RESEARCH ARTICLE

Statistical identification of differentially labeled peptides from liquid chromatography tandem mass spectrometry

HyungJun Cho[1, 2], *David M. Smalley*[3, 4, 5], *Dan Theodorescu*[3, 4], *Klaus Ley*[5, 6, 7] *and Jae K. Lee*[8]

[1] Department of Statistics, Korea University, Seoul, Korea
[2] Department of Biostatistics, Korea University, Seoul, Korea
[3] Mellon Prostate Cancer Institute, University of Virginia School of Medicine, Charlottesville, Virginia, USA
[4] Department of Urology, University of Virginia School of Medicine, Charlottesville, Virginia, USA
[5] Robert M. Berne Cardiovascular Research Center, University of Virginia School of Medicine, Charlottesville, Virginia, USA
[6] Department of Biomedical Engineering, University of Virginia School of Medicine, Charlottesville, Virginia, USA
[7] Department of Molecular Physiology and Biological Physics, University of Virginia School of Medicine, Charlottesville, Virginia, USA
[8] Department of Public Health Sciences, University of Virginia School of Medicine, Charlottesville, Virginia, USA

LC-MS/MS with certain labeling techniques such as isotope-coded affinity tag (ICAT) enables quantitative analysis of paired protein samples. However, current identification and quantification of differentially expressed peptides (and proteins) are not reliable for large proteomics screening of complex biological samples. The number of replicates is often limited because of the high cost of experiments and the limited supply of samples. Traditionally, a simple fold change cutoff is used, which results in a high rate of false positives. Standard statistical methods such as the two-sample *t*-test are unreliable and severely underpowered due to high variability in LC-MS/MS data, especially when only a small number of replicates are available. Using an advanced error pooling technique, we propose a novel statistical method that can reliably identify differentially expressed proteins while maintaining a high sensitivity, particularly with a small number of replicates. The proposed method was applied both to an extensive simulation study and a proteomics comparison between microparticles (MPs) generated from platelet (platelet MPs) and MPs isolated from plasma (plasma MPs). In these studies, we show a significant improvement of our statistical analysis in the identification of proteins that are differentially expressed but not detected by other statistical methods. In particular, several important proteins – two peptides for β-globin and three peptides for von Willebrand Factor (vWF) – were identified with very small false discovery rates (FDRs) by our method, while none was significant when other conventional methods were used. These proteins have been reported with their important roles in microparticles in human blood cells: vWF is a platelet and endothelial cell product that binds to P-selectin, GP1b, and GP IIb/IIIa, and β-globin is one of the peptides of hemoglobin involved in the transportation of oxygen by red blood cells.

Received: December 20, 2006
Revised: July 2, 2007
Accepted: July 2, 2007

Keywords:
False discovery rate / Isotope-coded affinity tag / Liquid chromatography / Local pooled error / Paired data

Correspondence: Professor Jae K. Lee, Division of Biostatistics and Epidemiology, Department of Public Health Sciences, University of Virginia School of Medicine, PO Box 800717, Charlottesville, VA 22908, USA
E-mail: jaeklee@virginia.edu
Fax: +1-434-924-8437

Abbreviations: FDR, false discovery rate; **ICAT,** isotope-coded affinity tag; **LOWESS**, locally weighted scatterplot smoothing; **LPE,** local pooled error; **MP,** microparticle; **vWF,** von Willebrand factor

1 Introduction

One of the critical demands in current proteomic research is the comparison of two or more complex samples in order to determine which proteins are differentially expressed. While such investigations were initially performed with no peptide labeling, many recent studies differentially label either proteins or peptides with isotope labels, combine the samples, and then perform LC-MS/MS analysis on the mixture. A growing number of isotope labels have been employed in recent years, including isotope-coded affinity tag (ICAT), isobaric tags for relative and absolute quantification (iTRAQ), and stable isotope labeling with amino acids in cell culture (SILAC) [1]. Quantification of individual peptides is then obtained from the relative ion intensities of the corresponding pair of isotope peaks by standard MS/MS analysis [2]. This approach greatly reduces the variations in MS data by comparing two different samples within the same analysis for more reliable relative quantification [3]. However, the expression intensities of individual peptides are often highly correlated between the pair-labeled MS data outputs from each MS experimental and instrumentation setting. Therefore, it is important to utilize statistical analysis methods that can directly consider such a correlated structure of pair-labeled LC-MS/MS data for rigorous discovery of differentially expressed peptides.

Currently, after obtaining individual intensity values or ratios on relative intensities of LC-MS/MS data, simple and traditional approaches such as fold change discovery and paired *t*-tests are often used to determine which peptides and proteins are differentially expressed. Since the fold change method uses only the ratios of signal intensities from two samples, it cannot provide rigorous evaluation of statistical significance on identified differentially expressed proteins and is often subject to high false-positive and false-negative error rates. A large number of false positives will be selected for peptides detected at low intensity levels where the S/N is high. On the other hand, many significant changes of high-intensity peptides (with relatively small fold changes) will be missed by simply employing a fold-change cutoff criterion. Conventional statistical methods such as two-sample or paired *t*-tests are derived from the signal intensity differences divided by the error variability estimates that are obtained based on each individual protein's intensity values in replicated experiments and samples. For example, statistical analysis was conducted using the two-sample *t*-test to examine differential expression in LC-MS/MS data of the paired nipple aspirate fluid samples from 18 women from tumor-bearing and contralateral disease-free breasts of patients with unilateral early-stage breast cancer, which were differentially labeled with ICAT reagent [2]. While this approach can provide statistical significance of differential expression of MS data, these statistics are often unreliable and severely underpowered because the error estimates are inaccurate, especially when there are a small number of replicates, *e.g.*, duplicate or triplicates [11].

Other statistical methods have also been employed in proteomic data analysis. An unsupervised learning technique was used to cluster spectra with LC-MS/MS data [4] and the analysis of variance (ANOVA) was used to examine the dependence of samples on proteins using 2-DE data [5]. Various supervised techniques have been used to analyze MALDI-TOF MS data [6–10]. While these methods allow statistical evaluation on the differential expressions of thousands of candidate peptides, they have been found to be still underpowered for the MS data analysis with a small number of replicates.

The local pooled error (LPE) test is specifically designed to strengthen statistical power originally in the analysis of small sample microarray data [11]. This approach has also been successfully applied to another proteomics study in determining the differences between plasma control and disease samples by parallel analysis of unlabeled protein samples analyzed by triplicate LC-MS/MS data [12]. In our current study, explicitly considering the correlative characteristics of pair-labeled MS data, we further refine the concept of the LPE method and introduce a weighted pooled test, L_w, for the analysis of paired isotope-labeled samples. We first compare L_w with the alternative statistical approaches by an extensive simulation study. We then use it to compare proteome of platelet microparticles (MPs) and plasma MPs to discover several novel protein biomarkers that were differentially expressed between the two MP proteomes.

2 Materials and methods

2.1 Paired *L*-statistic

The paired *t*-test can be a powerful statistical method for identification of differentially-expressed proteins in LC-MS/MS data when normality is assumed and a good number of replicates, *e.g.*, >10 are obtained. The paired *t*-test can be summarized as follows. Suppose x_{ij} and y_{ij} are (a peptide's) ion intensities for two conditions x and y, where replicates $i = 1, 2, \ldots, n$ and peptides (or proteins) $j = 1, 2, \ldots, m$. Note that prior to analysis the data may be $\log_2$-transformed in order to remedy the highly right-skewed distribution of protein intensity values [13]. Then for each protein *j*, the paired *t*-test statistic is:

$$t_j = \frac{d_{\mathrm{j}}}{\sqrt{(s_{\mathrm{j}}^2/n)}}$$

where $d_{\mathrm{ij}} = (x_{\mathrm{ij}} - y_{\mathrm{ij}})$, $d_{\mathrm{j}} = \sum_{\mathrm{i}} (x_{\mathrm{ij}} - y_{\mathrm{ij}})/\mathrm{n}$, and $s_{\mathrm{j}}^2 = \sum_{\mathrm{i}} (d_{\mathrm{ij}} - d_{\mathrm{ij}})^2/(n-1)$. The statistical significance of each protein can then be obtained from the observed *t*-statistics of which the *p*-value is often adjusted for multiple comparisons [14, 15]. Note that the sample variance s_j^2 is derived based only on the replicated observations of peptide *j*, which can be considerably variable and inaccurate with a small sample size. Due to this, the paired *t*-test is often underpowered and unreliable when data is lowly replicated. Thus, we modify the paired *t*-test and propose the so-called *L*-statistic in order to more reliably identify differentially expressed peptides from pair-labeled LC-MS/MS data. Similar to the paired *t*-test, our *L*-statistic is defined as:

$$L_{\mathrm{j}} = \frac{\delta_{\mathrm{j}}}{\sqrt{\tau_{\mathrm{j}}^2}},$$

where δ_j is the median of paired differences $[(x_{ij} - y_{ij}), i = 1, 2, \ldots, n]$, which reduces the effect of outliers. Borrowing the error information of adjacent-intensity proteins, variance (τ_j^2) is estimated based on LPE estimates in the following manner. Evaluate the ranks of (x_{ij}) and (y_{ij}). If the rank of x_{ij} is more than an α-percent, *e.g.*, 5% away from the rank of y_{ij}, the pair (x_{ij}, y_{ij}) will be excluded from the following calculation of the baseline error distribution (they are considered as potentially differentially expressed peptides).

(i) Let $M_{ij} = x_{ij} - y_{ij}$ and $A_{ij} = (x_{ij} + y_{ij})/2$ be the difference and average of x_{ij} and y_{ij}, respectively. Compute each pooled sample variance of M on each of the predetermined quantiles of A, *e.g.*, 100 1% quantile intervals ($q = 0.01$).

(ii) Obtain the baseline error distribution by applying a nonparametric smoothing spline technique to the above pooled sample variances.

This three-step estimation of LPE is used to avoid inaccurate error estimates in high intensity regions where a direct nonparametric estimation based on fixed-width windows often results in poor error estimation due to sparse observations. Once such a baseline error distribution is obtained, an estimate of the variance of each protein, τ_j^2 can then be extrapolated from this distribution corresponding to the mean of $A_{1j}, A_{2j}, \ldots, A_{nj}$.

By pooling information from other proteins with similar expression levels, this algorithm should provide reliable estimates of baseline variances with limited replication [11]. To determine a threshold of *L*-statistics for statistical significance false discovery rates (FDRs) can be used as described later.

2.2 Paired L_w-statistic

The variance estimate for the above *L*-statistic is based solely on the pooled error variance of adjacent intensity proteins. While the LPE estimate has a shrinkage effect toward the mean of (local) error variances, this effect is not sensitive enough to capture the innate biological variability of individual proteins among different biological subjects. Thus, in order to optimize the error estimates between individual and LPEs, we introduce the L_w-statistic which uses a weighted variance estimate between the two variance estimates. That is, the L_w-statistic on the paired LC-MS/MS data with a weight, *w*, is defined as:

$$L_{\mathrm{wj}} = \frac{\delta_{\mathrm{j}}}{\sqrt{\left((1-w)\tau_{\mathrm{j}}^2 + w s_{\mathrm{j}}^2/n\right)}}$$

where $0 \leq w \leq 1$ is the weight parameter between individual variance estimates or pooling variance estimates. For example, if *w* becomes smaller, the L_w-statistic relies more on individual variances while it relies more on pooling variances if *w* becomes larger. Both variances contribute equally if $w = 0.5$ and $L_w = L$ if $w = 0$.

The weight can be set to 0.5 if no information is available. However, to be more objective, we propose two approaches to estimate the weight. The first uses the randomness of statistics over any intensity level as follows; choose *w* such that $\min_w \{|b_1|: L_{wj} = b_0 + b_1 A_j\}$, where b_0 and b_1 are estimates obtained by regressing A_j on L_w. This approach is based on the assumption that differentially expressed peptides are distributed equally in high and low expression intensity levels. The second approach minimizes L_w-statistics of insignificant peptides as follows, choose *w* such that $\min_w \Sigma_{j \in \mathrm{J}} |L_{wj}|$, where J is a set of insignificant peptides selected by the rank-invariance rule. This rule selects peptides with large rank differences of expression values between two conditions. In practice, we select 50% of rank-invariant peptides and evaluate them for $w = 0, 0.01, 0.02, \ldots, 0.99, 1$.

2.3 Statistical significance by FDR

Raw *p*-values corresponding to the above *L*- or L_w-statistics can be obtained for all observed proteins if an underlying data distribution is assumed to be well-behaved, *e.g.*, a Gaussian distribution. However, this kind of distributional assumption may not be warranted in the ICAT-labeled LC-MS/MS data. In addition, an adjustment of statistical significance is required to take into account a large number of candidate proteins in terms of testing multiple hypothesis comparisons. For this, we control the FDR to determine a threshold of *L*- or L_w-statistics based on a rank-invariant resampling technique as follows:

(i) Order expression values of all replicates and peptides by the mean of x_{ij} and y_{ij} and divide them into *q* quantiles (*e.g.*, $q = 0.01$). Note that pairs for each peptide should be maintained.

(ii) In each quantile, compute ranks of expression values and their differences within each condition, *i.e.*, rank (x_{ij}) – rank (y_{ij}).

(iii) In each quantile, retain replicates having their rank differences less than the median of rank differences, and randomly sample from the remaining pairs to make a null data set with the same size (*m* proteins by *n* replicates) as the raw data.

(iv) Repeat the above procedure independently many times (*e.g.*, $B = 100$).

Using the resampled null data sets, we estimate FDR in the following manner. Suppose *L*-statistics and L^0-statistics are computed from the raw data and the resampled null data sets of size B. Given a critical value Δ, the estimate of FDR is defined as:

$$\mathrm{FDR}(\Delta) = \pi_0(\lambda) R^0(\Delta)/R(\Delta)$$

where $R^0(\Delta) = \#\{L^0_{ib}|L^0_{ib}\geq\Delta,\ i = 1, \ldots, G,\ b = 1, \ldots, B\}/B$ is the average number of differentially expressed peptides in the null data and $R(\Delta) = \#\{L_i|L_i \geq \Delta,\ i = 1, \ldots, G\}$ is the number of significant peptides in the raw data. The estimate of a correction factor with the λ-quantile m_λ of L^0_{ib} is $\pi_0(\lambda) = \#\{L_i|L_i<m_\lambda\}/\#\{L^0_{ib}|L^0_{ib}<m_\lambda\}$ which is required because of the different numbers of true null peptides in the raw data and the null data.

2.4 Simulation data

To investigate the performance of the proposed method, we conducted an extensive simulation study. Since the results of the simulation study rely heavily on the simulation settings, we have tried to assimilate real LC-MS/MS with the following two approaches: case-based and model-based.

2.4.1 Case-based approach

To generate simulated data with individual variations in addition to heterogeneous baseline variations, we utilized a real LC-MS data set. The means and variances within partitioned intervals of intensity levels were thus computed using the real data set described below, and LOWESS (locally weighted scatterplot smoothing) [16] was used to fit the variances against the means. In order to see the effects of different data sets, we used several different sets in our preliminary study, and found that this did not have a significant effect on our overall results (data not shown). The LOWESS line shows an overall decreasing trend (Fig. 1a). Assuming this LOWESS fit represents the true relationship between the mean and variance, we generated simulated data with three replicates from $x_{ij} \sim N(\mu_j, \sigma_{\mu j}^2) + \varepsilon_{ij}$ and $y_{ij} \sim N(\nu_j, \sigma_{\nu j}^2) + \varepsilon_{ij}$ where μ_j was randomly taken among the proteins in the real data for $i = 1, 2, 3$ and $j = 1, 2, \ldots, 1000$, while $\nu_j = \mu_j$ for $j = 1, 2, \ldots, 500$ and $\nu_j = \mu_j + \mathrm{C}$ for $j = 501, 502, \ldots, 1000$. The error term ε_{ij} is generated from $N(0, 1)$ and is added to both x_{ij} and y_{ij}, so x and y have a correlation within each pair i. The constant C is +1 or –1 for a two-fold change and +2 or −2 for a four-fold change because the data are assumed to be $\log_2$-transformed, *i.e.*, $\log_2(2) = 1$ and $\log_2(4) = 2$. The corresponding variances are obtained from the LOWESS line.

2.4.2 Model-based approach

The number of quantified proteins is typically small (*e.g.*, less than 200) because the current quantification of such data is typically performed manually. Therefore, even though the above case-based approach could provide simulated data very close to the actual ICAT-labeled LC-MS/MS data, the following mathematical model was also used to generate simulated data with a finer resolution: $f(\eta) = –0.001\eta + 1.1$, where $\eta = 0.1, 0.2, \ldots, 9.90, 10.0$, closely assimilating the nonlinear decreasing relationship between true mean and variance (Fig. 1b). We generated simulated data with 1000 proteins and three replicates from $x_{ij} \sim N(\mu_j, \sigma_{\mu j}^2) + \varepsilon_{ij}$ and $y_{ij} \sim N(\nu_j, \sigma_{\nu j}^2) + \varepsilon_{ij}$ where μ_j is randomly taken from the η values for $i = 1, 2, 3$ and $j = 1, 2, \ldots, 1000$, while $\nu_j = \mu_j$ for $j = 1, 2, \ldots, 500$ and $\nu_j = \mu_j + \mathrm{C}$ for $j = 501, 502, \ldots, 1000$. The error term ε_{ij} and C were determined as in the case-based approach. The corresponding variances are obtained from the above function, *i.e.*, $\sigma_{\mu j}^2 = f(\mu_j) = -0.001\mu_j + 1.1$ and $\sigma_{\nu j}^2 = f(\nu_j) = -0.001\nu_j + 1.1$.

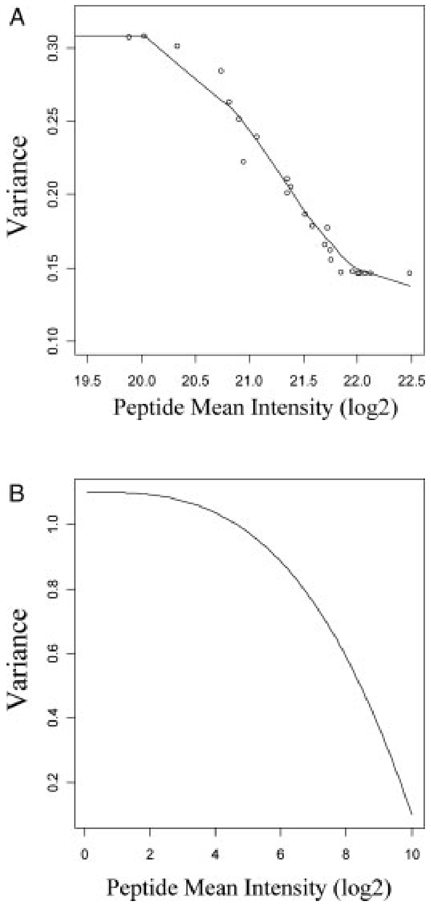


Figure 1. Mean-variance plots for generating simulated data: (A) case-based approach and (B) model-based approach. Simulated data are generated using the assumed relationship of means and variances, *i.e.*, a mean is randomly selected and its corresponding variance is selected; this procedure is repeated independently.

2.5 LC-MS/MS proteomics data for human plasma- and platelet-MPs

Platelets were isolated and platelet MPs were generated and isolated as previously described [17]. Briefly, human blood was collected by venipuncture into 1/10 volume of an acid-citrate-dextrose (85 mM trisodium citrate, 83 mM dextrose, and 21 mM citric acid) solution. Platelet-rich plasma (PRP) was obtained by centrifugation at 110 × *g* for 15 min. Platelets were pelleted by centrifugation at 710 × *g* for 15 min and the supernatant, platelet-poor plasma (PPP), was retained for isolation of plasma MPs (see below). The platelet pellet was washed three times, resuspended in 10 mL of Tyrode's buffer, and centrifuged one additional time at 110 × *g* to remove remaining red blood cells and dead cells. To generate platelet-derived MPs, ADP (10 μM final concentration) was added to the platelet suspension for 10 min. Platelets were removed by centrifugation (710 × *g* for 15 min) and platelet-derived MPs were pelleted by centrifugation at 150 000 × *g* for 90 min at 10°C.

Plasma-derived MPs were isolated by gel filtration chromatography followed by ultracentrifugation as previously described [18]. Briefly, PPP was centrifuged twice to remove residual cells and cell debris at 710 × *g* and 25°C for 15 min. This plasma was then applied to a Sephacryl® S-500 HR (GE Healthcare, Piscataway, NJ) gel filtration column and MP-containing fractions were concentrated by ultracentrifugation at 150 000 × *g* for 90 min at 10°C. Platelet- and plasma-derived MPs were resuspended in a minimal volume of PBS (phosphate-buffered saline, pH 7.4) and a small aliquot was taken for protein analysis using the Micro BCA Protein Assay (Pierce Biotechnology, Rockford, IL). Solutions containing equivalent protein amounts of paired samples (plasma MP and platelet-derived MP) were lyophilized and resuspended in a 1% SDS denaturing buffer from the ICAT labeling kit (Applied Biosystems, Foster City, CA). Samples were labeled and processed as instructed with the following modifications. The initial labeling reaction was performed at half the recommended volume, protein amount, and ICAT reagent because of the low amount of protein obtained from each plasma MP preparation. The differentially-labeled proteins were mixed, and electrophoresed approximately 1 cm into a 7.5% acrylamide SDS-PAGE using a Minigel system (BioRad, Hercules, CA) at 150 V. The acrylamide gel section containing the proteins was cut out as a single lane, extracted, enriched for biotin-containing peptides, and the biotin tag was cleaved as instructed. The samples were lyophilized and reconstituted to 20 μL with 0.1% acetic acid for MS analysis. This procedure was repeated three times with plasma MPs labeled with the light ICAT reagent for two of these samples and labeled with the heavy ICAT reagent in the third.

Samples were loaded onto a 360 μm od × 75 μm id microcapillary fused-silica precolumn packed with irregular 5–20 μm C18 resin. After sample loading, the precolumn was washed with 0.1% acetic acid for 15 min to remove any buffer salts or gel contaminants. The precolumn was then connected to a 360 μm od × 50 μm id analytical column packed with regular 5 μm C18 resin constructed with an integrated electrospray emitter tip. Samples were gradient eluted at a flow rate of 60 nL/min with an 1100 series binary HPLC solvent delivery system (Agilent, Palo Alto, CA) directly through an ESI source interfaced to a Finnigan LTQ mass spectrometer (Thermo Electron, San Jose, CA). The HPLC gradient used was initially 100% A, 5% B at 5 min, 50% B at 220 min, 100% B at 240 min, and restored to 100% A at 280 min (solvent A = 0.1 M acetic acid, solvent B = 70% acetonitrile in 0.1 M acetic acid). The LTQ mass spectrometer was operated in the data-dependent mode in which an initial MS first scan recorded the mass to charge (*m/z*) ratios of ions over the mass range 300–2000 Da, and then the ten most abundant ions were automatically selected for subsequent collisionally activated dissociation and an MS/MS spectrum recorded. All MS/MS data were searched against a human protein database downloaded from the NCBI (www. ncbi.nlm.nih.gov) on Aug 24, 2004 using the SEQUEST® program (Thermo Electron). For ICAT-labeled peptides, a static modification of 227.127 was used for the light isotope label and an additional 9 Da for the heavy ICAT-labeled peptides. Peptide identifications were made based on fully tryptic peptides, using a first-pass filtering of standard criteria as previously described [19], including crosscorrelation values ≥2.0 (+1 charge), 2.2 (+2 charge), and 3.5 (+3 charge). Protein assignments required at least two MS/MS spectra matches that passed the above criteria. Manual validation of at least one MS/MS spectrum–peptide sequence match *per* protein was performed for all proteins that were determined to be differentially expressed.

Ion intensity of peptides was determined using MSight®, freely available from the Swiss Institute of Bioinformatics (www.expasy.org/Msight) [20]. The experiment was conducted using three different samples and 94 peptides from 27 unique proteins were manually quantified.

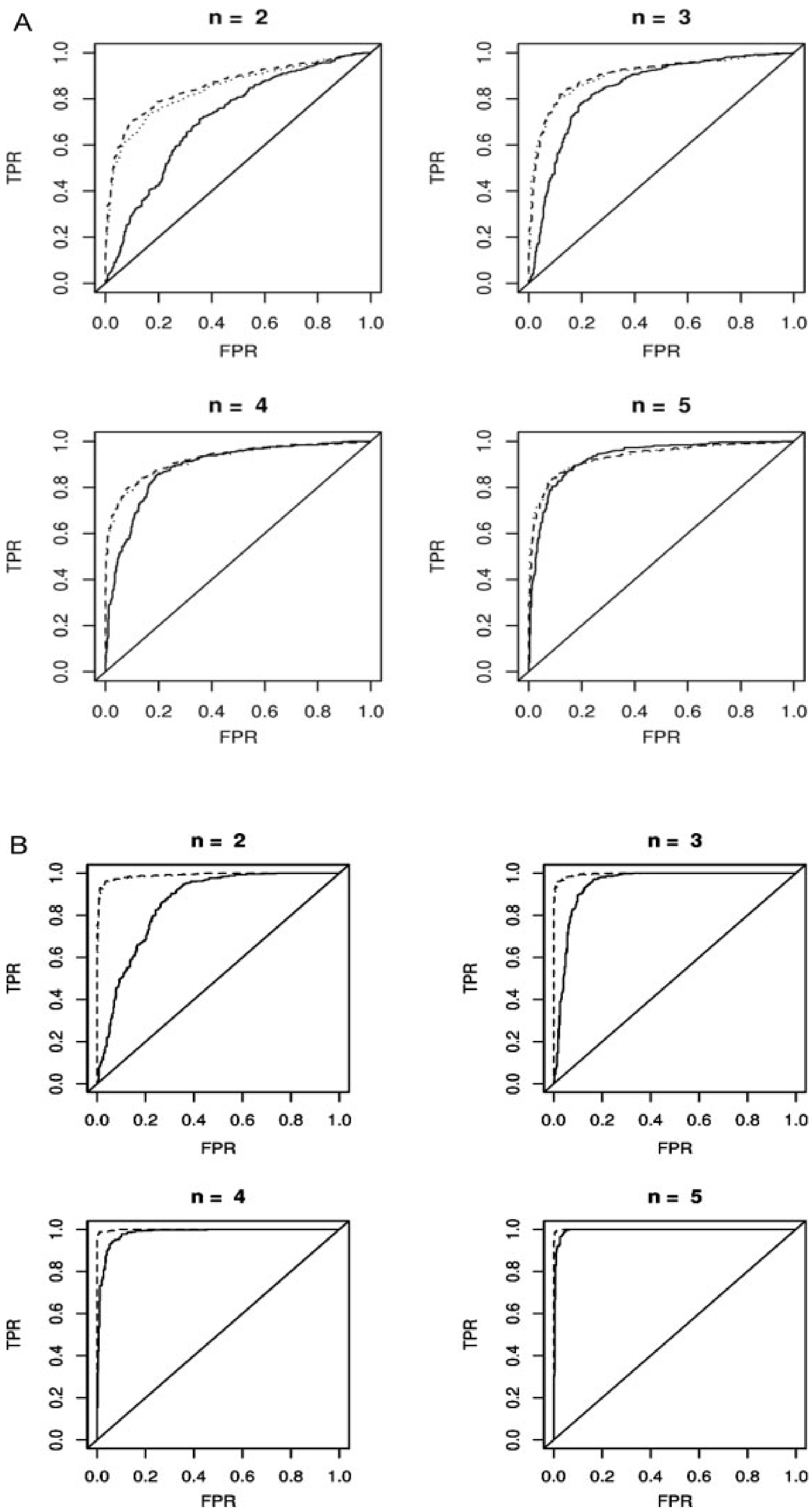
3 Results

3.1 Analysis of simulated data

We compared our proposed methods with the paired *t*-test, the most powerful (unbiased) statistical approach under the normality assumption and the most closely comparable to our approach. Other methods, such as the two-sample *t*-test and fold-change discoveries, were found to perform very poorly for this kind of paired LC-MS/MS data analysis without considering a high correlation between the paired samples of which results are not shown here.

3.1.1 Case-based simulated data

The performance of the paired *t*-test, *L*-test, and $L_{0.5}$-test is shown with the receiver operating characteristics (ROC) curves for the case-based simulated data in Fig. 2a and b. The paired *L*-test and $L_{0.5}$-test significantly outperformed the paired *t*-test when the number of replicates was two or three (*e.g.*, with FPR = 0.05 and $n = 2$, paired *t*-test has a power less than 0.2 compared to power >0.75 of $L_{0.5}$-test); however, the difference between the two tests became smaller with more replicates and there was no significant difference with five or more replicates. The results for the simulated data with four-fold changes were similar to those of the two-fold change cases. Note that *L*-test and $L_{0.5}$-test performed very similarly in these cases – almost identical in the four-fold change cases (Fig. 2b) since their error estimates were very similar under this simulation setting. However, $L_{0.5}$-test would out-


A
n = 2
n = 3
n = 4
n = 5
B
n = 2
n = 3
n = 4
n = 5
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0

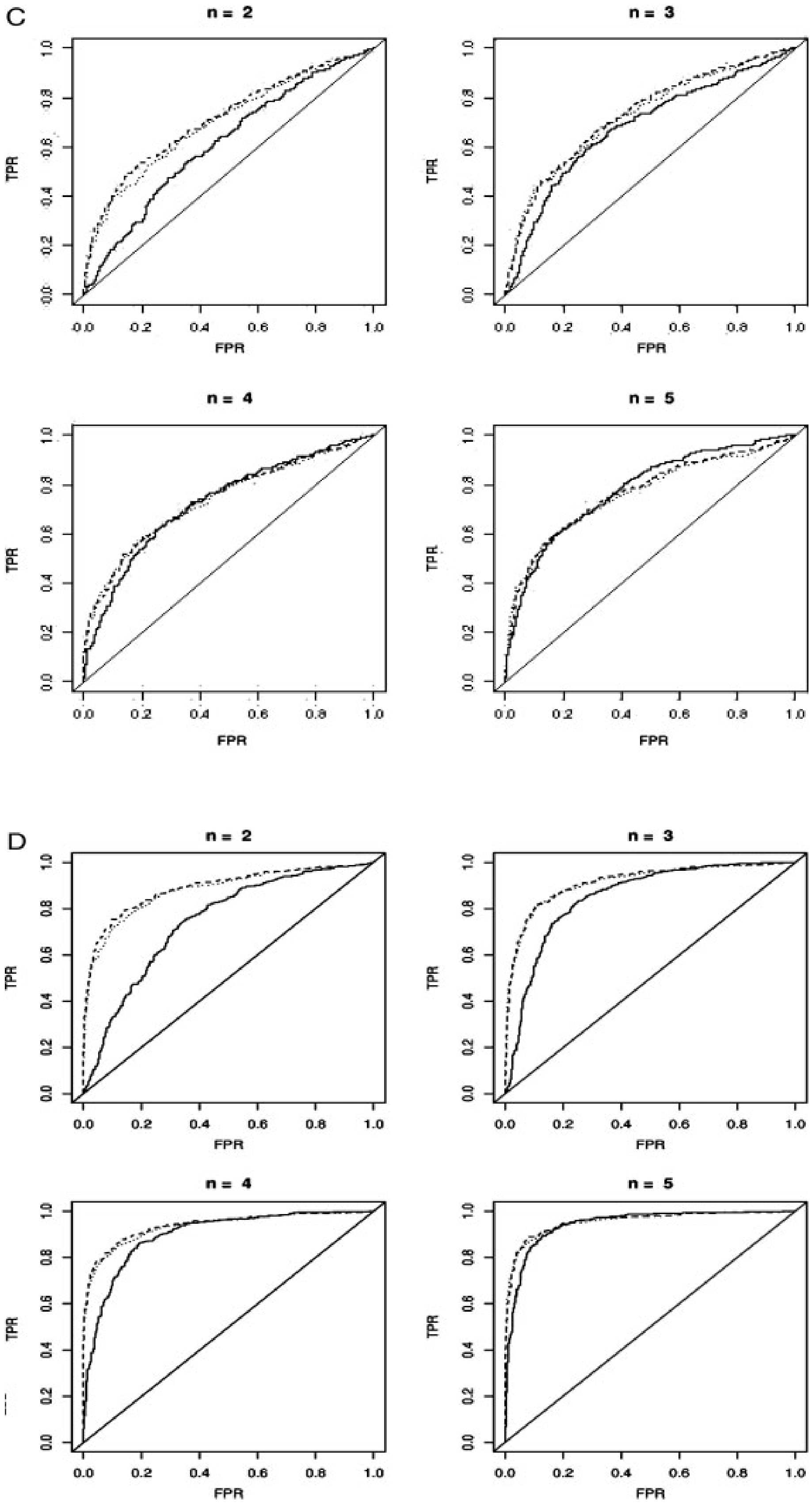


Figure 2. ROC curves for simulated data with $n = 2$, 3, 4, or 5 replicates. The dotted lines (— and ···) represent the paired *L*- and $L_{0.5}$-tests, respectively, the solid line represents the paired *t*-test; (A) case-based simulated data with two-fold change, (B) case-based simulated data with four-fold change, (C) model-based simulated data with two-fold change, and (D) model-based simulated data with four-fold change.

perform *L*-test if more heterogeneous errors exist among the proteins with similar expression intensities from *e.g.*, different biological replicates.

Table 1a shows the averaged areas under the ROC curve (AUCs) of 100 simulations for the various tests applied to the case-based simulated data with two- or four-fold changes. When the data with two-fold changes were analyzed using a small number of replicates, results from the paired *L*- and $L_{0.5}$-test were better whereas the differences between *L* (or L_w) tests and paired *t*-tests were negligible with five or more replicates. Thus the paired *L*-test and $L_{0.5}$ test would be safe to use in both cases. The table also shows AUCs for the paired L_w-tests with estimated weights. Using the L_w-test with the weight estimated by minimizing *L*-statistics of rank-invariant proteins yielded better results than those obtained when the weight was estimated at random. Further, the test yielded results greater than or equal to those obtained using the paired *L*-test. When differentially expressed proteins had four-fold changes, all tests, as expected, performed consistently better than when done with two-fold changes. The comparison results among the tests with four-fold changes could be interpreted with two-fold changes.

3.1.2 Model-based simulated data

Figures 2c and d display the ROC curves for the paired *t*-test, *L*-test, and $L_{0.5}$-test applied to the simulated data generated from the mathematical model with the finer resolution as shown in Fig. 1b. We found that as in the case-based simulated data, the paired *L*-test showed better results than the paired *t*-test with small sample sizes. For example, the paired *t*-test has a power <0.2 compared to $L_{0.5}$-test's power >0.9 with $n = 2$ and FPR = 0.05. These differences became smaller when larger sample sizes were used. Furthermore, when the number of replicates is equal to or greater than five, the results obtained using a paired *t*-test are slightly better than those obtained using the paired *L*-test for the simulated data with two-fold changes. In these cases, there was not much difference between results obtained using the paired *L*-test and L_w-test (with min *L*) because the optimal weight (*w*) for the data was estimated close to zero. Table 1b summarizes the averaged areas under the ROC curve (AUCs) of 100 simulations for the various tests applied to the model-based simulated data with two- or four-fold changes.

3.2 Analysis of human plasma and platelet MPs data

The paired LC-MS/MS data for human plasma MPs consist of three replicates for 94 peptides (27 proteins). Each replicate represents ion abundances for a pair of plasma and platelet MPs. The paired data with a small number of replicates (*e.g.*, $n = 3$) were analyzed using our proposed method, compared with the fold change and the paired *t*-test approaches. In computing fold changes, raw fold changes were adjusted to generate an overall value of 0.00 in the $\log_2$ scale, exclud-

Table 1a. The area under the ROC curve (AUC) generated by each method for the case-based simulated data with two- or four-fold changes and $n = 2, 3, 4, 5$, or 10 replicates

Method (paired)	Two-fold changes					Four-fold changes				
	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 10$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 10$
t-test	0.730	0.839	0.900	0.934	0.990	0.867	0.954	0.983	0.991	0.998
L-test	0.856	0.877	0.936	0.938	0.939	0.991	0.993	0.997	0.997	0.998
$L_{0.5}$-test	0.849	0.876	0.933	0.937	0.988	0.991	0.993	0.997	0.997	0.998
L_w-test (random)	0.851	0.861	0.933	0.934	0.987	0.984	0.988	0.994	0.997	0.998
L_w-test (min *L*)	0.856	0.872	0.942	0.942	0.987	0.991	0.993	0.997	0.997	0.998

The numbers displayed represent the average AUCs of 100 simulations.

Table 1b. The area under the ROC curve (AUC) generated by each method for the model-based simulated data with two or four-fold changes and $n = 2, 3, 4, 5$, or 10 replicates

Method (paired)	Two-fold changes					Four-fold changes				
	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 10$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 10$
t-test	0.622	0.690	0.741	0.784	0.900	0.754	0.857	0.912	0.945	0.992
L-test	0.701	0.711	0.767	0.776	0.884	0.890	0.902	0.949	0.954	0.992
$L_{0.5}$-test	0.694	0.712	0.764	0.775	0.881	0.883	0.901	0.946	0.953	0.992
L_w-test (random)	0.695	0.701	0.758	0.770	0.879	0.879	0.892	0.941	0.949	0.991
L_w-test (min *L*)	0.701	0.711	0.767	0.776	0.884	0.890	0.902	0.949	0.954	0.992

The numbers displayed represent the average AUCs of 100 simulations.

ing von Willebrand factor (vWF) because it was evident that there was a significant enrichment in the plasma MPs. For this set, our estimated weight for L_w-statistic was close to zero, so that our results below are summarized with *L*-statistics.

Figure 3a shows the $\log_2$-ratio of ion abundances for plasma MP and platelet MP against each of 94 peptides that were ordered by the *L*-statistics. The first several peptides were away from the center horizontal line indicating significant difference in ion abundances between plasma MP and platelet MP. For better illustration, the top ten peptides by *L*-statistics are displayed in Fig. 3b including the FDRs of *L*-statistics, FDRs (*q*-values, [21]) of *t*-statistics, and fold changes. Our *L*-statistics identified eight peptides significantly with $\text{FDR}<0.2$. Note that an FDR cutoff criterion can be used with a relatively large value, here 0.2, since such FDRs were rigorously adjusted for the random chances from multiple comparisons. On the contrary, no peptide was identified significantly (with 20% FDR) using the paired *t*-test. Also note that the last two peptides had large fold changes (three-fold changes or higher) even though they were not statistically significant both by the paired *t*-test and *L*-statistics. Therefore, our *L*-statistics were able to rigorously and sensitively identify significantly differentially expressed peptides in this example.

Table 2 summarizes the top ten peptides selected by the *L*-statistics, which were graphically shown with their confidence bounds in Fig. 3b. All ten peptides had a high fold change. However, fold change analysis alone is prone to a high false positive error as illustrated above for the last two peptides. Two peptides for β-globin and three peptides for vWF had very small FDRs when the paired *L*-test was used, while none was significant when the paired *t*-test was used. vWF is a platelet and endothelial cell product that binds to P-selectin, GP1b, and GP IIb/IIIa. β-Globin is one of the peptides of hemoglobin involved in the transportation of oxygen by red blood cells.

4 Discussion

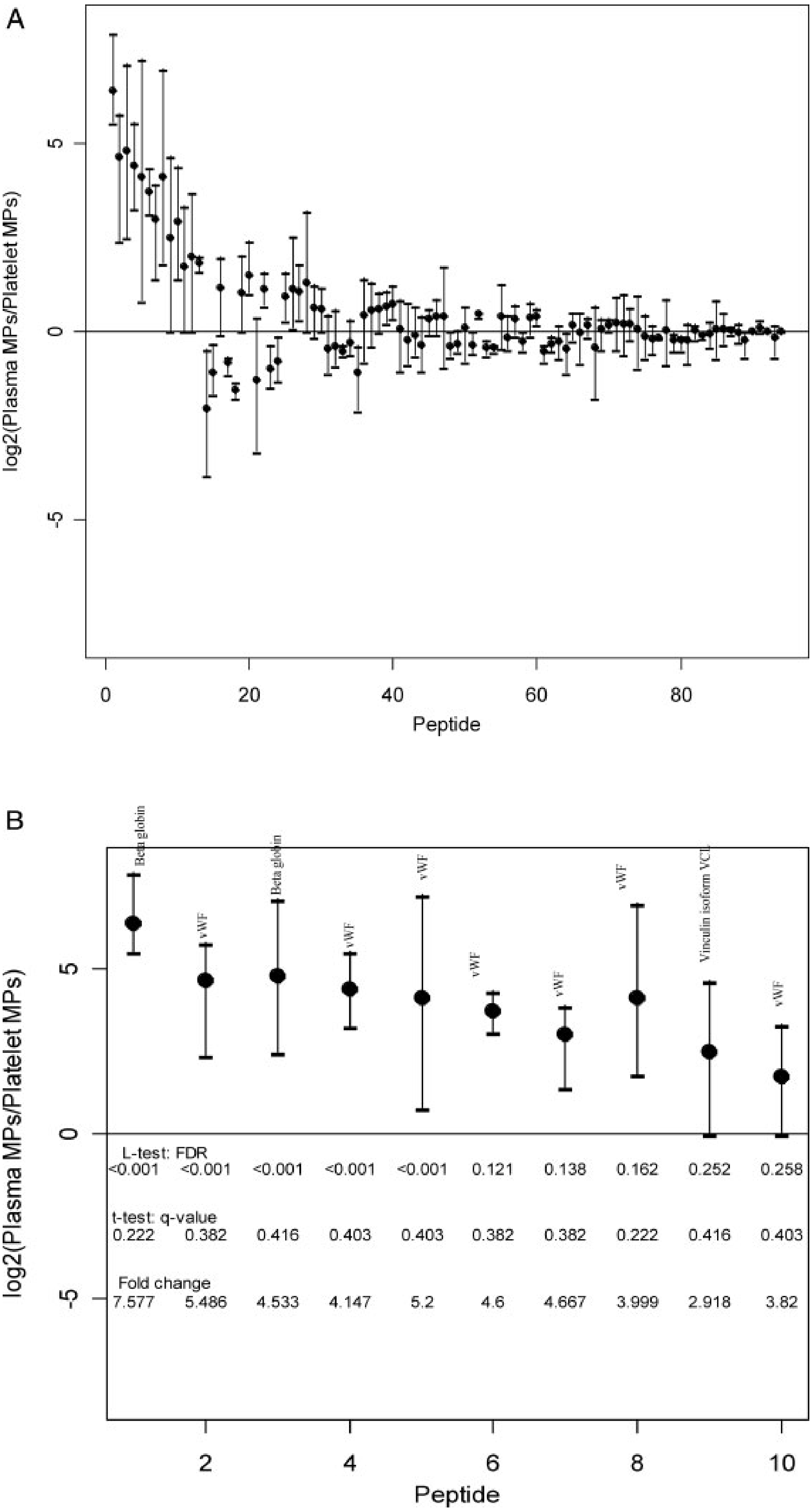
While the current LC-MS/MS technology is quite useful for analyzing purified samples of a small number of targeted proteins, identification and quantification of differentially expressed proteins (among thousands of candidate peptides) are still not reliable for proteomics screening of complex biological samples. In those experiments, replication is often limited because of the high cost of experiments and the limited supply of samples. Our proposed method is particularly useful in analyzing LC-MS/MS data from such experiments with small sample sizes. As shown by our simulation study, the proposed method effectively takes into account error variability arising both from biological replication and technical replication.

We previously developed the LPE test [11], which is useful in identifying differentially expressed genes with low-replicated microarray data. However, it was assumed that the samples are obtained independently from two conditions to apply the LPE test; hence, the LPE test is appropriate only for the analysis of *unpaired* microarray data rather than *paired* MS data such as the human plasma MPs data in our study. Traditional statistical methods such as paired two-sample *t*-tests were found to be statistically underpowered with low replication (*e.g.*, duplicate or triplicate) or non-normal intensities. Thus, we developed the *L*-statistic approach for reliably identifying differentially labeled peptides from LC-MS/MS by accounting for pairing of samples. As shown in our simulation and MP data applications, our newly developed *L*-statistic approach is significantly more powerful than the paired *t*-test in identifying differentially labeled peptides with low-replicated and paired MS data, and enables us to identify important proteins that have not been discovered by other approaches.

For the human plasma MP data, the ion intensities were extracted for 94 peptides because of limitations of the manual use of software. However, using this application, we were still able to show a significant improvement of our statistical analysis and identified several significantly differentially expressed proteins that were not detected using other statistical methods. The development of an automatic extraction technique can generate ion intensities of a much larger number of peptides and proteins, which in turn allows simultaneous examination of more biomarker candidates. Our proposed method can be strengthened and improved using such high-throughput data by borrowing error information from a large number of observations.

It should be noted, however, that if some missing intensity values exist for certain peptides in the paired samples, the paired *t*-test needs to be performed using a reduced number of pairs, even though one observation is available for some pairs, which will result in a less accurate estimation of variances when there is a small number of replicates. Additionally, if only one pair is left ($n = 1$), the paired *t*-test cannot be used. In contrast, the variance estimation for the paired *L*-test is robust even though some values are missing and while it is still desirable to have a larger sample size to generate higher statistical power, the *L*-statistics can effectively evaluate the statistical significance of differentially expressed proteins with a sample size of one. This is possible by borrowing error information from other proteins with similar expression levels. Also note that median statistics (in terms of paired difference evaluation) in our *L*- or L_w-statistics can be replaced by other robust tests such as Huber statistics with appropriate adjustments of variance.

We developed an open source software program (PLPE), which can be conveniently used in the R environment (http://www.r-project.org/). The PLPE package will also be included in the BioConductor website (http:// www.bioconductor.org/). A sample mean is used to estimate the difference of expression intensity levels for the paired *t*-test, which is highly sensitive when there are extreme values or outliers, particularly when the sample size is small. More



L-test: FDR	<0.001	<0.001	<0.001	<0.001	<0.001	0.121	0.138	0.162	0.252	0.258
t-test: q-value	0.222	0.382	0.416	0.403	0.403	0.382	0.382	0.222	0.416	0.403
Fold change	7.577	5.486	4.533	4.147	5.2	4.6	4.667	3.999	2.918	3.82

Figure 3. Peptide ion intensity for the human plasma MPs data; the peptides were ordered by FDRs of the paired *L*-statistics; (A) all 94 peptides and (B) top ten peptides.

Table 2. Analysis of human plasma MPs data. The top ten peptides selected by the paired *L*-test were displayed

Protein	Peptide	$\log_2$ – fold change	Paired *t*-test		Paired *L*-test	
			Statistic	FDR (*q*-value)	Statistic	FDR
β-Globin; hemoglobin β-chain	K.GTFATLSELHCDK.L	7.6	8.324	0.222	4.391	<0.001
vWF precursor	K.APTCGLCEVAR.L	5.5	4.130	0.382	4.342	<0.001
β-Globin; hemoglobin β-chain	R.LLGNVLVCVLAHHFGK.E	4.5	2.081	0.416	3.360	<0.001
vWF precursor	R.CLPSACEVVTGSPR.G	4.1	3.839	0.403	2.688	<0.001
vWF precursor	K.VEETCGCR.W	5.2	2.205	0.403	2.472	<0.001
vWF precursor	K.CLAEGGK.I	4.6	2.709	0.382	2.056	0.121
vWF precursor	R.VTGCPPFDEHK.C	4.7	3.744	0.382	2.082	0.138
vWF precursor	R.SGFTYVLHEGECCGR.C	4.0	10.525	0.222	2.100	0.162
Vinculin isoform VCL	R.KIAELC*DDPK.E	2.9	1.938	0.416	1.676	0.252
vWF precursor	R.TATLCPQSCEER.N	3.8	1.840	0.403	1.713	0.258

p-Values were obtained from the paired *t*-test and FDRs from the paired *L*-test.

robust statistics, such as median and Huber M-estimator [22, 23], would be less sensitive to such extreme values. These different options can be chosen by each individual using our software program. In our current study, we used the median for the *L*- and L_w-statistics. If tighter estimates of expression values than medians are desired despite a cost of greater sensitivity to outliers, the Huber M-estimator can be used (θ_j of $\Sigma_i\varphi[(x_{ij} - y_{ij}) - \theta_j] = 0$ for each j where $\varphi(z) = z$ for $|z| \leq k$ and $\varphi(z) = k$ sign(z) for $|z|>k$ default $k = 1.5$). It should also be noted that we use the rank-invariance rule before estimating pooled variance because variance estimates can be distorted by a small number of highly differentially expressed proteins.

This research was supported by a Korea University Grant of H.Cho, US NIH grant 5R01HL081690 of J. K. L., and the Mellon Prostate Cancer Center and the Robert M. Berne Cardiovascular Research Center at the University of Virginia. We would like to thank the Swiss Institute of Bioinformatics (Geneva, Switzerland) for allowing use of their software, MSight®.

5 References

[1] Ong, S. E., Foster, L. J., Mann, M., Mass spectrometric-based approaches in quantitative proteomics. *Methods* 2003, *29*, 124–130.

[2] Pawlik, T. M., Hawke, D. H., Liu, Y., Krishnamurphy, S. *et al.*, Proteomic analysis of nipple aspirate fluid from women with early-stage breast cancer using isotope-coded affinity tags and tandem mass spectrometry reveals differential expression of vitamin D binding protein. *BMC Cancer* 2006, *6*, 1–10.

[3] Ong, S. E., Mann, M., A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nat. Protoc.* 2006, *1*, 2650–2660.

[4] Beer, I., Barnea, E., Ziv, T., Admon, A., Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics* 2004, *4*, 950–960.

[5] Gustafsson, J. S., Ceasar, R., Glasbey, C. A., Blomberg, A., Rudemo, M., Statistical exploration of variation in quantitative two-dimensional gel electrophoresis data. *Proteomics* 2004, *4*, 3791–3799.

[6] Lee, K. R., Lin, X., Park, D. C., Eslava, S., Megavariate data analysis of mass spectrometric proteomics data using latent variable projection method. *Proteomics* 2003, *3*, 1680–1686.

[7] Markey, M. K., Tourassi, G. D., Floyd, C. E., Jr., Decision tree classification of proteins identified by mass spectrometry of blood serum samples from people with and without lung cancer. *Proteomics* 2003, *3*, 1678–1679.

[8] Wagner, M., Naik, D., Pothen, A., Protocols for disease classification from mass spectrometry data. *Proteomics* 2003, *3*, 1692–1698.

[9] Purohit, P. V., Rocke, D. M., Discriminant models for high-throughput proteomics mass spectrometer data. *Proteomics* 2003, *3*, 1699–1703.

[10] Tatay, J. W., Feng, X., Sobczak, N., Jiang, H. *et al.*, Multiple approaches to data-mining of proteomic data based on statistical and pattern classification methods. *Proteomics* 2003, *3*, 1704–1709.

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

[12] Allet, N., Barrillat, N., Baussant, T., Boiteau, C. *et al.*, *In vitro* and *in silico* processes to identify differentially expressed proteins. *Proteomics* 2004, *4*, 2333–2351.

[13] Yang, M. C., Ruan, Q. G., Yang, J. J., Eckenrode, S. *et al.*, A statistical method for flagging weak spots improves normalization and ratio estimates in microarrays. *Physiol. Genomics* 2001, *7*, 45–53.

[14] Westfall, P. H., Young, S. S., *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment*, John Wiley & Sons, New York 1993.

[15] Benjamini, Y., Hochberg, Y., Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 1995, *57*, 289–300.

[16] Cleveland, W. S., LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *Am. Stat.* 1981, *35*, 54.

[17] Garcia, B. A., Smalley, D. M., Cho, H., Shabanowitz, J. *et al.*, The platelet microparticle proteome. *J. Proteome Res.* 2005, *4*, 1516–1521.

[18] Smalley, D., 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.

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

[20] Palagi, P. M., Walther, D., Quadroni, M., Catherinet, S. *et al.*, MSight: An image analysis software for liquid chromatography-mass spectrometry. *Proteomics* 2005, *5*, 2381–2384.

[21] Storey, J. D., Tibshirani, R., Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.*, 2003, *100*, 9440–9445.

[22] Huber, P. J., Robust estimation of a location parameter. *Ann. Math. Stat.* 1964, *35*, 73–101.

[23] Huber, P. J., *Robust Statistics*, Wiley, New York 1981.