removed from the mixtures by buffer exchange against 50 mM $(NH_4)_2CO_3$ at pH 8.3 using a 5 kDa molecular weight cut-off spin filter (Millipore, Billerica, MA). Modified trypsin (Promega Corp.,

Madison, WI) of 1% weight equivalence of the proteins was then added to the mixtures with incubation at 37 °C. A total of 20 μg of material, equivalent to 1.5 μL of the starting serum was injected into the LC-MS in a 20 μL volume containing 0.1% formic acid. The serum metabolome was desalted with a C-18 SPE cartridge (Sep-Pak, Waters Corporation, Milford, MA). This sample, obtained from 100 μL of serum, and in a final 20 μL volume after desalting, was then injected into the LC-MS.

2.2. Instrumental

A binary HP 1100 series HPLC was directly coupled to a Micromass (Manchester, UK) LCT™ ESI-time-of-flight (TOF) mass spectrometer equipped with a nanospray source (New Objective, Woburn, MA). PicoFrit fused-silica capillary columns (5 μm BioBasic C18, 75 μm × 10 cm, New Objective, Woburn, MA) were run at a flow rate of 300 nL/min after flow splitting. An on-line trapping cartridge (Peptide CapTrap, Michrom Bioresources, Auburn, CA) allowed fast loading onto the capillary column. Gradient elution of the proteome sample was achieved using 100% solvent A (0.1% formic acid in H_2O) to 40% solvent B (0.1% formic acid in acetonitrile) over 100 min. Separation of the metabolome was performed with a gradient of 10–25% of solvent B in 40 min, followed by 25–90% solvent B in 30 min. The throughput for both proteome and metabolome analysis was 50 samples per week per instrument.

2.3. Quantification method

All data analysis uses the MassView™ software developed at SurroMed [11–13]. Data, stored as a list of peaks for every scan, undergoes baseline subtraction, smoothing and de-isotoping, i.e., an isotopic pattern assignment. Isotopic assignment is based on template matching [14,15]. After a variety of patterns associated with different charge states from 1 up to 5 or 6 (and different masses for the *m*/*z* region) are systematically examined, the peak's charge state is designated by the best fit. For a molecular ion to qualify as a peak after baseline correction, smoothing and de-isotoping, a threshold of typically 15 or 20 counts is required. This ensures that all the signals being tracked have substantial ion counts. A chromatographic peak is then built by linking together a series of consecutive scans that contain a signal at a given *m*/*z* with a window of ±0.10 Da, allowing an occasional scan with no detectable peak if the signal is low. The tallest peak in the chromatogram yields an intensity value for the molecular ion. A list of de-isotoped peaks is obtained for a given LC-MS run, each peak distinguished by its characteristic monoisotopic *m*/*z*, retention time, charge state and maximum intensity.

Retention times for each fileach file are 'time-warped' with respect to a reference file by first choosing a set of common peaks between the two files that have close proximity in *m*/*z* (±0.10 Da) and retention times (±3 min) [15]. Employing dynamic programming techniques, a warping function is derived that minimizes intensity differences at a given time and *m*/*z* between the two files [16]. A similar approach has been reported using only total ion chromatograms [17], but our approach of considering a large part of the data set is more effective.

Intensity normalization is performed by choosing one file as a reference and normalizing all other files one at a time. The single normalization constant for each file is taken as the median of the ratios of intensities for all components between the file in question and the reference file.

After peak lists are corrected by normalizing retention times and intensities, they are correlated between all samples by a process called 'component building'. If a peak from another sample is within the user-adjustable *m*/*z* and retention time windows, they are considered to represent the same component. Differential quantification is performed by comparing the intensity of the component between groups using a standard two-sided *t*-test or a non-parametric test, as appropriate [13]. In this manner a standard deviation can easily be calculated for each component. The variance of this measurement is defined as the square of the standard deviation, and the coefficient of variance is the ratio of the standard deviation and the mean value for a given group. The significance of any observed change can be determined by its *p*-value and significantly changing molecules can then be identified.

3. Results and discussion

Typical data from the LC-MS analysis of human serum proteome and metabolome fractions are significantly complex (Fig. 1). A proteome sample obtained from 1.5 μL of serum, after depletion of serum albumin and immunoglobulin G, and containing about 20 μg of serum proteins displays over 5000 molecular ions (not counting isotopes). A typical metabolome sample is obtained from 100 μL of human serum and yields approximately 1500 molecular ions after analysis. Considering an average of 3 isotopic peaks for each molecular ion, this amounts to over 15,000 isotopic peaks per sample.

3.1. Data analysis

The measurement of thousands of molecular ions per clinical sample and their comparison between large numbers of patients has been made possible by the development of the described computer application, MassView™ software (Section 2.3). As explained (Section 2.3), a list of de-isotoped peaks is obtained for each LC-MS run. Each molecular ion is characterized by its monoisotopic *m*/*z*, retention time, charge state and maximum intensity. The software allows us to visualize LC-MS data in the retention time and mass-to-charge ratio dimensions, as shown in Fig. 2 (from the data of Fig. 1).

Although mass calibration ensures accuracy of ±0.10 Da over long acquisition times, it is not unusual for a LC-MS

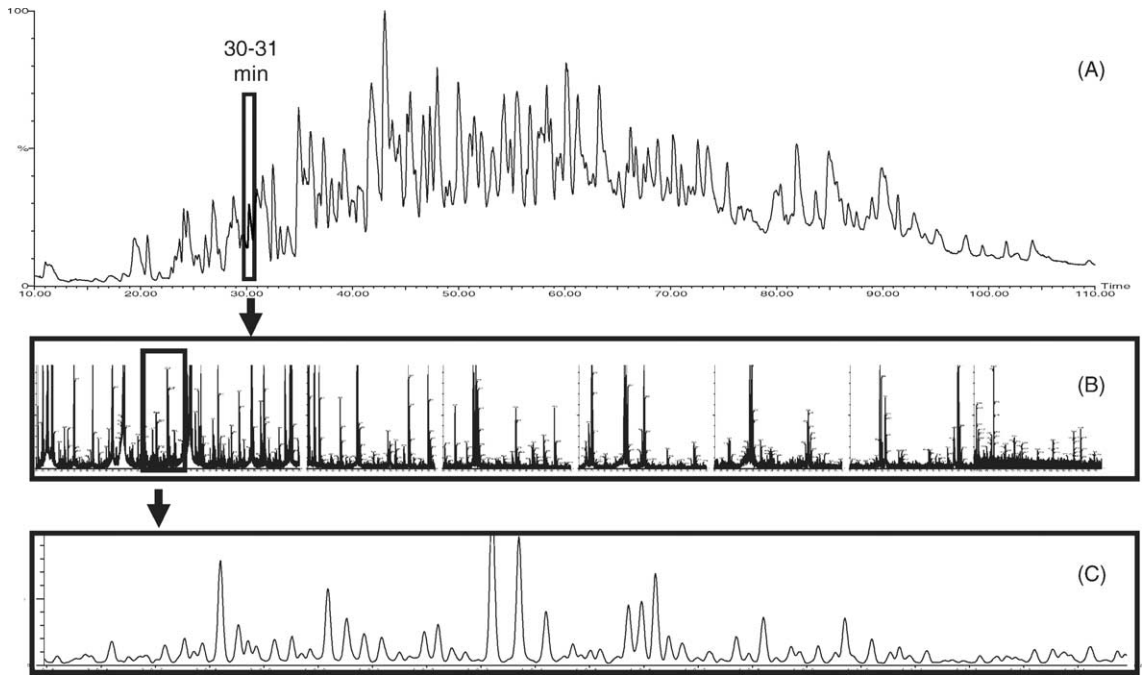


Fig. 1. The complexity of the human serum proteome sample (20 μg, from ∼1.5 μL of serum) after removal of abundant proteins is shown in an LC-MS run. (A) The base peak chromatogram. (B) A condensed mass spectrum integrated over 1 min of the run shows hundreds of peaks. (C) A magnified region spanning less than 20 Thompsons shows several molecular ions. Over 5000 such molecular ions can be detected in a typical run.

system to show non-linear chromatographic retention time drifts of up to ±3 min over the typical 100 min LC-MS run per sample. After dynamic time warping, we find that peak lists from different files are successfully corrected to within a half minute of each other's retention times.

Due to biological variations such as diet or water intake, as well as any drift over LC-MS sensitivity, there is a need to normalize individual serum sample analyte-intensities. A single normalization constant obtained from the median of intensity ratios (Section 2.3), as opposed to the mean, allows

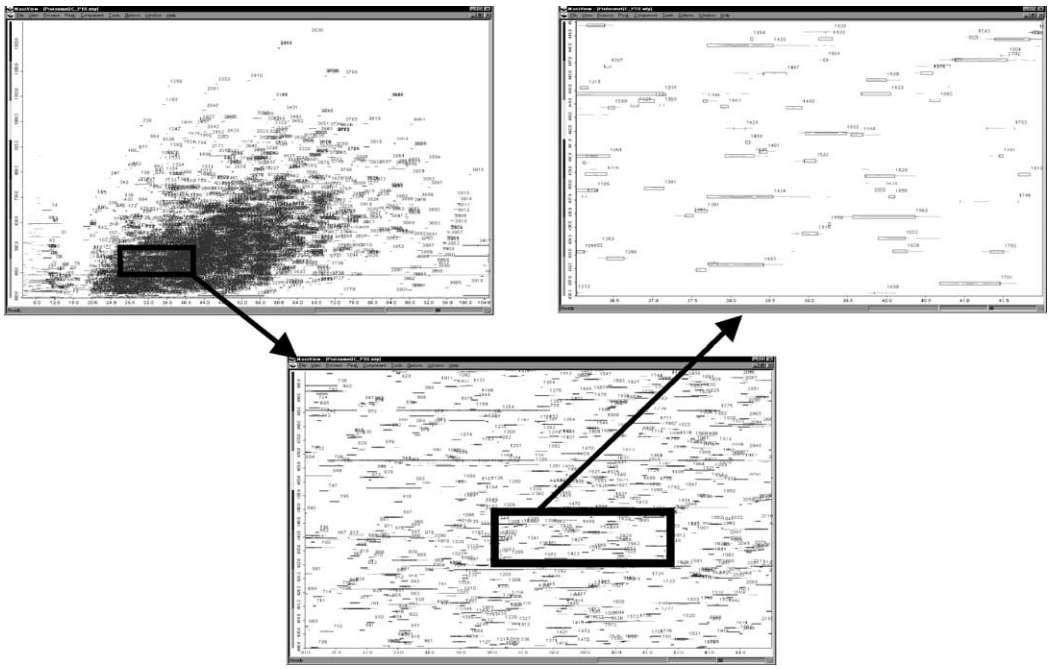


Fig. 2. A two-dimensional view of the same data shown in Fig. 1, de-isotoped, processed and displayed using SurroMed's proprietary MassView™ software. Retention time is shown on the *X*-axis and the *Y*-axis indicates mass-to-charge ratios. The same 5000 or more de-isotoped molecular ions shown in Fig. 1, with charge-states and intensities assigned, can be detected and quantified in a typical run.

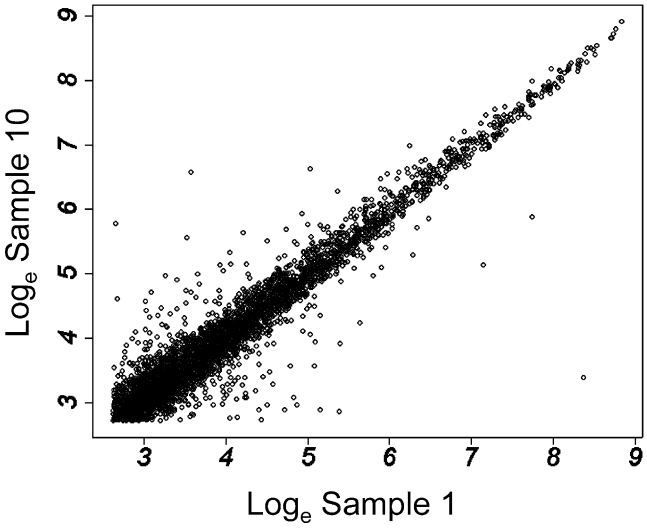


Fig. 3. This scatter plot shows the component intensities for two arbitrarily chosen samples. All intensities are log transformed.

for some strongly changing peaks such as putative biomarkers. Typical normalization factors are found to be between 0.80 and 1.20.

3.2. Validation of proteomic and metabolomic differential expression platform

As part of system validation, the reproducibility of the platform must be examined. Several steps of sample preparation can produce variation, namely, fractionation, albumin and IgG depletion, reduction, alkylation, buffer-exchange and tryptic digestion. In addition, contributions to measurement of variance could presumably come from the chromatographic separation process itself, from variability in sample injection volumes, and moreover from variations in ion suppression in the ionization process, as well as transmission and detection in the mass spectrometer. Our validation study of the platform included characterization of two encompassing sources of variability: sample preparation and the instrumental LC-MS analysis. The following experiment was devised to measure these two different and presumably independent general sources of variation.

To characterize first the contribution of LC-MS measurement to the total variance of our platform (independent of sample preparation), 20 samples were pooled after independently being prepared, and were realiquoted and analyzed on the LC-MS platform. MassView^{TM} software was used to quantify all molecular ions observed above an acceptable threshold (15 counts) for these 20 samples. An average coefficient of variance of 22.8%, and a median coefficient of variance of 15.1% was measured for the ∼5000 (de-isotoped) molecular ions measured per run in each of these 20 samples. Fig. 3 is a scatter plot comparing two arbitrarily chosen runs from this experiment. Due to effective normalization and robust LC-MS analysis, the data is close to, and scattered around the diagonal. Fig. 4A shows the distribution of CVs for these 20 samples, with intensity measurements made for 5000 molecular ions in each sample.

Secondly, 20 aliquots of the same, pooled human serum were *individually prepared* in parallel and subjected to LC-MS analysis. This measures the total variability in the platform, i.e., variability in sample preparation plus LC-MS measurement. The MassView^{TM} software was used for quantification of these 20 runs as well. An average coefficient of variation of 30.0%, and a median coefficient of variation of 25.4% was measured for the entire platform. Fig. 4B displays the distribution of CVs of normalized intensities for a total of approximately 5000 molecular ions from measurements on these 20 individually prepared proteome samples. Each histogram in Fig. 4 corresponds to a total of approximately 100,000 individual molecular ion measurements and more than 300,000 individual isotopic peak measurements. As far as we are aware, this is the largest number of intensity measurements reported to this date in any LC-MS study.

These measurements were then used to estimate the variability introduced by sample preparation alone. Assuming independence of the two sources of variation, which appears to be a reasonable assumption, i.e., the instrument variance and sample preparation variance, the total variance measured will be, $\sigma^2_{\text{total}} = \sigma^2_{\text{instrument}} + \sigma^2_{\text{sample\_preparation}}$, where σ is the standard deviation, defined as the square root of the variance,

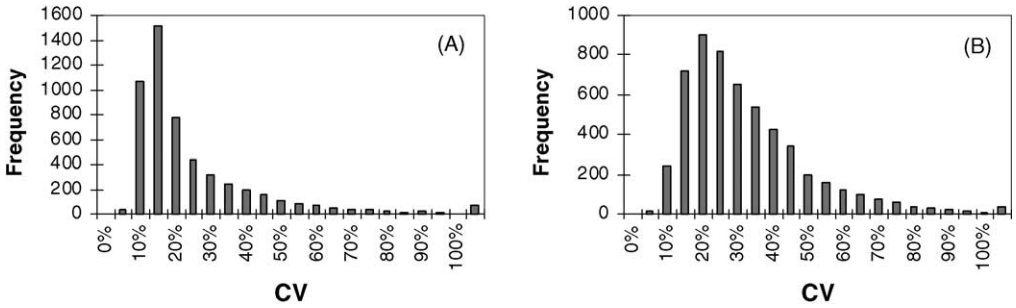


Fig. 4. The frequency distribution of coefficients of variation (CVs) for proteome sample preparation and LC-MS analysis. Reproducibility of LC-MS analysis of identical samples (A) and of individually prepared samples (B) is shown. After sample preparation, when the samples are pooled and then run 20 times on LC-MS, the median CV is 15.1% for ∼5000 molecular ions. When 20 identical samples are prepared individually using standard operating procedures, the median CV for a similar number of molecular ions is 25.4%. This indicates that the median CV for sample preparation of the proteome is 20.4%.

σ_{total} the total standard deviation, $\sigma_{instrument}$ the standard deviation of LC-MS instrumental analysis, and $\sigma_{sample\_preparation}$ the standard deviation of sample preparation alone. Using the already obtained values for variance of the total platform, and that of the LC-MS or instrument alone, the average CV of sample preparation for the proteome was calculated to be 19.5% and median CV is 20.4%. Thus, the individual contributions of sample preparation and the LC-MS analysis are comparable for the proteome.

For the metabolome (data not shown), similar CV distributions were obtained. More than 1500 molecular ions were quantified, with 16.4% (average) and 13.5% (median) CVs for variance of the platform, 30.9% (average) and 23.7% (median) CVs for individually prepared samples. Assuming the same independence of the two sources of variation, the average CV of metabolome sample preparation is 26.2% (average) and 19.5% (median).

These results indicate that the platform presented is able to quantify up to 5000 proteome molecular ions and more than 1500 metabolite molecular ions with overall median CVs of between 23 and 25% for the entire platform. As seen clearly in the histograms, several molecular ions are measured with great reproducibility, as low as 5 or 10% while others are measured with greater CVs up to 40% or more. Thus, we expect to be able to detect changes as small as 20% in some proteins or metabolites that are measured with very low CVs, and should overall be able to detect two-fold changes in almost any molecular ion.

3.3. Validation of dynamic range and sensitivity

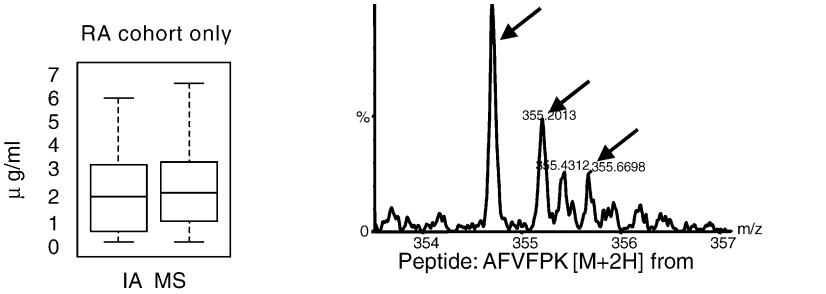
In addition to the ability to quantify molecular ions reproducibly, profiling as large a dynamic range as possible is important. Ability to quantify relatively low-intensity yet important metabolites and proteins may in fact be crucial for biomarker discovery. This application of our platform uses serum, known as one of the most challenging of body fluids or tissues for proteomic analysis due to its complexity and large dynamic range of analytes present [18]. To gain some depth in analysis, as described earlier, we remove the two most abundant proteins, human serum albumin and immunoglobulin G; these account for more than 80% of the total serum proteins [19]. Although limited in the dynamic range of our measurement by the capacity of the MS detector, we are able to observe a dynamic range of over 3 orders of magnitude (see Fig. 3). We also find that our CVs do not increase dramatically at lower intensities, enabling differential profiling of analytes with relatively low signal-to-noise ratios.

The sensitivity of this method and another validation by comparison with immuno-assay measurements is demonstrated in Fig. 5. For a group of individuals diagnosed with RA, the variation in C-reactive protein (CRP) concentrations is compared. C-reactive protein was observed between 1.0 and 3.0 μg/mL with considerable signal-to-noise ratio. The two methods agree not only in the range and mean intensities of CRP concentrations, but also in the effect size compared with a group of normal individuals.

After our platform was carefully characterized and validated for its reproducibility, dynamic range and sensitivity, we proceeded to conduct a differential profiling study in rheumatoid arthritis.

3.4. Differential profiling in rheumatoid arthritis

SurroMed is undertaking a longitudinal rheumatoid arthritis study that will enroll over 250 patients over three years and measure disease progression by routine clinical laboratory tests, cellular phenotyping as well as metabolic and proteomic profiling by this mass spectrometric method. We present here our initial results of a comparison of the metabolic and proteomic profiles of a subset of individuals enrolled in this study who are healthy (controls) and those diagnosed with RA.



	Ratio of Mean Value (RA/Normal)	Effect Size
IA	4.6	0.83
MS	4.1	0.88

Fig. 5. Comparison of mass spectrometry results with immunoassay measurements for C-reactive protein (CRP) in rheumatoid arthritis and normal subjects. Similar fold-changes and effect sizes are obtained from both immuno assay (IA) and mass spectrometry (MS). Low-concentration proteins from serum can be detected with significant signal-to-noise ratios, as seen in the detailed mass spectra for CRP. This data was collected on 20 μg of protein from the serum of a human subject and integrated over approximately 10 s elution time out of a total 100 min gradient LC-MS run. Effect size is the difference in means of the two groups divided by the pooled standard deviation.

Sera from 19 selected rheumatoid arthritis patients and 19 age- and sex-matched healthy controls were prepared as described before. The MassView™ software platform was able to distinguish isotopic patterns, build peak lists for a given sample, connect these peaks between all 38 samples and quantify each observed peak (de-isotoped molecular ion) in all samples being studied. More than 5000 molecular ions were observed in the proteome and ∼1500 molecular ions were observed in the metabolome of these 38 samples. Each molecular ion was quantified by measuring its signal intensity at the maxima of elution. The distribution of CVs for the proteome as well as the metabolome for both the control and RA groups is shown in Fig. 6. For the RA group, median CV of the distribution is 35.0% for the proteome and 29.9% for the metabolome. For the 19 normal individuals, the median CV is 33.8% for the proteome and 29.0% for the metabolome.

Before we consider analysis of differences between the diseased and normal population, it is interesting to try and estimate the biological variation observed in this group of 19 normal individuals. Assuming, quite fairly, that our sample preparation and LC-MS analysis methods are independent of biological diversity inherent in these individuals, we can estimate the biological variation, since variance measured in a population of normal individuals, $\sigma^2_{\text{normals}} = \sigma^2_{\text{instrument+sample\_preparation}} + \sigma^2_{\text{biological\_variation}}$. This implies that biological variation, as measured over 19 healthy donors using median CVs, is 22.3 and 16.7%, and using average CVs, is 24.0 and 14.9% for the proteome and metabolome, respectively. In the case of this group of RA patients, biological variation is only slightly larger. Table 1 summarizes these results along with results on CVs calculated for platform validation.

For differential profiling, peptides and metabolites that undergo significant relative intensity (concentration) changes between the RA and control cohort, as judged by a statistical *t*-test or non-parametric test, were tracked by their component number, mass-to-charge ratio, retention time and *p*-value. Of the ∼5000 peptide molecular ions compared, 95 showed differences at $p < 0.001$, where only five would be expected by chance, assuming independence between components. Table 2 summarizes the number of significant changes observed for a given *p*-value.

Differential profiling results in RA are visualized in Fig. 7. All 409 components (molecular ions) observed to change with a significance of $p < 0.01$ between the normal and RA group proteome are stacked horizontally. The plot uses a "*Z*" score for scaling, where $Z = (X_i - \langle X_i \rangle)/\sigma$, X_i is the individual measure and $\langle X_i \rangle$ the average of all measures divided by the standard deviation, σ. Cells in one row correspond to the same molecular ion. Columns correspond to individuals within the two panels of the RA and normal group. The color of a cell in any given row or column is the *Z* score calculated for a molecular ion represented by that row and for the individual corresponding to that column. The RA group can be easily distinguished from normals in having a very different pattern of expression of proteins. Individual profiles can be observed clearly in each column representing that individual, for both the RA and the control cohorts. Within the RA cohort, individuals have slightly different patterns, which will

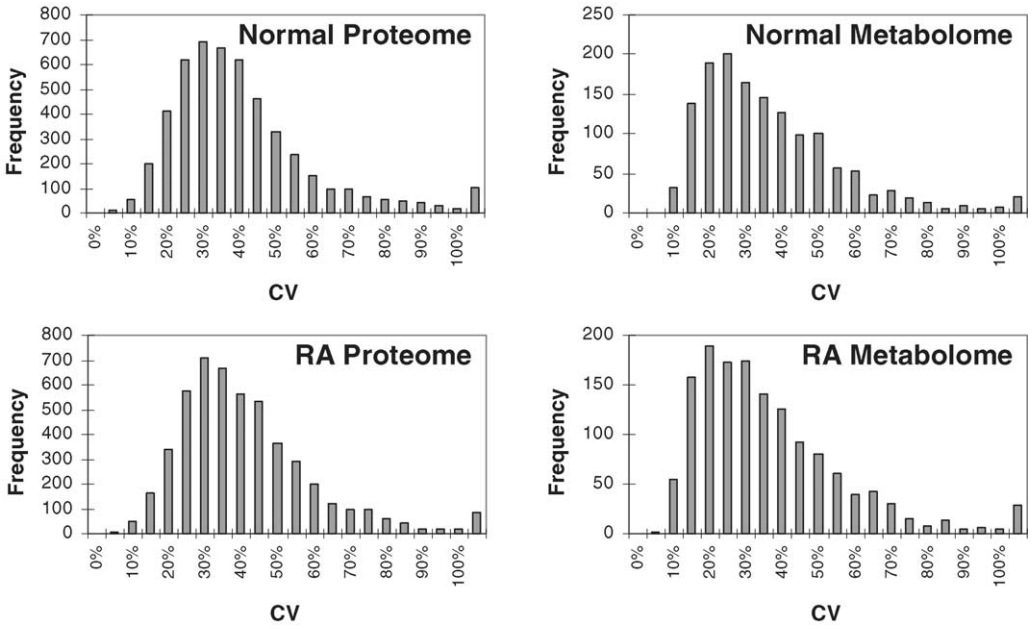


Fig. 6. The frequency distribution of coefficients of variation (CVs) for the proteome and metabolome from normal individuals (top panel) and individuals diagnosed with rheumatoid arthritis (RA; bottom panel). For 19 normal individuals, the median CV is 33.8% (proteome), and 29.0% (metabolome). For 19 individuals diagnosed with rheumatoid arthritis, the median CV is 35.0% (proteome) and 29.9% (metabolome). More than 5000 proteome molecular ions and ∼1500 metabolome molecular ions (de-isotoped) are differentially quantified between these 19 normal and 19 RA samples.

Table 1
Summary of raw and derived %CV statistics for all studies

	# Samples	Proteome %CV		Metabolome %CV	
		Average	Median	Average	Median
Instrumental (LC-MS)	20	22.8	15.1	16.4	13.5
Instrumental preparation (sample preparation) + instrumental	20	30.0	25.4	30.9	23.7
Derived sample preparation alone	20	19.5	20.4	26.2	19.5
RA cohort (biological + sample preparation + instrumental)	19	39.0	35.0	34.5	29.9
Derived biological variation (RA cohort)	19	24.9	24.1	15.3	18.2
Normal cohort (biological + sample preparation + instrumental)	19	38.4	33.8	34.3	29.0
Derived biological variation (normal cohort)	19	24.0	22.3	14.9	16.7

A total of 5000 molecular ions were quantified from the proteome and ∼1500 molecular ions for the metabolome. Average and median coefficients of variation (%CV) are listed (CV is defined as the ratio of standard deviation to the mean for each molecule for a given group).

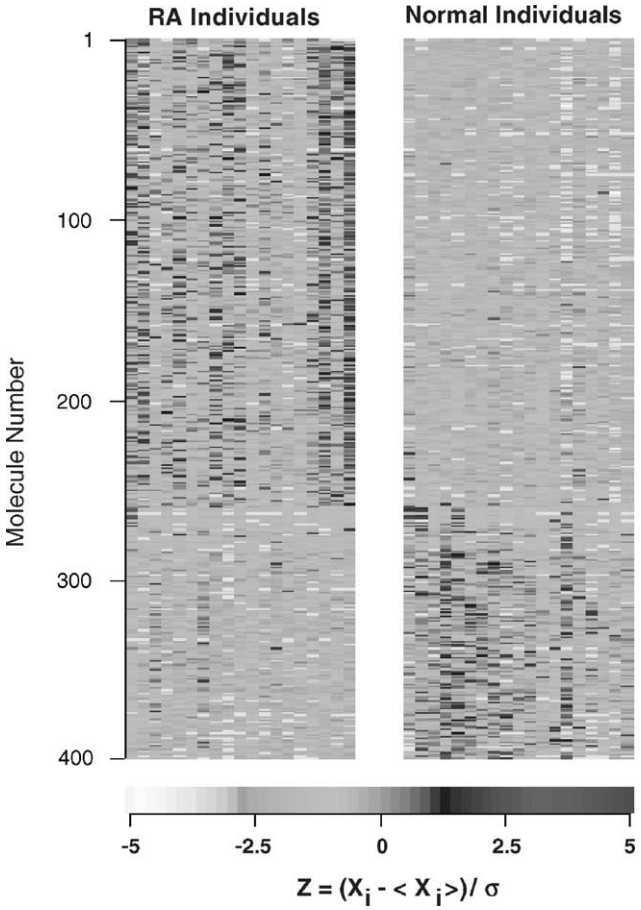


Fig. 7. A heat map (in grayscale) demonstrating differential profiling of proteins in 19 individuals with RA (left) and 19 normal individuals (right). Cells in one row correspond to the same component. Columns correspond to individuals. Each cell is assigned a color corresponding to its *Z*-score. The *Z*-parameter is defined as the difference between the individual measure X_i and the average of all measures $\langle X_i \rangle$, divided by the standard deviation σ; $Z = (X_i - \langle X_i \rangle)/\sigma$. Each component (for RA and controls together) is scaled to zero mean and unit variance in order to apply a mapping to a grayscale. A vertical ordering based on the size of the effect, taking sign into account was performed prior to plotting. Normal individuals clearly show a different pattern of their proteomic profile from RA patients. In addition, different individuals show slightly different patterns within a given group.

be correlated to degrees of severity of disease in future work. It is interesting to note that approximately two-thirds of the most significant differences seen arise from overexpression of particular molecules in the RA group. These ions are not observed to overexpress in even one of the 19 normal individuals.

The metabolome data also shows significant differences between the two groups. Of the ∼1500 molecular ions measured, 62 showed significant differences with a *p*-value of 0.05 or less. We observe more molecular ions with significant changes in the proteomic profile than in the metabolome profile for this comparison of diseased and normal individuals. It is noteworthy that although more molecular ions are measured in the proteome, the proteome sample contains several peptides per protein due to tryptic digestion and sometimes more than a single charge state of the given peptide is observed.

Once significantly changing molecules are marked, they can be targeted for identification by subsequent tandem mass spectrometry (MS/MS) analysis. For peptides, results are searched against protein or DNA expressed sequence tag (EST) databases [20,21] using a commercial program such as SEQUEST™ (ThermoFinnigan, Inc.) or Mascot™

Table 2
Summary of statistics from differential profiling of the proteome of normal individuals and those diagnosed with RA

# Molecular ions quantified	5025
%CV (average)	
Normal group	38.4
RA group	39.0
%CV (median)	
Normal group	33.8
RA group	35.0
# Significant changes at	
$p < 0.001$	95
$p < 0.005$	265
$p < 0.01$	409
$p < 0.05$	908

A total of ∼5000 molecular ions are quantified from the proteome. Of these, 908 molecular ions change significantly (p value < 0.05).

(Matrix Science), or alternatively, de novo sequencing of the MS/MS spectra can be used with a variety of commercial software. We have so far identified more than 1500 of these molecular ions. Several changes observed in RA are as expected from previous literature. This includes overexpressed alpha-1-antitrypsin, alpha-1-acid glycoprotein, alpha-2 glycoprotein, haptoglobin, alpha-1-antichymotrypsin and ceruloplasmin. In addition, the RA group shows a decrease in the concentrations of H2 polypeptide, transferrin and retinal-binding protein, when compared with the normal group. We have also observed proteins not previously identified with RA that are undergoing further clinical validation.

4. Conclusions

Biomarkers are important molecular signatures of the phenotype of an organism that aid in early disease detection and drug response, and may serve as viable targets for designing drugs. Advances in mass spectrometry techniques and the inherent ability to detect a multitude of analytes present in a complex biological matrix at a given time by virtue of their different molecular masses and retention times have clearly made LC-MS a practical platform for biomarker discovery, including investigations that are not hypothesis driven. In addition to improving mass spectrometric technologies for identifying proteins and metabolites present in a given biological fluid, technologies such as presented here that allow accurate quantification of detected molecular ions, are essential for biomarker discovery.

Toward this effort, we have presented a platform for differential quantification of proteins and metabolites that is generally applicable to all complex biological fluids and tissues. The platform has been validated for the human serum proteome and metabolome by measuring the independent contribution of sample preparation and LC-MS analysis to the total variance. We have quantified biological diversity as observed in 19 healthy individuals by comparing their serum proteome and metabolome profiles. Above all, we have demonstrated the capability of differential expression profiling of serum proteins and metabolites for a clinical study.

Acknowledgements

We are grateful to Sophia Chen and Lander Hill for sample preparation. We thank Praveen Kumar for his many contributions to MassView^{TM} MS software. We are also grateful to Weixun Wang for identifying C-reactive protein, and to Haihong Zhou, Thomas Shaler, Jeffrey Satkofsky, Gary Frenzel and Jennifer Thompson for their many contributions to the SurroMed Proteomics and Metabolomics Platform.

References

[1] E. Stein, Am. J. Cardiol. 87 (Suppl.) (2001) 21A.
[2] V. Murthy, A. Karmen, J. Clin. Lab. Anal. 11 (1997) 125.
[3] N.L. Anderson, N.G. Anderson, Electrophoresis 19 (11) (1998) 1853.
[4] E. Brockstedt, M. Peters-Kottig, V. Badock, C. Hegele-Hartung, M. Lessl, Endocrinology 141 (2000) 2574.
[5] J. Fenn, M. Mann, C.K. Meng, S.F. Wang, C.M. Whitehouse, Mass Spectrom. Rev. 9 (1) (1990) 37.
[6] T.R. Covey, E.D. Lee, A.P. Bruins, J.D. Henion, Anal. Chem. 58 (14) (1986) 1451A.
[7] S.P. Gygi, B. Rist, S.A. Gerber, F. Turecek, M.H. Gelb, R. Aebersold, Nat. Biotechnol. 17 (1999) 994.
[8] J. Ji, A. Chakraborty, M. Geng, X. Zhang, A. Amini, M. Bina, F. Reignier, J. Chromatogr. B 745 (2000) 197.
[9] G. Cagney, A. Emili, Nat. Biotechnol. 20 (2002) 163.
[10] C. Muller, P. Schafer, M. Stortzel, S. Vogt, W. Weinmann, J. Chromatogr. B 773 (2002) 47.
[11] W. Wang, H. Zhou, H. Lin, S. Roy, T.A. Shaler, L. Hill, S. Norton, P. Kumar, M. Anderle, C.H. Becker, Anal. Chem. 75 (2003) 4818.
[12] C.H. Hastings, S.M. Norton, S. Roy, Rapid Commun. Mass Spectrom. 16 (2002) 462.
[13] M. Anderle, S. Roy, H. Lin, C. Becker, K. Joho, Bioinformatics (2004) in press.
[14] R. Brunelli, T. Poggio, Pattern Recognit. 30 (1997) 751.
[15] H. Sakoe, C. Chiba, IEEE Trans. Accoust. Speech Signal Process ASSP-26 (1978) 43.
[16] L. Rabiner, B.-H. Juang, Fundamentals of Speech Recognition, Prentice-Hall Publishers, Englewood Cliffs, NJ, 1993 (Chapter 4).
[17] C.P. Wang, T.L. Isenhour, Anal. Chem. 59 (1987) 649.
[18] N.L. Anderson, N.G. Anderson, Mol. Cell. Proteomics 2 (1) (2003) 50.
[19] K. Diem, C. Lentner, Scientific Tables, Geigy Pharmaceuticals, Division of Ciba-Geigy Corporation, Ardsley, NY, 1970, p. 580.
[20] J.K. Eng, A.L. McCormack, J.R. Yates, J. Am. Soc. Mass Spectrom. 5 (1994) 976.
[21] A. Link, J. Eng, D.M. Schieltz, E. Carmack, G.J. Mize, D.R. Morris, B.M. Garvik, J.R. Yates, Nat. Biotechnol. 17 (1999) 676.