BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 20 2011, pages 2866–2872
doi:10.1093/bioinformatics/btr479

Gene expression

Advance Access publication August 18, 2011

Improved quality control processing of peptide-centric LC-MS proteomics data

Melissa M. Matzke[1], Katrina M. Waters[1], Thomas O. Metz[1], Jon M. Jacobs[1], Amy C. Sims[2], Ralph S. Baric[2], Joel G. Pounds[2] and Bobbie-Jo M. Webb-Robertson[1,*]

[1]Pacific Northwest National Laboratory, PO Box 999, Richland, WA 99352 and [2]Department of Microbiology and Immunology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Associate Editor: Trey Ideker

ABSTRACT

Motivation: In the analysis of differential peptide peak intensities (i.e. abundance measures), LC-MS analyses with poor quality peptide abundance data can bias downstream statistical analyses and hence the biological interpretation for an otherwise high-quality dataset. Although considerable effort has been placed on assuring the quality of the peptide identification with respect to spectral processing, to date quality assessment of the subsequent peptide abundance data matrix has been limited to a subjective visual inspection of run-by-run correlation or individual peptide components. Identifying statistical outliers is a critical step in the processing of proteomics data as many of the downstream statistical analyses [e.g. analysis of variance (ANOVA)] rely upon accurate estimates of sample variance, and their results are influenced by extreme values.

Results: We describe a novel multivariate statistical strategy for the identification of LC-MS runs with extreme peptide abundance distributions. Comparison with current method (run-by-run correlation) demonstrates a significantly better rate of identification of outlier runs by the multivariate strategy. Simulation studies also suggest that this strategy significantly outperforms correlation alone in the identification of statistically extreme liquid chromatography-mass spectrometry (LC-MS) runs.

Availability: https://www.biopilot.org/docs/Software/RMD.php

Contact: bj@pnl.gov

Supplementary information: Supplementary material is available at *Bioinformatics* online.

Received on March 7, 2011; revised on June 27, 2011; accepted on July 29, 2011

1 INTRODUCTION

The majority of statistical strategies to assess peptide/protein differential abundances from liquid chromatography-mass spectrometry (LC-MS) proteomic experiments are based on analysis of variance (ANOVA) methodologies applied to peak intensities (i.e. abundance measures) of proteolytic peptides (Bukhman *et al.*, 2008; Daly *et al.*, 2008; Karpievitch *et al.*, 2009; Oberg and Vitek, 2009; Oberg *et al.*, 2008). However, the ANOVA approach relies upon accurate estimates of sample variance, and proteomics studies not only have inherent variability associated with the biological samples, but potentially diverse process-based sources of variability. That is, the accurate estimate of sample variances is often difficult to obtain. For example, sample preparation protocols and instrument variations associated with the LC column (particularly important for multi-column platforms) and mass spectrometer can cause variations in peak intensities, as well as peptides identified across MS analyses within an experiment. Data quality is especially important when the number of biological samples is small, often the case in proteomics experiments, and extreme values can negatively influence all subsequent data analysis outcomes.

Identification of statistical outliers in univariate data is an established but highly debated statistical topic (Barnett and Lewis, 1994; Hawkins, 1980). There are many consecutive outlier procedures, focusing on one suspect value at a time, that have been proposed and implemented across many fields of application, such as Grubbs' test and Dixon's Q-test (Dixon, 1950; Grubbs, 1950). Because these methods iteratively remove outlier points, the false positive rate (i.e. Type 1 error) is inflated (Jain, 2010). In contrast, recursive outlier detection procedures detect the presence of any number of outliers and control Type 1 errors. For example, Jain (2010) presents a recursive version of Grubbs' test and Caroni and Prescott (1992) derived a sequential application of Wilks's multivariate outlier test. There are downfalls to these recursive procedures: (i) they are designed for univariate data, and if applied to multivariate data, will likely fail to detect statistically influential extreme values, and (ii) they are negatively affected by masking (i.e. the inability to detect an outlier in the presence of another outlier) and swamping (i.e. identify non-outliers as outliers) effects.

The identification of statistical outliers in multivariate data, such as microarray and proteomic data, is non-trivial. The multiple dimensions of the data often subject outliers to masking (Filzmoser *et al.*, 2008). The microarray community, however, has made considerable progress in applying statistical metrics to assess the quality of microarray data (Kauffmann *et al.*, 2009; Kemmeren *et al.*, 2005; Lee *et al.*, 2006; Wilson and Miller, 2005). Of particular applicability to proteomics data are the ideas presented by Kauffmann *et al.* They note that a poor quality array will impede the statistical and biological significance of the analysis due to the added noise. This is also true for proteomics data. That is, poor quality peptide abundance data will hinder downstream statistical analysis, including normalization, and subsequent biological interpretations.

*To whom correspondence should be addressed.

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

For proteomics data, a routine but non-probabilistic approach used for the identification of outlier LC-MS analyses (i.e. runs) during data preprocessing is through a correlation matrix plot (Metz *et al*., 2008). The sample correlation coefficient is calculated among technical replicates and biological replicates. Those runs with a relatively low correlation are removed from the dataset. The determination of 'low' correlation is subjective, and varies across analysts, experiments and time. Correlation may be examined via a heat map in which a color palette represents the numeric value, the color palette choice as well as the range of correlation values it covers can be highly subjective and extremely influential on the selection of which runs should be removed from the dataset. In addition, the sample correlation coefficient can only be computed across peptides with common identifications between runs (i.e. it does not account for missing data), it does not account for the multivariate nature of LC-MS runs, nor is there any statistical certainty associated with the exclusion of a run.

Advanced statistical approaches to outlier detection in proteomics data have focused either on the identification of outlier spectra maps (Rudnick *et al.*, 2010; Schulz-Trieglaff *et al.*, 2009) or on peptide/protein abundances independent of LC-MS run behavior (Cho *et al.*, 2008; MacCoss *et al.*, 2003; Xia *et al.*, 2006). Rudnick *et al.* (2010) described a large set of metrics for the quantitative assessment of system performance and evaluation of technical variability among inter- and intra-laboratory LC-MS/MS proteomics experiments. However, the use of these metrics to assess the quality of an individual LC-MS/MS run is not addressed. Schulz-Trieglaff *et al.* (2009) applied a multivariate method to perform a quality assessment of raw LC-MS maps using 20 quality descriptors. The goal of their approach was to identify and remove outlier runs using unprocessed spectra before noise filtering, peak detection or centroiding was performed. Cho *et al.* (2008) presented a peptide outlier detection method using quantile regression to account for the heterogeneity of variance between replicate LC-MS/MS runs. Peptide intensity ratios were plotted on an *MA* plot, where *M* is the difference in peptide abundance values and *A* is the average peptide intensity value. MacCoss *et al.* (2003) developed a correlation algorithm to detect outlier peptides using fractional changes between sample and reference intensities. Xia *et al.* (2006) proposed a two-stage method, combining Dixon's Q-test and a median absolute deviation (MAD) modified *z*-score test, for outlier detection of peptide ratios. These latter methods focus on assessing individual peptides for extreme behavior rather than the distribution of peptide abundance values for an entire LC-MS run.

Our goal is to statistically identify runs that exhibit extreme peptide abundance distribution properties, and thus will likely impact downstream statistical analyses. Consequently, we are not focused on outliers specific to the spectral properties. We describe a statistical strategy to identify and remove extreme LC-MS runs with a high level of statistical certainty, thus removing subjectivity from the filtering process. The approach, based on a robust Mahalanobis distance (rMd), assesses the reproducibility of the distribution of peptide abundance values across replicate runs of the same biological sample as well across related biological samples. Statistical methods, which limit the influence of extreme observations, are applied to obviate assumptions about underlying probabilistic models (Hoaglin *et al.*, 2000). We demonstrate the approach by applying it to simulated and real LC-MS datasets.

2 METHODS

Our approach to detect and ascertain if an individual LC-MS run within an experiment, is a statistical outlier with a four-step process. The algorithm was implemented in MATLAB (version 7.10.0.499, R2010a, The MathWorks Inc.: Natwick, MA, USA).

2.1 Summarize each LC-MS run as five metrics

Five statistical metrics were chosen to describe the distribution of observed peptide abundance values in a single LC-MS run. These metrics described below capture selected aspects of the peptide abundance distribution such as shape and scatter. The location of each distribution is not directly considered since it could potentially be a false indicator of outlingness. In addition, location can easily be corrected by a simple overall normalization factor. The metrics are vectorized for each run, represented as $\vec{x}$; initially reducing the dimension of each run from p peptides to q metrics with the resulting dataset dimensionality of $(n\times q)$ where n is the number of LC-MS runs.

2.1.1 Metric 1: correlation coefficient The sample correlation coefficient, r_{ij}, is calculated for peptide abundance values between all LC-MS runs ($i=1,\ldots,n$; $j=1,\ldots,n$) resulting in an $n\times n$ matrix. The correlation coefficient metric for the i-th run, R_i, which is used for the robust principal component analysis, is the average correlation within a common grouping (e.g. treatment group, G), and has dimension $(n\times 1)$. For the i-th run this is computed as,

$$R_i=\frac{1}{N_{G(i)}}\sum_{j\in G(i)} r_{ij} \tag{1}$$

where $N_{G(i)}$ is the total number of runs in the group associated with run i. The average correlation among biological replicates, rather than among technical replicates, is used due the small number of technical replicates, if any at all, observed in a typical LC-MS experiment.

2.1.2 Metric 2: fraction of missing peptide abundance data The fraction of missing abundance data in the i-th $(1,\ldots,n)$ LC-MS run is defined as,

$$Fm_i=\frac{\sum_{j=1}^{p} a_{ij}}{p} \tag{2}$$

where $a_{ij}=1$ if the j-th peptide abundance is absent for the i-th run; otherwise, $a_{ij}=0$.

2.1.3 Metric 3: median absolute deviation of peptides within a LC-MS run The MAD (Hoaglin *et al.*, 2000) is a robust measure of the spread of the data, and is used as an estimate of the sample standard deviation if scaled by a factor of 1.483. The MAD of the i-th LC-MS run is defined as,

$$\mathrm{MAD}_i=\mathrm{med}\left|x_j-\mathrm{med}(X)_i\right| \tag{3}$$

That is, within a run, each abundance value for peptide j is compared with the median peptide abundance values of the run i.

2.1.4 Metric 4: skew The asymmetry of a distribution is described by skew. In our application to the i-th $(1,\ldots,n)$ LC-MS run, p is the number of peptides observed in the i-th run, $\bar{x}$is the average peptide abundance value of all peptides observed in the i-th run and S is the sample standard deviation of the i-th run.

$$\mathrm{Skew}_i=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i-\bar{x}}{S}\right]^3 \tag{4}$$

2.1.5 Metric 5: kurtosis The peakedness, or 'heavy-tailedness', of a distribution is described by kurtosis. The same parameters are used as skew. Kurtosis is calculated as,

$$\mathrm{Kurtosis}_i=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i-\bar{x}}{S}\right]^4-3 \tag{5}$$

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

2.2 Obtain a robust estimate of the covariance matrix

The purpose of robust principal component analysis (rPCA) in our method is to obtain the eigenvalues and eigenvectors to calculate a robust covariance matrix, which will be used in the calculation of the rMds. We employ a rPCA algorithm developed by Croux *et al.* that is based on the projection-pursuit approach to estimate the eigenvalues, and subsequent scores obtained from the projections of the metrics on the eigenvectors (Croux and Ruiz-Gazen, 2005; Li and Chen, 1985). The robust covariance estimate is defined as,

$$C_{Sn}=\sum_{k=1}^{p}\lambda_{Sn,k}\nu_{Sn,k}\nu^{t}_{Sn,k} \tag{6}$$

for which S_n is the robust scale estimator used by the projection-pursuit index $\lambda_{Sn,k}$ is the k-th eigenvalue and $\nu_{Sn,k}$ is the k-th eigenvector (Croux and Ruiz-Gazen, 2005). The rPCA algorithm uses the L_1-median value to center the data, and (MAD*1.483) as the robust scale estimate.

2.3 Identify outlier LC-MS run(s) using the rMd

A widely accepted measure of distance in multivariate data is the Mahalanobis distance because it accounts for not only the average value, but also the covariance structure of the measured variables (Mahalanobis, 1936). The distance of an individual LC-MS run from the center of the data is measured by a rMd. For a q-dimensional multivariate vector $\vec{x}_i$ for $i=1,\dots,n$, the rMd is defined as,

$$D_M(x)=\sqrt{(\vec{x}_i-\vec{m})^T C_{Sn}^{-1}(\vec{x}_i-\vec{m})} \tag{7}$$

where C_{Sn}, a robust estimate of the covariance matrix, is obtained from the robust principal component analysis of the $n\times q$ quality matrix, and $\vec{m}_i$ is a vector of medians of the five metrics.

2.4 Statistical assessment of the rMds

The rMd squared values associated with the peptide abundances vector (rMd-PAV) is the score used to assess whether an individual LC-MS run is an outlier. The rMd-PAV scores are approximately chi-square distributed with q degrees of freedom (χ^2_q). Therefore, outlier LC-MS runs are defined by a large rMd-PAV score such that the calculated squared distance exceeds a critical value of the χ^2_q distribution specified a priori.

2.5 Proteomics data processing

We present two independent real datasets to demonstrate the application of this outlier discovery strategy to LC-MS proteomics data. Human cell culture samples were analyzed with an Exactive mass spectrometer (Thermo Electron Corp.), and mouse plasma samples were analyzed with an LTQ-Orbitrap mass spectrometer (Thermo Electron Corp.). Nanoelectrospray ionization was used in the analysis of all samples. Spectra were collected at 400–2000 m/z with a resolution of 100 k and analyzed using the accurate mass and elution time (AMT) tag approach (Smith *et al.*, 2002). The mass de-isotoping process was performed using Decon2LS (Jaitly *et al.*, 2009), and the matching process was performed using VIPER (Monroe *et al.*, 2007). Features from the LC-MS analyses were matched to AMT tags to identify peptides, using an initial tolerance of ±3 p.p.m. for mass and 2.5% for the LC normalized elution time (NET). The human cell culture peptide datasets were further processed to remove peptides identified with low confidence, using the uniqueness filter Statistical Likelihood Confidence (SLiC) (Anderson *et al.*, 2006) score of 0.35 and a DelSLiC of 0.2. In circumstances where a peptide was identified in some LC-MS analyses, but not others, the missing data were coded as 'NaN'. All peptide abundance values were transformed to the log10 scale. Minimum occurrence data filters were used to identify those peptides for which the amount of data present was not adequate for differential abundance analysis (Webb-Robertson *et al.*, 2010). The sample complexity of the sham controls (SCs) in each of the designed experiments is the same with respect to original biological material.

3 RESULTS

Simulations of size 500 based on the p-variate standard normal distribution $Np(\mathbf{0},\mathbf{I})$, and an empirically influenced p-variate normal distribution $Np(\mu,\Sigma)$ were performed to examine a range of outlier configurations. In addition, we assessed the performance of the multi-dimensional outlier detection method against the conventional method of using a Pearson's correlation coefficient [previously described in Section 2.1 as metric 1—Equation (1)] to ascertain whether a LC-MS run is an outlier. Simulation is useful to investigate the properties of rMd-PAV, however; since simulation of expected distribution parameters in real proteomics data is not well understood, these results are presented in Supplementary Material (Rocke *et al.*, 2009).

The results of the multi-dimensional outlier detection analysis are displayed in a simple yet effective graphic in which rMd-PAV scores are plotted for each LC-MS run and compared with a reference line representing the χ^2 critical value. For improved visualization, the rMd-PAV scores and the χ^2 critical value are transformed to the log 2 scale. The red horizontal line represents the $\log_2(\chi^2_{0.9999,5})$ critical value. That is, at a significance level of 0.0001, a LC-MS run may be classified as a statistical outlier if the calculated test statistic $\geq \chi^2_{0.9999,5}$ critical value, or equivalently, the χ^2 $P\leq 0.0001$. LC-MS runs with $\log_2$(rMd-PAV) scores above the red horizontal line are suspect and should be removed from the dataset.

3.1 Real data benchmark—expert identified outlier runs

Calu-3 cells, a human lung adenocarcinoma cell line, were infected with the severe acute respiratory syndrome coronona virus (SARS-CoV) at a multiplicity of infection of 5. Cell monolayers were inoculated with SARS for 40 min at 37°C, and sham-infected controls were inoculated with medium only. Following inoculation, monolayers were rinsed and incubated for times 0, 3, 7, 12, 24, 30, 36 and 48 h. At the indicated times post-infection, wells were washed three times with ice cold 150 mM ammonium bicarbonate buffer and cells lyzed for 5 min in ice cold 8 M urea. Samples were frozen at −80°C until assayed. Samples were analyzed in triplicate, except where noted in Supplementary Table S2, and the minimum occurrence filter returned a total of 26 776 peptides (Webb-Robertson *et al.*, 2010).

This study included three biological replicates per time point as well as a large number of LC-MS runs ($n=141$), thus the removal of runs with poor quality abundance data is essential to maintain statistical power in downstream analyses. An LC-MS expert at Pacific Northwest National Laboratory upon reviewing the chromatography maps for this study was able to designate 28 out of 141 (~20%) LC-MS analyses as suspect due to various reasons (e.g. electrospray instability, elution time, sample prep/collection problem). We performed the rMd-PAV analysis, and compared its performance with t correlation alone to identify statistical outliers (runs at the peptide abundance level) via a receiver operating characteristic (ROC) curve analysis.

The rMd-PAV approach identified 12 out of the 28 expert-designated suspect runs as statistical outliers at the 0.0001 significance level (Fig. 1a). Electrospray issues represent almost half (13/28) of the expert identified runs, while the statistical algorithm identified three of these runs. It is the most likely technical issue to occur and the most difficult to detect. One reason

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

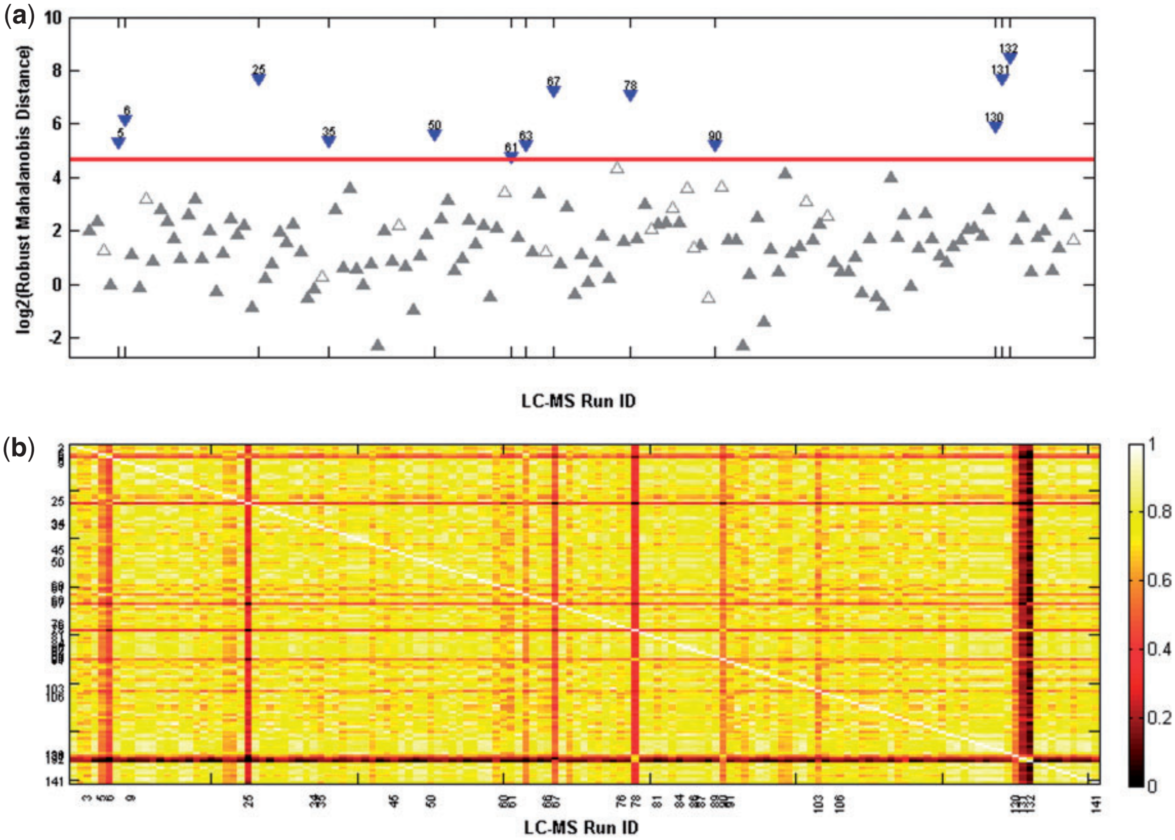

Fig. 1. Calu-3 cell-line experiment. (**a**) The rMd-PAV plot of the LC-MS runs. Runs identified as outliers (blue downward triangles) sit above the red horizontal line which represents the $\log_2\left(\chi^2_{0.9999,5}\right)$ critical value (i.e. $P = 0.0001$). The empty upward triangles below the red horizontal line represent runs identified as suspect by the MS expert that were not identified as statistical extreme. (**b**) The correlation plot of the LC-MS runs.

could be that the electrospray issue does not translate to a poor peptide abundance distribution, and thus an outlier. The other 15 runs identified by the MS expert are due to elution time (5/28; 4/5 identified by algorithm), chromatography (3/28; 1/3 identified by algorithm) and sample prep/collection (7/28; 4/7 identified by algorithm).

LC-MS runs that were expert designated as suspect, but did not exhibit different peptide abundance distributions from those runs that were not designated as suspect are identified in Figure 2a as unfilled triangles. Although the MS expert identified these runs as suspicious, the peptide abundance distributions are indistinguishable from those runs that were not designated as suspect.

In addition, we reviewed the sample correlation coefficient between all the study runs (Fig. 1b). Based on a subjective visual inspection of this graph, 6 out of the 28 expert-designated suspect LC-MS runs (#6, 25, 67, 78, 131 and 132) would have been dropped from the dataset. The rMd-PAV scores identified six additional runs as statistical outliers. This method did not identify any of the extreme runs due to electrospray issues; it did identify 3/5 runs labeled as suspect due to elution time, 1/3 suspect runs due to chromatography and 2/7 runs due to sample prep/collection issues.

A ROC analysis was completed to compare all levels of sensitivity and specificity. A comparison of the ROC curves for the rMd-PAV scores and the correlation metric alone by a Wilcoxon signed

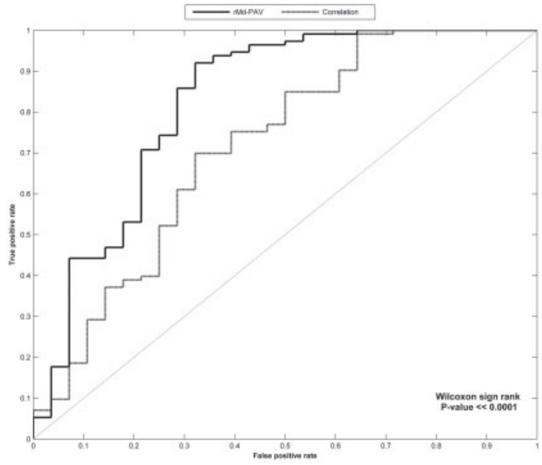

Fig. 2. The ROC curves from the rMd-PAV and correlation alone outlier analyses of the calu-3 cell-line experiment.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

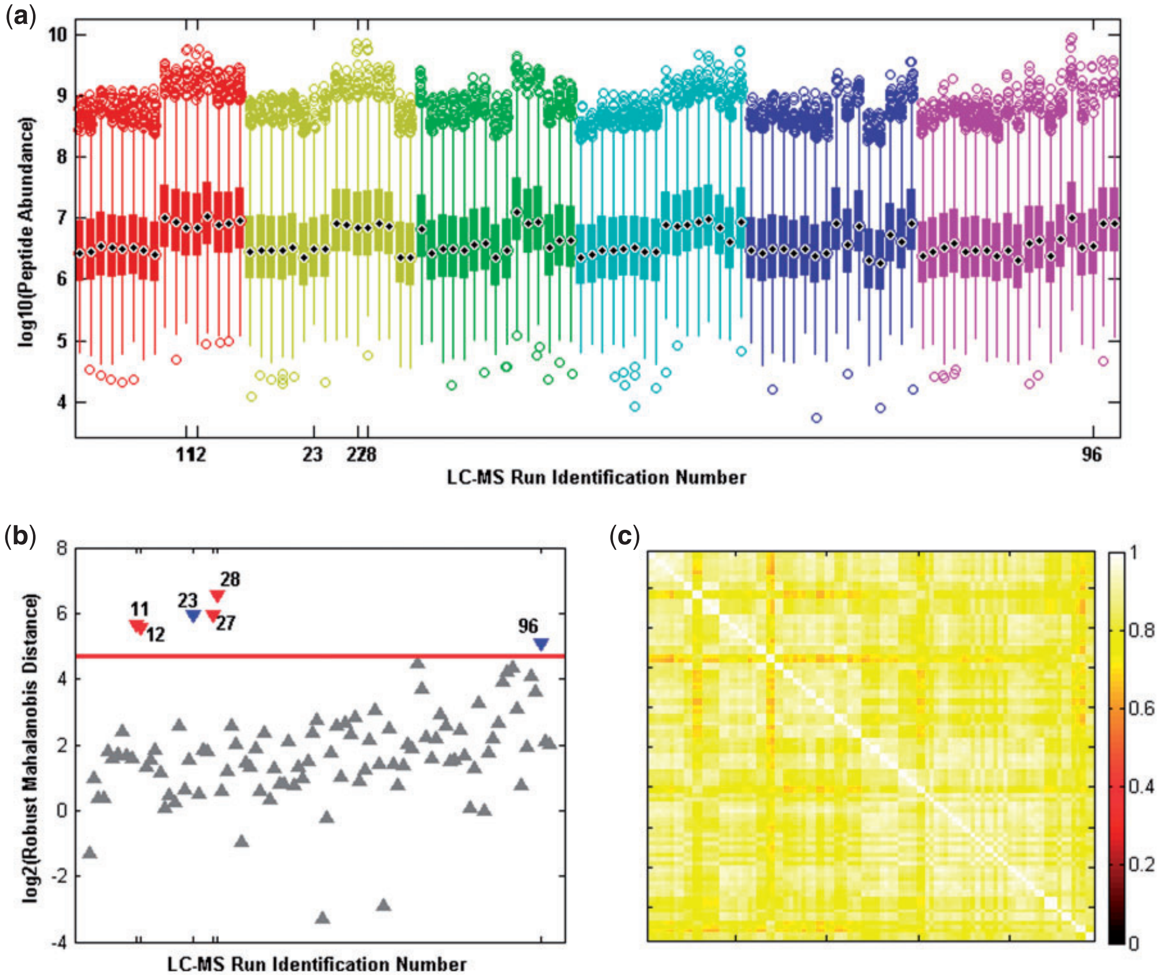

Fig. 3. Cigarette smoke exposure experiment. (**a**) Box plots of peptide abundance values observed in LC-MS runs ($n=98$) for the mouse plasma dataset. The color indicates experimental group membership. (**b**) The rMd-PAV plot of the LC-MS runs. Those runs identified as outliers sit above the red horizontal line which represents the $\log_2\left(\chi^2_{0.9999,5}\right)$ critical value (i.e. $P \leq 0.0001$). The downward triangles represent outlier runs—red represents all technical replicates from a biological sample, and blue represents individual technical replicates within a sample. (**c**) The run-by-run (r_{ij}) correlation plot of the LC-MS runs.

rank test results in statistically significant differences between the curves in favor of rMd-PAV ($P < 0.0001$, Fig. 2). Therefore, for this benchmark dataset we observe that rMd-PAV scores are superior to correlation alone for the identification of statistical outlier runs in LC-MS peptide abundance data.

3.2 Case study—cigarette smoke exposure data

Groups ($N=8$ biological replicates) of regular weight (RW) and diet-induced obese (OB) C57BL/6 mice (15 weeks old) were exposed to either filtered air (SCs), mainstream (MS) or side stream (SS) cigarette smoke by nose-only inhalation exposure for 5 h/day for 8 days. Target cigarette smoke exposure concentrations were 250 µg wet-weight total particulate matter (WTPM)/L of air for the MS exposures and 85 µg WTPM/L for the SS exposures. RW mice are defined as those mice fed a regular diet (PMI 5002 Rodent Diet®, Richmond, IN, USA; ~5kal% fat) throughout the study. DIO mice were fed a high-calorie/high-fat diet (D12492 Rodent Diet, Research Diets Inc., New Brunswick, NJ, USA; 60kal% fat) starting at 6 weeks of age and continued throughout the study. Immediately following the last exposure, each animal was removed from the exposure unit and anesthetized. Blood was collected into tubes containing potassium ethylenediaminetetraacetic acid (EDTA) (Tyco Healthcare Group LP, Mansfield, MA, USA) and centrifuged to obtain plasma for analysis by LC-MS/MS. Samples were analyzed in duplicate, except where noted in Supplementary Table S3 and a minimum occurrence filter returned a total of 3655 peptides (Webb-Robertson *et al.*, 2010).

As in any data analysis problem, visual inspection of complex data before statistical analysis is vital. Box plots are a simple and statistically robust techniques that are informative concerning distributional properties (e.g. skew and kurtosis), and provide visual guidance when interpreting analysis results. Peptide abundance data for each example has been displayed versus a LC-MS run order

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

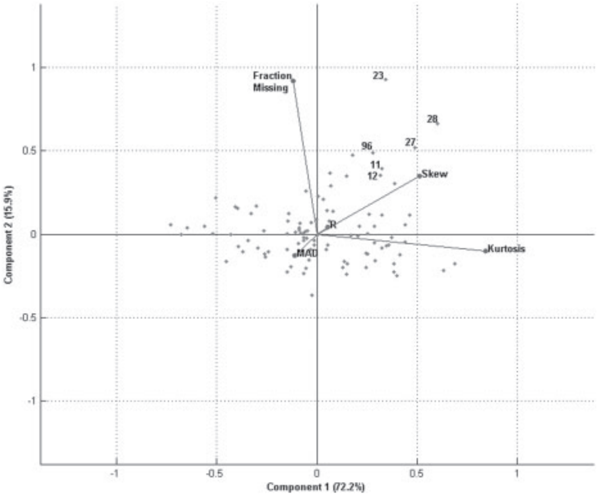

Fig. 4. Cigarette smoke exposure experiment. The score plot of the first two latent variables resulting from the rPCA of the data. It suggests the runs labeled on the plot are outliers due to the fraction of missing peptide abundance values, and the skewness and kurtosis of the peptide abundance distribution within a run.

identification number (not true LC run order) using a box plot. The box plot of the mouse plasma data (Fig. 3a) shows a fair amount of variability from run to run making visual determination of statistical outlier runs difficult.

The rMd-PAV approach identified 6 out of the 98 LC-MS runs as statistical outliers at the 0.0001 confidence level (Fig. 3b). Singleton technical replicates were removed (run id #23 and 96), in addition to two complete biological samples (run id #11 and 12—obese SC sample; run id #27 and 28—obese MS inhalation sample). Of the six runs identified as statistical outliers, it is unlikely any would have been removed using run-by-run correlation coefficient, r_{ij}, as the median correlation of all runs is 0.86 (Fig. 3c), and ranging from 0.72 to 0.87 across the pool of identified outlier runs. Using a more reflective score of correlation, R_i, which for the i-th run is the average correlation among the biological replicates within a group, the rMd-PAV identified runs would not have been removed from the dataset as the median correlation is 0.88 ranging from 0.73 to 0.85 across the identified outlier runs.

An additional benefit of the rPCA is the ability to explore the behavior of the metrics (e.g. skew, kurtosis, fraction missing, etc.) used to describe the peptide abundance distributions for the LC-MS runs within an experiment. Explaining high-dimensional data in two or three latent variables (i.e. principal components) is highly desirable. With only a few latent variables, data can be graphically displayed and the key contributing attributes to the total explained variation is easily interpreted. The relationship among the five metrics for peptide abundance data can be understood by examining the score plots of the latent variables. In addition, the behavior of the outlier runs can be understood relative to the non-outlier runs (i.e. average).

The most dominant manner in which these runs deviate is *Kurtosis, Skew* and *Fraction Missing Data*, as observed in the score plot associated with the rPCA (Fig. 4). The score plot is unique to an experiment, and thus is an excellent tool to further understand statistical differences in the peptides distributions among the LC-MS runs. The first score plot to consider is a comparison of the first two rPCA components (i.e. latent variables). In combination they account for >88% of the total variation in the data, and suggest differences among kurtosis, skew and fraction of missing abundance data explain most of the variation in the data. The plot shows the rMd-PAV identified runs located at the extreme ends of the observed data with respect to the first and second latent variables. Using the angle between vectors as a visual guide, for this data, it can be deduced the *Fraction Missing Data* and *Skew* of the peptide abundance distribution are correlated. In total, the first three components account for ∼95% of the variation observed in the data. While a two-dimensional view of the data is helpful in understanding relationships among variables, outliers and non-outliers, it is the relationship among the data under the full dimensionality that is the basis for the evidence of outlier runs.

4 DISCUSSION

Outlier detection in multivariate data is a non-trivial statistical task often subject to the masking effect (Filzmoser *et al.*, 2008; Rocke and Woodruff, 1996). Caution should always be taken when removing data from any dataset, large or small, and data should not be removed solely on the grounds of a statistical outlier test. Rather, the results of any statistical outlier algorithm used should always be reviewed in the context of the research goal and the experiment. Often the extreme data values are of interest and may explain technical difficulties in the process (e.g. sample preparation issues, technical difficulties with instrumentation and a mislabeling of samples). However, as with any statistical analysis and especially those dealing with small sample sizes, reviewing the outcome of the analysis is imperative. Specifically, graphical methods allow the analyst to review the analysis in a stepwise manner. For example, as our first step, we first plot the peptide abundances observed in the experiment for each LC-MS run using a box plot. Then to understand how the abundance distributions vary across the LC-MS runs we examine the scores plot resulting from the robust PCA.

5 CONCLUSION

We have presented a novel approach to the identification of statistical outliers in LC-MS proteomics peptide abundance data. The value of the multivariate outlier discovery strategy utilizing rMd-PAV scores is the use of an objective probabilistic model to assess statistical certainty of the exclusion of runs within an experiment in the context of the complete dataset. Proteomics has placed considerable effort on assuring the quality of the peptide identification with respect to spectral processing (Piening *et al.*, 2006; Rudnick *et al.*, 2010; Schulz-Trieglaff *et al.*, 2009; Stead *et al.*, 2008); however, quality assessment of the subsequent data matrix has focused on subjective visual inspection of run-by-run correlation, or individual peptide components. The quality of the LC-MS peptide abundance data matrix is essential to the identification of robust biomarkers. Moreover, statistical evaluation of the data relies upon tools often based on linear models, such as ANOVA which require accurate estimates of variance (Bukhman *et al.*, 2008; Daly *et al.*, 2008; Karpievitch *et al.*, 2009; Oberg *et al.*, 2008). Without proper identification of statistical outlier runs the estimates of variance will be inflated, which may have a considerable effect on the identification of significant peptides and proteins.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

ACKNOWLEDGEMENTS

Portions of this work were performed in the Environmental Molecular Science Laboratory, Office of Biological and Environmental Research, National Scientific User Facility at Pacific Northwest National Laboratory (PNNL) in Richland, WA, a U.S. Department of Energy (DOE).

Funding: This work was supported by the National Institutes of Health, the National Institute of General Medical Sciences (R01 GM084892 to B.-J.M.W.-R.); the National Institute of Environmental Health Sciences (U54 ES 016015 to J.G.P.); and the National Institute of Allergy and Infectious Disease (HHSN272200800060C to K.M.W.). PNNL is operated by Battelle Memorial Institute for the DOE under contract number DE-AC05-76RLO1830.

Conflicts of Interest: none declared.

REFERENCES

Anderson,K.K. *et al.* (2006) Estimating probabilities of peptide database identifications to LC-FTICR-MS observations. *Proteome Sci.*, **4**, 1.

Barnett,V. and Lewis,T. (1994) *Outliers in Statistical Data*. John Wiley & Sons, West Sussex, England.

Bukhman,Y.V. *et al.* (2008) Design and analysis of quantitative differential proteomics investigations using LC-MS technology. *J. Bioinform. Comput. Biol.*, **6**, 107–123.

Caroni,C. and Prescott,P. (1992) Sequential application of Wilks's multivariate outlier test. *J. R. Stat. Soc. Ser. C (Appl Stat)*, **41**, 355–364.

Cho,H, *et al.* (2008) OutlierD: an R package for outlier detection using quantile regression on mass spectrometry data. *Bioinformatics*, **24**, 882–884.

Croux,C. and Ruiz-Gazen,A. (2005) High breakdown estimators for prinicpal components: the projection-pursuit approach revisited. *J. Multivariate Anal.*, **95**, 206–226.

Daly,D.S. *et al.* (2008) Mixed-effects statistical model for comparative LC-MS proteomics studies. *J. Proteome Res.*, **7**, 1209–1217.

Dixon,W.J. (1950) Analysis of extreme values. *Ann. Math. Stat.*, **21**, 488–506.

Filzmoser,P. *et al.* (2008) Outlier identification in high dimensions. *Comput. Stat. Data Anal.*, **52**, 1694–1711.

Grubbs,F.E. (1950) Sample criteria for testing outlying observations. *Ann. Math. Stat.*, **21**, 27–58.

Hawkins,D.M. (1980) *Identification of Outliers*. Chapman and Hall, New York, NY.

Hoaglin,D.C. *et al.* (2000) *Understanding Robust and Exploratory Data Analysis*. John Wiley & Sons, Inc., New York.

Jain,R.B. (2010) A recursive version of Grubbs' test for detecting multiple outliers in environmental and chemical data. *Clin. Biochem.*, **43**, 1030–1033.

Jaitly,N. *et al.* (2009) Decon2LS: An open-source software package for automated processing and visualization of high resolution mass spectrometry data. *BMC Bioinformatics*, **10**, 87.

Karpievitch,Y. *et al.* (2009) A statistical framework for protein quantitation in bottom-up MS-based proteomics. *Bioinformatics*, **25**, 2028–2034.

Kauffmann,A. *et al.* (2009) arrayQualityMetrics–a bioconductor package for quality assessment of microarray data. *Bioinformatics*, **25**, 415–416.

Kemmeren,P. *et al.* (2005) Predicting gene function through systematic analysis and quality assessment of high-throughput data. *Bioinformatics*, **21**, 1644–1652.

Lee,E.K. *et al.* (2006) arrayQCplot: software for checking the quality of microarray data. *Bioinformatics*, **22**, 2305–2307.

Li,G. and Chen,Z. (1985) Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo. *J. Am. Stat. Assoc.*, **80**, 759–766.

MacCoss,M.J. *et al.* (2003) A correlation algorithm for the automated quantitative analysis of shotgun proteomics data. *Anal. Chem.*, **75**, 6912–6921.

Mahalanobis,P. (1936) On the generalized distance in statistics. *Proc. Indian Natl Sci. Acad.*, **12**, 49–55.

Metz,T.O. *et al.* (2008) Application of proteomics in the discovery of candidate protein biomarkers in a diabetes autoantibody standardization program sample subset. *J. Proteome Res.*, **7**, 698–707.

Monroe,M.E. *et al.* (2007) VIPER: an advanced software package to support high-throughput LC-MS peptide identification. *Bioinformatics*, **23**, 2021–2023.

Oberg,A.L. and Vitek,O. (2009) Statistical design of quantitative mass spectrometry-based proteomic experiments. *J. Proteome Res.*, **8**, 2144–2156.

Oberg,A.L. *et al.* (2008) Statistical analysis of relative labeled mass spectrometry data from complex samples using ANOVA. *J. Proteome Res.*, **7**, 225–233.

Piening,B.D. *et al.* (2006) Quality control metrics for LC-MS feature detection tools demonstrated on Saccharomyces cerevisiae proteomic profiles. *J. Proteome Res.*, **5**, 1527–1534.

Rocke,D.M. and Woodruff,D.L. (1996) Identification of outliers in multivariate data. *J. Am. Stat. Assoc.*, **91**, 1047–1061.

Rocke,D. *et al.* (2009) Papers on normalization, variable selection, classification or clustering of microarray data. *Bioinformatics*, **25**, 701–702.

Rudnick,P.A. *et al.* (2010) Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses. *Mol. Cell Proteomics*, **9**, 225–241.

Schulz-Trieglaff,O. *et al.* (2009) Statistical quality assessment and outlier detection for liquid chromatography-mass spectrometry experiments. *BioData Min.*, **2**, 4.

Smith,R.D. *et al.* (2002) An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics*, **2**, 513–523.

Stead,D.A. *et al.* (2008) Information quality in proteomics. *Brief. Bioinform.*, **9**, 174–188.

Webb-Robertson,B.M. *et al.* (2010) Combined statistical analysis of peptide intensities and peptide occurrences improves identification of significant peptides from MS-based proteomics data. *J. Proteome Res.*, **9**, 5748–5756.

Wilson,C.L. and Miller,C.J. (2005) Simpleaffy: a BioConductor package for affymetrix quality control and data analysis. *Bioinformatics*, **21**, 3683–3685.

Xia,Q. *et al.* (2006) Quantitative proteomics of the archaeon Methanococcus maripaludis validated by microarray analysis and real time PCR. *Mol. Cell Proteomics*, **5**, 868–881.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016