RESEARCH ARTICLE

ReSASC: A resampling-based algorithm to determine differential protein expression from spectral count data

Kristina M. Little[1,2], Jae K. Lee[3] and Klaus Ley[1]

[1] Division of Inflammation Biology, La Jolla Institute for Allergy and Immunology, La Jolla, CA, USA
[2] Department of Biomedical Engineering, University of Virginia, Charlottesville, VA, USA
[3] Department of Public Health Sciences, University of Virginia, Charlottesville, VA, USA

Received: May 18, 2009
Revised: July 27, 2009
Accepted: September 29, 2009

Label-free methods for MS/MS quantification of protein expression are becoming more prevalent as instrument sensitivity increases. Spectral counts (SCs) are commonly used, readily obtained, and increase linearly with protein abundance; however, a statistical framework has been lacking. To accommodate the highly non-normal distribution of SCs, we developed ReSASC (resampling-based significance analysis for spectral counts), which evaluates differential expression between two conditions by pooling similarly expressed proteins and sampling from this pool to create permutation-based synthetic sets of SCs for each protein. At a set confidence level and corresponding *p*-value cutoff, ReSASC defines a new *p*-value, p', as the number of synthetic SC sets with $p > p_{cutoff}$ divided by the total number of sets. We have applied ReSASC to two published SC data sets and found that ReSASC compares favorably with existing methods while being easy to operate and requiring only standard computing resources.

Keywords:
Bioinformatics / Differential protein expression / Label-free / Spectral counts / Statistical models

1 Introduction

LC coupled to MS/MS is the leading method used to determine the protein composition of complex biological samples [1]. Increasingly, this technology has been employed for biomarker discovery research, wherein the protein composition of cell or bodily fluid samples from a group of patients with a given disease or condition of interest (henceforth termed ''cases'') is compared to that from matched healthy ''controls'' [2–4]. The goal of this type of experiment is to effectively detect and quantify differences between samples. While a single MS experiment is capable of identifying thousands of proteins – perhaps dozens or hundreds of which may be up- or down-regulated between the two groups – the lists of possible candidates must be further tested by alternate methods to confirm statistically significant differences [5]. Currently stable-isotope labeling (ICAT [6] and iTRAQ [7]) is widely used to measure relative protein abundance [8]; however, there are several difficulties associated with its use, including increased costs and time required for sample processing, as well as reduced sensitivity of protein and peptide detection [9].

In response to these difficulties, label-free methods have been developed to quantify differential expression in MS experiments. Two of the most relevant label-free methods are the comparison of peptide ion peak current area and spectral counting, both of which have been shown to correlate well with protein abundance [10, 11]. Spectral counts (SCs) – the total number of MS/MS spectra matched to a given protein – have the advantage that they are theoretically and computationally simple, and that they are provided directly by database searching algorithms such as

Correspondence: Dr. Klaus Ley, Division of Inflammation Biology, La Jolla Institute for Allergy and Immunology, 9420 Athena Circle, La Jolla, CA 92037, USA
E-mail: klaus@liai.org
Fax: +1-858-752-6986

Abbreviations: LPE, Local-pooled-error; **LSPAD,** localized statistics of protein abundance distribution; **ReSASC,** resampling-based significance analysis for spectral counts; **SC,** spectral count; **SpI,** spectral index

© 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

SEQUEST ([12]; Thermo Electron, Waltham, MA, USA) or MASCOT [13]. On the other hand, their very nature makes statistical analysis difficult: SCs are discrete, and their underlying distribution is not normal – in fact, it is highly skewed due to the overwhelming number of zero entries (*e.g.* when a protein is not found in a given MS experiment or in a given condition). Consequently, standard parametric tests such as Student's *t*-test are inappropriate for this type of data, especially when the number of runs is small [14].

Many experiments that employ SCs as a means for relative quantification rely on simple fold cut-off criteria, either with or without some sort of normalization procedure [15, 16]. Although unsuitable for the distribution of SCs, some publications have used *t*-tests or other parametric statistical tests [17]. Some groups join SCs with other label-free metrics such as the number of unique peptides detected *per* protein or the SEQUEST-based *XCorr* parameter [18]. With the recent increased interest in novel statistical measures geared specifically toward SC data, several new approaches have been developed [19–21]. Choi *et al.* introduced a test based on Bayes estimation of generalized linear mixed effects models, known as QSpec [22], while Fu *et al.* have proposed a new metric, termed the spectral index (SpI) that combines a protein's SC with the number of runs in which the given protein was detected [23].

QSpec employs a statistical method known as hierarchical Bayes. The Bayes factor is the ratio of the average likelihoods of two models. In the case of QSpec, this describes the ratio of a model that includes a differential expression term, M_2, to a model that does not include such a term, M_1. So for a set of SCs, X, from a given protein, the Bayes factor $B(X)$ is defined as

$$B(X) = \frac{p(X|M_2)}{p(X|M_1)} = \frac{\int p(\boldsymbol{\Theta}_2|M_2)p(X|\boldsymbol{\Theta}_2, M_2)\mathrm{d}\boldsymbol{\Theta}_2}{\int p(\boldsymbol{\Theta}_1|M_1)p(X|\boldsymbol{\Theta}_1, M_1)\mathrm{d}\boldsymbol{\Theta}_1} \quad (1)$$

where $p(X|M_i)$ is the averaged likelihood for model i. Therefore, a large Bayes factors (*e.g.* greater than 10) implies that M_2 better describes the data, indicating statistically significant differential expression for a given protein (for more details on Bayesian statistics, see [24]). Because the Bayes factor can be overestimated in the case of large SCs – *e.g.* high-abundance proteins – the authors also imposed a minimum SC-based fold-change requirement of 0.5 between conditions. Additionally, the authors also determined a ''minimum threshold'' Bayes factor, B^*, to control the global false discovery rate at approximately 5%.

The SpI is based on two metrics: a protein's relative abundance as measured by SCs across repeated replicates, and its reproducibility as measured by the number of replicates for each condition in which the protein was detected. A protein's SpI is defined as

$$SpI = \left(\frac{\bar{S}_D}{\bar{S}_D + \bar{S}_C} \times \frac{N_D^D}{N_D^T}\right) - \left(\frac{\bar{S}_C}{\bar{S}_D + \bar{S}_C} \times \frac{N_C^D}{N_C^T}\right) \quad (2)$$

where S_D and S_C correspond to the mean SC values for the disease state and control state, respectively, and N^D and N^T correspond to the number of runs in which a protein was detected in the given state and the total number of runs conducted for the given state, respectively. The SpI can range from -1 to $+1$, with values close to -1 or $+1$ indicating that a protein is down- or up-regulated in the disease sample relative to the control sample. Random permutation analysis is performed to create a null distribution of SpI's, from which an appropriate SpI cutoff is chosen corresponding to the desired confidence level.

A reproducibility study by Durr *et al.* performed on rat lung endothelial cell plasma membranes concluded that ten replicate runs would be necessary to reach completeness of protein detections with 95% confidence [25]. These results show that SCs reflect the variability of MS as a technique. Factors that influence variability include a protein's mass, hydrophobicity, cellular location, and number of trypic peptides, among others; machine-based factors such as operation in data-dependent mode and exclusion window; and processing parameters such as which PTMs are allowed in database searches [26]. As such, low-abundance proteins may or may not be detected in a given MS run, and higher abundance proteins are detected at varying levels in different levels from run to run.

Here we propose a novel statistical method, called ReSASC (resampling-based significance analysis for spectral counts) for SC-based LC-MS/MS analysis. ReSASC makes only one assumption about SC data: its variability. ReSASC is based on the idea that if a protein is truly differentially expressed between two conditions, then that differential expression would be maintained if the samples were to be re-analyzed. For instance, if ten samples were run in each of two conditions, and then subsequently re-run, one would expect that truly differentially expressed proteins would be detected in both analyses. Obviously the SC values from the new experiment would vary slightly from the previous one, but ReSASC assumes that the pattern of differential expression is maintained. Conceptually, ReSASC is comparable to local-pooled error (LPE; [27]) and Rank Products [28], both of which were introduced for microarray analysis. Like LPE, ReSASC pools protein expression data based on their run-to-run variability; like rank products, ReSASC uses permutation analysis to estimate the significance of a protein's apparent up- or down-regulation. Effectively, ReSASC samples similarly expressed proteins to create synthetic experiments and then compares the SC from these experiments to see if the initially detected statistical differences are conserved. We compare ReSASC to QSpec and the SpI, and show that these three SC-based methods perform comparably.

2 Materials and methods

2.1 Data sets

Five data sets were analyzed by ReSASC: three for validation and two for comparison to existing methods. The first

© 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

(hereafter referred to as ''*replicates*'') consists of seven replicates of human plasma analyzed over seven non-consecutive days. Human blood was collected from one healthy donor at a single time point *via* venipuncture into one-tenth volume of ACD (85 mM trisodium citrate, 83 mM dextrose, and 21 mM citric acid) solution. Platelet poor plasma was obtained by centrifugation and residual cells were removed. Proteins were electrophoresed approximately 1 cm into a 7.5% acrylamide SDS-PAGE using a Mini-gel system (Bio-Rad Laboratories, Hercules, CA, USA) at 150 V. The acrylamide gel section containing the proteins was cut out and placed in fixative (50% methanol, 12% acetic acid, and 0.05% formalin) for 2 h. The in-gel tryptic digestion of the lanes and the peptide extraction were performed as described [29]. The extracted peptide solutions were lyophilized and reconstituted to 20 μL with 0.1% acetic acid for MS analysis. Two-microgram samples were loaded onto a 360 μm od × 75 μm id microcapillary fused silica precolumn packed with irregular 5–20 μm C18 resin. The samples were gradient eluted at a flow rate of 60 nL/min with a Surveyor binary HPLC solvent delivery system (Agilent, Palo Alto, CA, USA) directly through an ESI source interfaced to an LTQ-FT ion trap mass spectrometer (Thermo Electron). The LTQ mass spectrometer was operated in data-dependent mode in which an initial MS scan recorded the m/z values of ions over the mass range 300–2000 Da, and then the ten most abundant ions were automatically selected for subsequent MS/MS analysis. All MS/MS data were searched against a human protein database downloaded from the European Bioinformatics Institute (http://www.ebi.ac.uk) using SEQUEST. A static modification of 57.02150 Da for cysteine residues and variable modifications of 15.9949 and 14.01550 for methionine residues and cysteine residues, respectively, were allowed. The parent mass tolerance was set to 10 ppm and the mass tolerance of daughter ions was set at 0.5 Da. For the human plasma replicates, peptide identifications were made based on fully tryptic peptides, using a first-pass filtering of standard criteria as previously described [30], including cross correlation values of 2.0, 2.2, and 3.3 for charged states of +1, +2, and +3, respectively. Higher charged states were not considered. Proteins were also required to be detected at least twice ($SC \geq 2$) across all runs.

The second and third data set are simulated data sets provided by Choi *et al.* [22]. Four biological replicates of yeast strain BY 4741 were grown in different media, either ^{14}N or ^{15}N [31]. Cells were collected, washed, soluble proteins extracted by centrifugation, desalted by trichloroacetic acid precipitation, urea-denatured, reduced, alkylated, and digested by endoproteinase Lys-C and by modified trypsin. For each replicate, 500 μg of protein extract were analyzed by three-phase LC coupled to an LTQ linear ion trap equipped with an ESI source (Thermo Electron). A 12-step MudPIT [32] run was performed with the spectrometer operating in data-dependent mode such that the five most abundant ions were chosen for subsequent MS/MS analysis. Tandem mass spectra were searched against the NCBI *Saccharomyces cerevisiae* database by SEQUEST, allowing a mass tolerance of 3 amu for precursor ions and 0 amu for fragment ions. No variable modifications were searched. Detected peptides were thresholded by the following criteria: $\Delta Cn \geq 0.1$; minimum XCorr of 1.5, 2.5, and 3.0 for charged states of +1, +2, and +3, respectively, and maximum Sp rank of 10. As this preparation would not be expected to result in any differential protein expression, Choi *et al.* created two synthetic data sets by randomly shuffling the rows of the data set and inserting either a twofold or fourfold increase to the SCs of the first 200 proteins from the ^{14}N samples. (Hereafter, these data sets are referred to as ''synthetic 2-fold'' and ''synthetic 4-fold,'' respectively.)

The fourth data set (hereafter referred to as ''yeast'') consists of four replicate LC-MS/MS runs of yeast strain BY4741 at two different stages of cell growth: log phase and stationary phase [31]. Samples were processed and analyzed in the same manner as the above simulated data sets. The third replicate from the log phase was found to be a statistical outlier that could not be corrected by normalization procedures and thus was not included in the analysis.

The final data set (hereafter referred to as ''cystic fibrosis'') consists of broncheoalveolar lavage (BAL) fluid obtained from eight patients with cystic fibrosis and four healthy controls [23]. Protein samples were reduced, alkylated, and digested overnight with trypsin. After concentration, desalting, drying and re-suspension, the tryptic digests were separated using 2-D HPLC (Thermo Electron), and analyzed by an LCQ Deca XP+ion trap mass spectrometer (Thermo Electron) equipped with an ESI source. The spectrometer was operated in data-dependent mode with the three most abundant ions being selected for subsequent MS/MS analysis. Dynamic exclusion criteria allowed repeated selection of the same precursor ion twice within a 30-s window, followed by a 72-s exclusion period. SEQUEST was used to search all identified tandem mass spectra against the Human International Protein Index, allowing modified thiol residues and at most one incomplete cleavage site. For a protein to be considered identified, it must have been detected by at least two unique peptides in at least one sample.

2.2 Pre-filtering

Before applying ReSASC, the data were filtered by requiring a protein to be detected at least once on average in at least one of the conditions. This filtering criterion is of comparable stringency as the two-unique peptide minimum criterion, but is less sensitive to non-reproducible outliers. Of 649, 1508, and 641 proteins identified in the human plasma, yeast, and BAL fluid, respectively, 228, 906, and 560 passed this filtering criterion.

© 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

2.3 Statistical analysis

ReSASC was coded and run in Matlab version R2007b (The MathWorks, Natick, MA, USA). Correlation coefficients correspond to Spearman rank coefficients. All values are listed as median (inter-quartile range) unless otherwise stated. The ReSASC p_{cutoff} values were based on $k = 0.95$. For ReSASC analysis, a p'-value of 0.05 was considered significant.

3 Results

3.1 ReSASC

ReSASC is conceptually related to the LPE test, which estimates overall variance of in a gene expression microarray experiment by pooling similarly expressed genes. A standard Bland-Altman plot for a gene microarray experiment is shown in Fig. 1A, with the corresponding LPE-estimated variance across the range of gene expression levels shown in Fig. 1B. A comparable Bland-Altman for a typical proteomics SC data set (Fig. 1C) exemplifies the difficulties in extrapolating LPE to this type of data. There are far fewer proteins detected in a MS experiment compared to a microarray experiment (by approximately one order of magnitude), and there is distinct aliasing in the plot due to the discrete nature of SC data. These issues combine to make variance estimation by LPE difficult. Like LPE, ReSASC also attempts to estimate pooled properties, but is permutation-based and non-parametric. ReSASC is a novel procedure that is used to simulate LC-MS/MS experiments and then, based on those simulations, to determine if a protein is truly differentially expressed between conditions. A flow chart for the ReSASC logic is shown in Supporting Information Fig. 1. For each condition, ReSASC determines a protein's median value and "scatter" values, S_i, that describe the range around its mean value, μ_p (e.g. for each SC value, x_j, of a given protein, p, $S_{p,x_j} = x_j/\mu_p$). Similar to SpI in [23], these values estimate a protein's abundance level and reproducibility between runs. Next, for each unique median value, a sampling window is determined incrementally. This process is depicted in Fig. 2A and B. The purpose of this incremental expansion is to determine a group of similarly expressed proteins from which hypothetical SCs can be sampled. A radius, r_i, around the given median value, M_i, is chosen to include the closest neighboring median value (*e.g.* $r_i = \min\{(M_i - M_{i-1}),(M_{i+1} - M_i)\}$). If there are no outlier S values within this window – when an outlier is defined as lying beyond one SD from either the first or third quartiles of the data – then the radius is incrementally increased in the same way described previously. This process is repeated until there is an outlier value, at which point the immediately prior radius is kept. In this way, each median value contains multiple proteins, all similarly scattered around their average values. Figures 2C and D give a graphical depiction of this data with multiple boxes that show the radius for various median values.

Because few LC-MS/MS experiments are based on ten or more runs, it is difficult to be confident that any given protein is absent based on the limited data available. Indeed, a major caveat to consider when using SC data is the difference between detection and identification, that is, differentiating between proteins that are present in the sample but not detected by MS and proteins that are not present, both of which will have SCs of zero. To address this issue, the SCs for

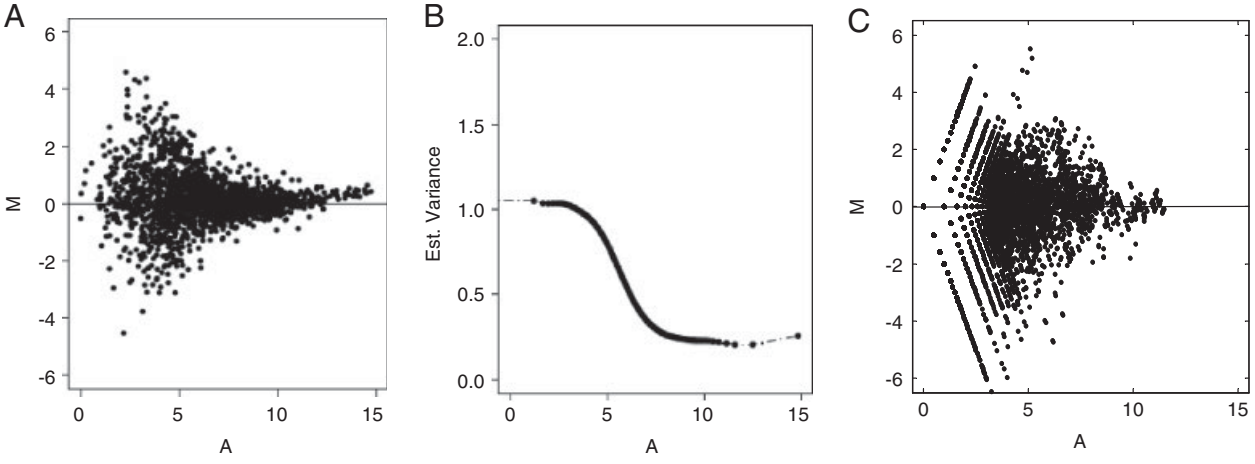


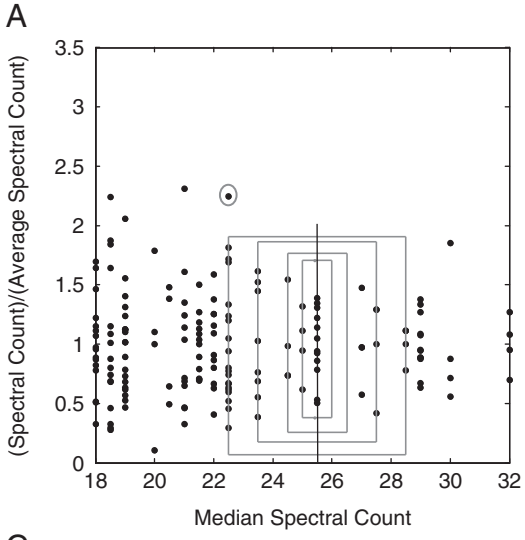
Figure 1. (A) Typical M-A plot based on Affymetrix gene chip data [27]. (B) Experiment-wide variance estimate based on data in (a) determined by local-pooled error (LPE, [27]). $A = \log_2(\text{expression average}) = \log_2\sqrt{(x_1 x_2)}$; $M = \log_2(\text{ratio}) = \log_2(x_1/x_2)$. (C) SC-based proteomics M-A plot, using data from [22]. Note that for SC data, $x = (\text{SC}+1)$ to eliminate zero values. Due to the small number of proteins detected in a MS experiment (1508 compared to 12 488 genes in (A)), the plot is underpopulated, making variance estimation difficult. The discrete nature of SC data also causes aliasing, which hampers normalization difficult and makes statistical analysis by LPE impossible. Note that the dynamic range of SC-based proteomics is approximately 1/8th that of genomic data (2^{12} *versus* 2^{15}).

© 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

all proteins with at least one zero-value SC are compiled and the data is modeled by a continuous distribution. MATLAB's non-linear least-squares curve fitting algorithm was used to fit the data to multiple distributions and calculate the squared 2-norm ($\sum f[(x)-y(x)]^2$, where $f(x)$ is the least-squares fitted function[x] evaluated at x and $y(x)$ is the actual relative frequency of SC = x in the compiled data) of the residuals from each fit. Based on this analysis, it was determined that the data is best approximated by an exponential distribution (see Fig. 3A and B).

From the measured SC data, 100 synthetic data sets were created as shown below. Preliminary experiments showed that taking $n = 100$ synthetic sets yields results that are highly correlated to those based on larger n (data not shown). In all data sets tested, the Spearman correlation coefficient between the results for $n = 100$ and $n = 1000$ was greater than 0.90, with $p < 0.001$. In each of these synthetic data sets, the SCs for proteins with non-zero median value are determined by sampling from within the appropriate radius window. Figure 4A shows the number of proteins whose SCs were sampled at each unique median value. For proteins with zero median value, a random sample was taken from the fitted exponential distribution. For each set of synthetic SCs, the non-parametric Wilcoxon rank-sum test was performed between the two groups, resulting in 100 p-values *per* protein. One important feature of the ReSASC method is that it allows synthetic experiments to consist of more runs than the original experiment. To increase the power of the non-parametric Wilcoxon test, ReSASC's synthetic data sets are populated to a minimum of ten runs in each condition.

To determine the appropriate p-value to be considered, a null-distribution of p-values was produced. The runs from the different conditions (cases and controls) were randomly permuted and for each permutation, a Wilcoxon p-value was calculated for each protein. After 1000 permutations, a null distribution was constructed with the same number of p-values as with the synthetic data. Figure 4B shows the cumulative p-value distribution for the null (dashed line) and synthetic (black line) data sets, respectively. As expected, the p-values for the null data set are approximately



B

i	r_i	Q1	Q2	Q3	IQR
1	0.5	0.8224	0.9976	1.2647	0.4423
2	1.0	0.7609	0.9655	1.2647	0.5038
3	2.0	0.7368	0.9777	1.3000	0.5632
4	3.0	0.6790	0.9730	1.2917	0.6127

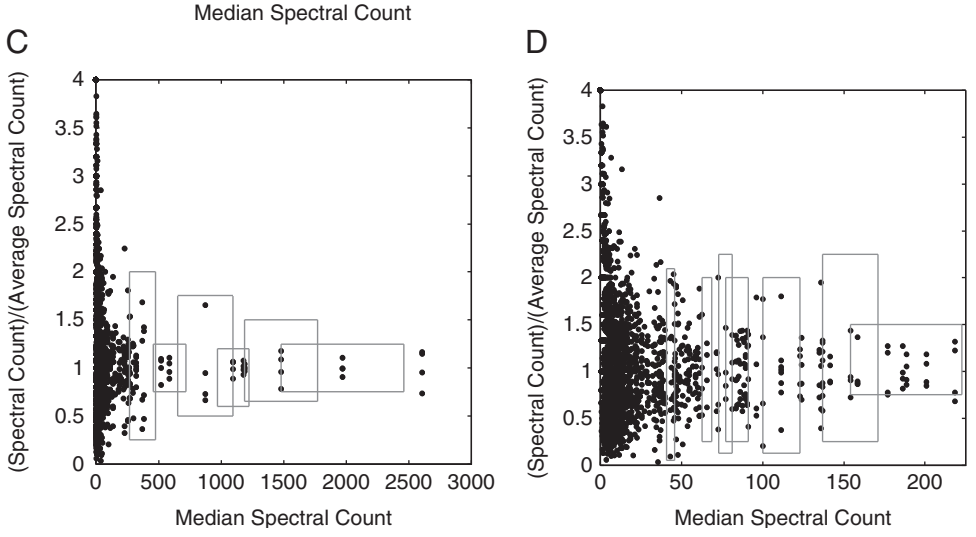


Figure 2. (A) Determination of appropriate sampling radius, r_i around median value $M_i = 25.5$. For the radii r_n ($n = 1 \ldots 4$), corresponding quartile boundaries and IQRs are given in (B). r_n is increased until an outlier (denoted by an open circle) exists in the data. At this point, r_i is defined as $n-1$. (C) Distribution of M_p *versus* $S_{p,i}$ for all p and i for a single condition [22]. Boxes correspond to sampling radii for $M_i = \{371.5, 587, 874, 1097, 1480, 1972\}$. (D) The same plot for M ranging from 0 to 225 with boxes imposed for $M_i = \{43, 65.5, 77, 84, 111.5, 154, 188.5\}$.

© 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

uniformly distributed while there are far more significant *p*-values in the synthetic data. To determine an overall confidence level, k, a *p*-value cutoff was taken based on the null distribution: $p_{\text{cutoff}}|(k) = (1-k)^{\text{th}}$ quantile of the null *p*-value distribution. Finally, each protein was given a new *p*-value, p', defined as

$$p' = \frac{\sharp p\text{-}values_{(simulation\text{-}based)} > p_{cutoff}}{100} \quad (3)$$

This p'-value corresponds to the ReSASC-based probability of a type I error, *i.e.* the identification is a false positive.

3.2 ReSASC is distinct from Wilcoxon test and fold-change cutoffs

Two sets of analyses were performed to validate the applicability of ReSASC and to show that it provides useful information beyond a Wilcoxon test alone or beyond a simple fold cutoff criterion. As a first validation step, ReSASC was run on the replicates data set whose runs were randomly separated into ''case'' and ''control'' sets. No proteins were found to be significantly different between these two groups at any relevant confidence level, specifically $k \geq 0.60$ (data not shown). Next ReSASC-based *p*-values were calculated for the yeast and cystic fibrosis data sets, and these values were correlated with the protein's corresponding average fold-change and with the protein's Wilcoxon *p*-value. Figures 5A and B show the respective scatter results of this analysis for the cystic fibrosis data set. Correlation analysis was highly significant for all comparisons (ReSASC p' *versus* Wilcoxon *p*-value: $R = 0.7619$, $p < 0.0001$ (yeast); $R = 0.8746$, $p < 0.0001$ (cystic fibrosis). ReSASC p' *versus* fold-change: $R = -0.6075$, $p < 0.0001$ (yeast); $R = -0.6901$, $p < 0.0001$ (cystic fibrosis)). In all cases, while the ReSASC p'-value is significantly correlated to the Wilcoxon metric, the two are in no instance

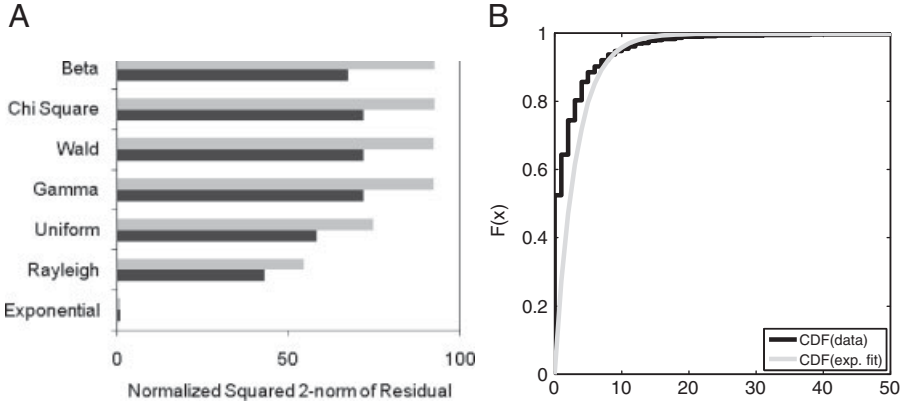


Figure 3. (A) Residual results after fitting the SC distribution from proteins with at least one zero value to various continuous probability distributions. All squared 2-norms of the residuals are scaled relative to the result from the exponential fit, which is at least 40 times smaller than the next best-fitting distribution. Fits are shown for synthetic two-fold (black) and synthetic (grey) data set. (B) Cumulative distribution functions for the SCs from the same group of proteins described previously, as well as the corresponding exponential fit.

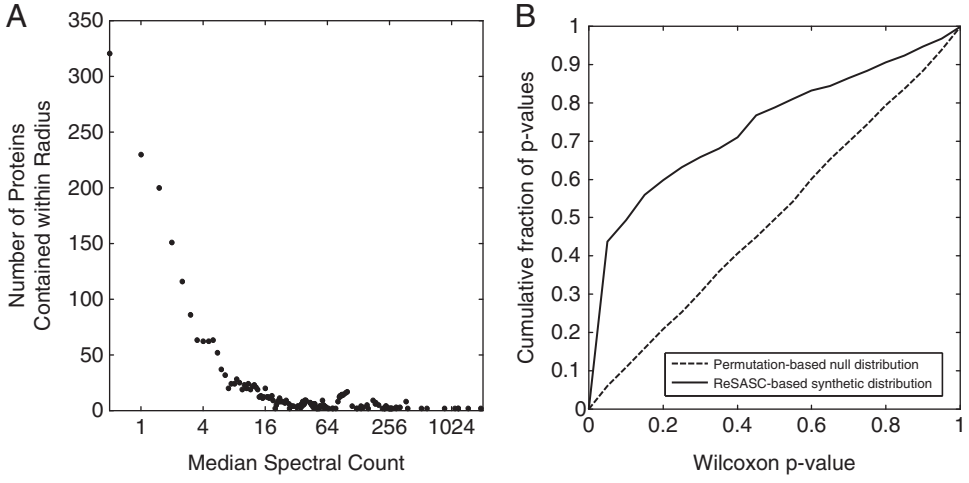


Figure 4. (A) Number of proteins contained in $M_i \pm r_i$, for each unique median value. This corresponds to the number of proteins whose SCs will be sampled to create a synthetic distribution. (B) Cumulative Wilcoxon *p*-value distributions for sampling-based synthetic data (black line) and permutation-based null data (dashed line). The 45° angle of the null data's distribution indicates that its *p*-values are uniformly distributed, as expected. On the other hand, the synthetic data has a much higher ratio of low *p*-values (40% ≤ 0.05 *versus* 5% of null data).

© 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

equivalent. In fact, the upper triangular distribution of the p_{Wilcoxon}–p' scatterplots indicate that the ReSASC p'-value is capable of excluding proteins that may have appeared significant based on a Wilcoxon test alone, or alternatively of identifying proteins that were only marginally significant.

Likewise, ReSASC can offer additional information beyond a simple fold-change cutoff. Previous studies have found that while SCs are effective in detecting differential expression between two groups, in general they are more effective in detecting large changes rather than small (*e.g.* twofold or less [17, 33, 34]). However, by incorporating the variable nature of SCs, ReSASC can identify as significant proteins with relatively small fold-changes. For example, in the cystic fibrosis data sets, the smallest significant fold-change was 1.7; and in the yeast data set, ReSASC was able to detect a fold-change of 1.2 in the highly expressed protein, glyceraldehyde-3-phosphate dehydrogenase, isozyme 1 (NP_012483.1; mean SC increased from 957 (log phase) to 1182 (stationary phase)).

3.3 ReSASC outperforms traditional statistical tests and performs comparably with SpI and QSpec

To compare the power of ReSASC with traditional statistical tests as well as with specialized tests for SCs (specifically QSpec and the SpI), receiver operator curves were constructed for each of the synthetic data sets. While there was no significant difference among any of the tests in detecting twofold changes (data not shown), Fig. 5C shows that all SC-based techniques outperformed the traditional *t*-test and Wilcoxon test in terms of their ability to detect a fourfold change. The caption reports the area under the curve (AUC) $\pm$ the width of the 95% confidence interval around that value.

3.4 Identification of differentially expressed proteins in yeast

Of the 906 proteins considered, 299 were found to be significantly different after QSpec analysis. Figure 6A shows a Venn diagram comparing the proteins determined to be differentially expressed between the two growth phases by ReSASC (with p_{cutoff} based on $k = 0.95$) and QSpec, respectively. ReSASC only claims an additional three proteins as significant relative to QSpec. (These three proteins are listed in Fig. 6C as well as five proteins identified by QSpec but not ReSASC, with ReSASC-based p'-value > 0.90.) ReSASC corroborates 128 (43%) of the proteins determined to be differentially expressed by QSpec. 26 (15%) of the proteins identified by QSpec but not ReSASC were identified in only one out of seven runs, and 59 (35%) were detected in at most two runs out of the seven, indicating that ReSASC is less affected by whether or not a protein was detected in one condition but not the other. Figure 6B demonstrates the relationship between the ReSASC-based p'-value and QSpec's Bayes factor ($B(X)$). The two metrics are significantly correlated ($R = -0.7538$; $p < 0.0001$). All identified proteins, along with their SCs, QSpec's Bayes factor, and ReSASC p'-value are listed in Supporting Information Table 1.

3.5 Identification of differentially expressed proteins in BAL fluid

Next, we applied ReSASC to the cystic fibrosis data set, which is more relevant for disease biomarker discovery. Figure 7A shows a Venn diagram comparing the distribution of proteins identified as significant by ReSASC and SpI, respectively. Of the 55 proteins identified only by SpI, 65% were detected in only one sample, with the average SC in the detected condition generally being very small (3.5, range = 2.75–4.75),

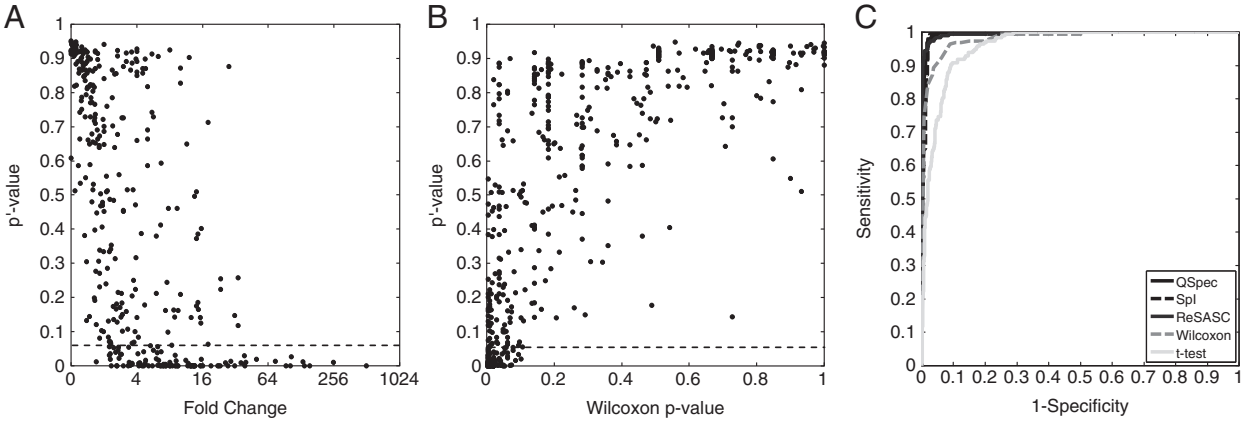
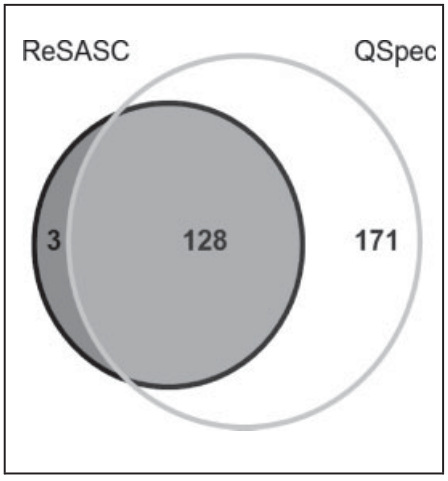


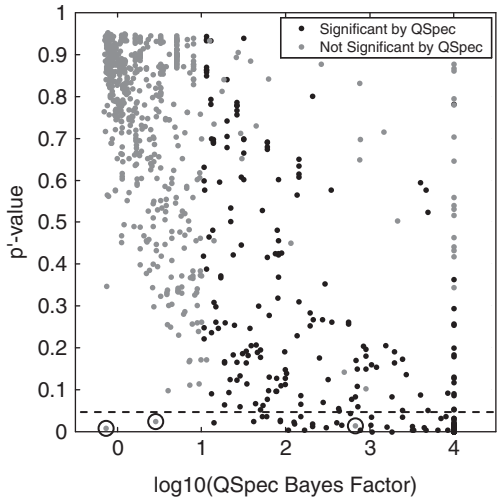
Figure 5. ReSASC p'-values (as defined in Fig. 2) *versus* (A) fold-change (Spearman correlation coefficient $R = -0.6075$, $p < 0.0001$) and (B) Wilcoxon *p*-value ($R = 0.7619$, $p < 0.0001$) for yeast data set [22]. (C) Receiver operator curve depicting the sensitivity and specificity of QSpec, SpI, ReSASC, Wilcoxon test, and *t*-test for detecting simulated fourfold changes in SC data. The corresponding AUC (95% CI width) are 0.996 (0.002), 0.994 (0.004), 0.996 (0.003), 0.982 (0.008), and 0.960 (0.011) for the respective tests.

© 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

A



B



C

Accession	LP_1	LP_2	LP_4	SP_1	SP_2	SP_3	SP_4	QSpec	ReSASC p'
NP_012483.1	949	960	963	1200	1096	1159	1273	FALSE	0.014
NP_013897.1	0	7	0	4	7	10	0	FALSE	0.025
NP_014873.1	0	18	0	2	9	12	3	FALSE	0.008
NP_015495.2	0	4	0	0	0	0	0	TRUE	0.944
NP_009815.1	0	0	0	9	0	0	0	TRUE	0.94
NP_009731.2	4	0	0	0	0	0	0	TRUE	0.938
NP_010131.1	0	0	4	0	0	0	0	TRUE	0.936
NP_012113.1	0	0	0	0	0	7	0	TRUE	0.935
NP_012269.1	0	4	0	0	0	0	0	TRUE	0.934

Figure 6. (A) Venn diagram showing the number of proteins determined to be differentially expressed in yeast between the log and stationary phases [31] by ReSASC and QSpec [22], respectively. The large overlap in ReSASC-detected proteins suggests that both methods find a common set of differentially expressed proteins. (B) ReSASC p'-values *versus* QSpec Bayes factors ($R = -0.7538$; $p < 0.0001$). The Bayes factor is less uniformly distributed than p', with 818 (90.3%) proteins having Bayes factors at either extreme (≤ 100 or ≥ 9900). At a p' significance level of 0.05 (dashed line), the three proteins identified by ReSASC but not by QSpec as differentially expressed are noted with open circles. (C) Table indicating the SCs from log phase (L1, L2, L4) and stationary phase (S1…4) of the three proteins identified by ReSASC but not QSpec (top), and five proteins identified by QSpec with ReSASC $p' > 0.9$ (bottom).

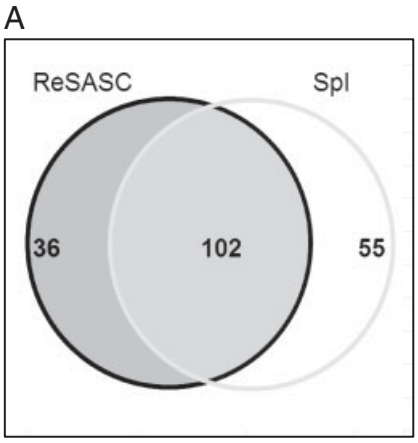
indicating a heightened sensitivity to whether or not a protein is detected independent of the SC values from the second condition. By contrast, ReSASC was able to identify relatively small fold-changes (3.5, range = 2.7–4.6) among proteins that were identified in both groups.

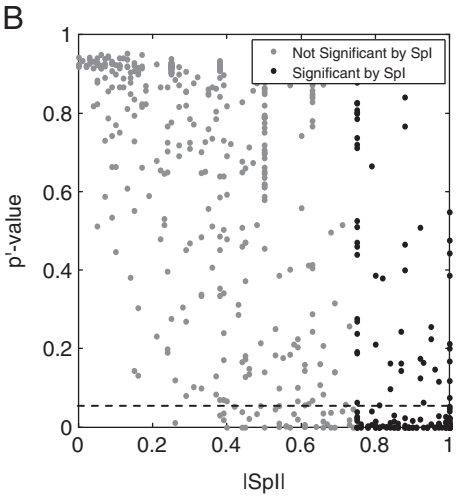
Figure 7B shows the relationship between each protein's ReSASC-based p'-value and the absolute value of its SpI. The two metrics are significantly correlated ($R = -0.7512$; $p < 0.0001$) as expected, since both metrics are based entirely on the proteins' SCs. The number of proteins identified as significant by both methods is 102, or 53% of the proteins identified as significant by either method. Several proteins from the cystic fibrosis data were measured by ELISA and the ReSASC p'-value was shown to correlate well with these proteins' p-values as determined by t-test. Of the six proteins listed, five were found to be statistically different by ELISA, five by SpI, and four by ReSASC. Only TIMP-1 differs between ReSASC and ELISA, with ReSASC claiming marginally significance with $p' = 0.045$. When a 99% confidence level is used, corresponding to $k = 0.99$ (as is used by Fu *et al.*), ReSASC agreed with ELISA results in all cases, thus validating the ReSASC approach by comparison with a gold standard ''ground truth''. These results are described in more detail in Fig. 7C. A list of all identified proteins together with their SpI and ReSASC-based p'-value is given in Supporting Information Table 2.

4 Discussion

As MS becomes a central tool for biomarker discovery, it is increasingly necessary to understand how reproducible its results are. Current literature on this topic generally focuses on reproducibility in terms of relative quantification, that is, the reproducibility of detected differences between multiple samples. Stable isotope labeling methods dominate such quantification, and there is an abundance of information detailing their reproducibility levels, which generally are relatively high. Although labeling methods are useful, they have been associated with reduced sensitivity of protein and peptide detections due to incomplete labeling. As such, several alternative methods – specifically spectral counting – have become increasingly popular although they are associated with lower levels of reproducibility. Liu *et al.* developed a non-linear model partially based on SCs to predict the total number of proteins within a specific abundance level that would be detected after a given number of replicates [35]. Such a model's explanation of run-to-run variability could possibly

© 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim





C

	Mean Spectral Count			ReSASC p'		Mean Conc. (ng/ml)		
Protein	CF (Cases)	Controls	SpI	k=0.95	k=0.99	CF (Cases)	Controls	p-value
MMP-9	113.0	0.0	1.00	<0.0001	<0.0001	748.0	1.6	0.0050
MMP-8	46.5	0.0	1.00	<0.0001	<0.0001	199.0	0.3	0.0001
MPO	332.0	3.2	0.99	<0.0001	<0.0001	3630.0	1.8	0.0060
TIMP-1	8.5	1.5	0.74	0.0450	0.1760	62.4	3.5	0.0700
ALB	1680.0	3060.0	-0.29	0.0800	0.2670	44500.0	56500.0	0.5000
ICAM-1	1.3	34.3	-0.96	<0.0001	<0.0001	4.0	13.2	0.0009

Figure 7. (A) Venn diagram depicting the distribution of proteins identified as significantly differentially expressed between patients with cystic fibrosis and healthy controls [23] by ReSASC and the spectral index (SpI) method [23], respectively. Of the 55 proteins identified only by SpI, 65% were detected in only one sample, with the median SC in the non-zero runs being small (3.5, range = 2.75–4.75). (B) ReSASC p'-values *versus* SpI. ReSASC $p' = 0.05$ is marked by a dashed line. |SpI| and p' are correlated ($R = -0.7512$; $p < 0.0001$) as expected since both metrics are based on the proteins' SCs. (C) Table comparing ReSASC p' (at 95% and 99% confidence levels), SpI (at 99% confidence level) and ELISA-based p-values (t-test) for detecting differential expression between the two groups.

be extrapolated to a statistical model for differential expression based on SCs, that is, if MS's variability could be effectively modeled, then that variability could be extrapolated to a theoretically robust statistical method based on SCs as output. However, Liu's model assumes information about mixture complexity, specifically abundance levels, the number of proteins at each abundance level, and the total number of proteins in the mixture. Generally, such information is unknown to the experimenter.

To avoid making these assumptions, ReSASC was developed as a non-parametric, permutation-based method to account for MS's inherent variability and determine differential expression between states. By pooling proteins that are similarly expressed – here defined to imply that the proteins in question have similar median SC values and the data set contains no outliers – ReSASC is able to exploit the variability information contained in SC data.

SC-based significance analysis has recently received significant attention. In addition to QSpec and SpI, Carvalho *et al.* recently tested various non-standard normalization procedures with different feature recognition tests and concluded that Z-normalization combined with a forward-support vector machine model (including a penalty function) performed optimally in identifying differences in yeast [19]. APEX (Absolute Protein Expression), introduced by Lu *et al.* also proposed machine learning techniques, but after the addition of a SC correction factor to account for the variability of peptide detection by MS [21]. LSPAD (Localized Statistics of Protein Abundance Distribution) performs Fisher's exact test, which effectively compares the log-transformed SCs from one protein to other similarly expressed proteins from the same experiment [20]. One advantage of LSPAD compared to the previously mentioned methods is that it does not require replicates since a protein's SCs are pooled from all runs in a given condition. However, in general all of these methods are statistically complex, and also require more information from an experiment than simply the proteins' SCs. As such, some are difficult to implement and understand. LSPAD – though easier to implement – is highly sensitive to outliers since it does not consider how reproducibly a protein was or was not detected.

ReSASC is the first method to apply a variation of the non-parametric Wilcoxon test to SCs. Up to this point, the Wilcoxon test has been used in MS experiments to detect differences based on labeling ratios, LC-MS feature ratios, and protein spots [33, 36, 37]. Non-parametric tests are by definition less powerful than their parametric counterparts, and they are not effective at taking fold-increases into account when determining statistical significance. For example, a control group consisting of SCs (0, 1, 1, 0, 1, 2) compared to a diseased group consisting of SCs (3, 3, 3, 4, 3, 3) would return the same p-value as the same control compared to a diseased group of SCs (100, 101, 105, 98, 100, 97). By simulating additional runs, ReSASC improves the power of the Wilcoxon test because synthetic runs are more likely to show no difference in the former than the latter case.

ReSASC has several limitations, most notably that it requires replicate samples. We believe that a minimum of

© 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

three replicates in both conditions are necessary for credible results since there is no distinction between the mean and the median of a single value or of two values. A retrospective power analysis supports this hypothesis, with estimated power reaching saturation at three runs (data not shown). Additionally, a protein's median value can only be approximated coarsely and is sensitive to outliers at low run numbers, which could in turn affect its sampling window and synthetic SCs. On the other hand, ReSASC is also affected by the distribution of SCs in a data set – specifically if many proteins are expressed at a similar level – so that the variability is low. Supporting Information Fig. 2 demonstrates ReSASC's robustness as the distribution of SCs changes to become less variable.

ReSASC's running time scales linearly with the number of proteins in a data set (mean R^2 based on linear regression is 0.992). As such, there is an obvious trade-off between run time and p'-value accuracy, since p'-values are best estimated when many proteins are present in each sampling window. Finally, ReSASC is currently only able to compare two states and therefore cannot be applied to experiments with more conditions or to time course data. In summary, ReSASC is an effective and robust tool for differential expression studies based entirely on SC data. Its increased stringency relative to existing metrics make it preferable for biomarker discovery studies, and its conceptual simplicity makes it easier to understand and operate compared to existing SC-based tests.

This work was supported by the American Heart Association [0815327F] and the wallace H. Coulter Foundation. ReSASC is currently being re-coded from MATLAB into the R programming language and we hope to submit the package to the BioConductor repository. The authors will run ReSASC for any researcher interested in using the method.

The authors have declared no conflict of interest.

5 References

[1] Cravatt, B. F., Simon, G. M., Yates, J. R., 3rd., The biological impact of mass-spectrometry-based proteomics. *Nature* 2007, *450*, 991–1000.

[2] Ackermann, B. L., Hale, J. E., Duffin, K. L., The role of mass spectrometry in biomarker discovery and measurement. *Curr. Drug Metab.* 2006, *7*, 525–539.

[3] Fisher, W. G., Rosenblatt, K. P., Fishman, D. A., Whiteley, G. R. *et al.*, A robust biomarker discovery pipeline for high-performance mass spectrometry data. *J. Bioinform. Comput. Biol.* 2007, *5*, 1023–1045.

[4] Li, H., DeSouza, L. V., Ghanny, S., Li, W. *et al.*, Identification of candidate biomarker proteins released by human endometrial and cervical cancer cells using two-dimensional liquid chromatography/tandem mass spectrometry. *J. Proteome. Res.* 2007, *6*, 2615–2622.

[5] Diamandis, E. P., Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. *Mol. Cell. Proteomics* 2004, *3*, 367–378.

[6] Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F. *et al.*, Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* 1999, *17*, 994–999.

[7] Chong, P. K., Gan, C. S., Pham, T. K., Wright, P. C., Isobaric tags for relative and absolute quantitation (iTRAQ) reproducibility: Implication of multiple injections. *J. Proteome Res.* 2006, *5*, 1232–1240.

[8] Regnier, F. E., Riggs, L., Zhang, R., Xiong, L. *et al.*, Comparative proteomics based on stable isotope labeling and affinity selection. *J. Mass Spectrom.* 2002, *37*, 133–145.

[9] Shen, Z. W., M; Briggs, S. P., Use of high-throughput LC-MS/MS proteomics technolgies in drug discovery. *Drug Discov. Today: Technol.* 2006, *3*, 301–306.

[10] Old, W. M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K. G. *et al.*, Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell. Proteomics* 2005, *4*, 1487–1502.

[11] Zhang, B., VerBerkmoes, N. C., Langston, M. A., Uberbacher, E. *et al.*, Detecting differential and correlated protein expression in label-free shotgun proteomics. *J. Proteome Res.* 2006, *5*, 2909–2918.

[12] Eng, J. K. M. A. L., Yates, J. R.,3rd, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 1994, *5*, 976–989.

[13] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999, *20*, 3551–3567.

[14] Hays, W. L., *Statistics*, Holt, Rinehart amd Winston, Inc., New York 1963.

[15] Stevenson, S. E., Chu, Y., Ozias-Akins, P., Thelen, J. J., Validation of gel-free, label-free quantitative proteomics approaches: applications for seed allergen profiling. *J. Proteomics* 2009, *72*, 555–566.

[16] Vertommen, D., Van Roy, J., Szikora, J. P., Rider, M. H. *et al.*, Differential expression of glycosomal and mitochondrial proteins in the two major life-cycle stages of Trypanosoma brucei. *Mol. Biochem. Parasitol.* 2008, *158*, 189-201.

[17] Hendrickson, E. L., Xia, Q., Wang, T., Leigh, J. A., Hackett, M., Comparison of spectral counting and metabolic stable isotope labeling for use with quantitative microbial proteomics. *Analyst* 2006, *131*, 1335–1341.

[18] Bridges, S. M., Magee, G. B., Wang, N., Williams, W. P. *et al.*, ProtQuant: a tool for the label-free quantification of MudPIT proteomics data. *BMC Bioinformatics* 2007, *8*, S24.

[19] Carvalho, P. C., Hewel, J., Barbosa, V. C., Yates, J. R., 3rd, Identifying differences in protein expression levels by spectral counting and feature selection. *Genet. Mol. Res.* 2008, *7*, 342–356.

[20] Li, R. X., Chen, H. B., Tu, K., Zhao, S. L. *et al.*, Localized-statistical quantification of human serum proteome associated with type 2 diabetes. *PLoS ONE* 2008, *3*, e3224.

© 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

[21] Lu, P., Vogel, C., Wang, R., Yao, X., Marcotte, E. M., Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* 2007, *25*, 117–124.

[22] Choi, H., Fermin, D., Nesvizhskii, A. I., Significance analysis of spectral count data in label-free shotgun proteomics. *Mol. Cell. Proteomics* 2008, *7*, 2373–2385.

[23] Fu, X., Gharib, S. A., Green, P. S., Aitken, M. L. *et al.*, Spectral index for assessment of differential protein expression in shotgun proteomics. *J. Proteome Res.* 2008, *7*, 845–854.

[24] Jeffreys, H., *The Theory of Probability*, Oxford University Press, Oxford 1961.

[25] Durr, E., Yu, J., Krasinska, K. M., Carver, L. A. *et al.*, Direct proteomic mapping of the lung microvascular endothelial cell surface *in vivo* and in cell culture. *Nat. Biotechnol.* 2004, *22*, 985–992.

[26] Dijkstra, M., Vonk, R. J., Jansen, R. C., SELDI-TOF mass spectra: a view on sources of variation. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 2007, *847*, 12–23.

[27] Jain, N., Thatte, J., Braciale, T., Ley, K. *et al.*, Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics* 2003, *19*, 1945–1951.

[28] Breitling, R., Armengaud, P., Amtmann, A., Herzyk, P., Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.* 2004, *573*, 83–92.

[29] Shevchenko, A., Wilm, M., Vorm, O., Mann, M., Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal. Chem.* 1996, *68*, 850–858.

[30] Washburn, M. P., Wolters, D., Yates, J. R.,3rd, Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* 2001, *19*, 242–247.

[31] Pavelka, N., Fournier, M. L., Swanson, S. K., Pelizzola, M. *et al.*, Statistical similarities between transcriptomics and quantitative shotgun proteomics data. *Mol. Cell. Proteomics* 2008, *7*, 631–644.

[32] Chen, E. I., Hewel, J., Felding-Habermann, B., Yates, J. R., 3rd, Large scale protein profiling by combination of protein fractionation and multidimensional protein identification technology (MudPIT). *Mol. Cell. Proteomics* 2006, *5*, 53–56.

[33] Chakravarti, B., Oseguera, M., Dalal, N., Fathy, P. *et al.*, Proteomic profiling of aging in the mouse heart: Altered expression of mitochondrial proteins. *Arch. Biochem. Biophys.* 2008, *474*, 22–31.

[34] Smalley, D. M., Root, K. E., Cho, H., Ross, M. M., Ley, K., Proteomic discovery of 21 proteins expressed in human plasma-derived but not platelet-derived microparticles. *Thromb. Haemost.* 2007, *97*, 67–80.

[35] Liu, H., Sadygov, R. G., Yates,J. R., 3rd, A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* 2004, *76*, 4193–4201.

[36] Cole, A. R., Ji, H., Simpson, R. J., Proteomic analysis of colonic crypts from normal, multiple intestinal neoplasia and p53-null mice: a comparison with colonic polyps. *Electrophoresis* 2000, *21*, 1772–1781.

[37] Titus, M. A., Schell, M. J., Lih, F. B., Tomer, K. B., Mohler, J. L., Testosterone and dihydrotestosterone tissue levels in recurrent prostate cancer. *Clin. Cancer Res.* 2005, *11*, 4653–4657.

© 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim