# Label-free, normalized quantification of complex mass spectrometry data for proteomics analysis

**Noelle M Griffin**[1,2], **Jingyi Yu**[1,2], **Fred Long**[1,2], **Phil Oh**[1,2], **Sabrina Shore**[1,2], **Yan Li**[1,2], **Jim A Koziol**[3], and **Jan E Schnitzer**[1,2]

[1]Proteogenomics Research Institute for Systems Medicine, 11107 Roselle Street, San Diego California 92121

[2]Sidney Kimmel Cancer Center, 10905 Road to the Cure, San Diego, CA. 92121

[3]The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA. 92027

Replicate mass spectrometry (MS) measurements and the use of multiple analytical methods can greatly expand the comprehensiveness of shotgun proteomic profiling of biological samples[1-3]. Such MS data contains inherent biases and variations which create computational and statistical challenges for quantitative comparative analysis[4]. We have developed and extensively tested a normalized label-free quantitative method termed the normalized spectral index, $SI_N$, which combines three MS abundance features: peptide and spectral count with fragment-ion (ms/ms) intensity. $SI_N$ largely eliminated variances between replicate MS measurements, permitting quantitative reproducibility and highly significant quantification of thousands of proteins detected in replicate MS measurements of the same and distinct samples. It accurately predicts protein abundance more often than current methods tested. Comparative immunoblotting and densitometry further validate this method. $SI_N$ enables comparative quantification of complex datasets from multiple shotgun proteomics measurements, which is potentially useful for systems biology and biomarker discovery.

Quantitative proteomics is widely used for examining differences in global protein expression between cellular states and in disease biomarker and target discovery[3, 5-8]. These methods are based on one MS feature of abundance such as spectral or peptide count or chromatographic peak area or height values using labelled or label-free approaches (see supplementary notes online for details). How to compare and quantify differential expression remains an important challenge for this field[9, 10]. To date, MS fragment ion

intensities appear only to be used for candidate-based quantification, such as the quantification of small molecules relative to a labelled version of the analyte of interest[11]. A similar approach is single or multiple reaction monitoring (SRM, MRM) where transitions from selected precursor to specific fragment ions are monitored and compared to a spiked standard[12, 13]. Fragment ion intensities are also used in iTRAQ quantification, where the intensity of the reporter fragment ion is directly related to the abundance of the precursor from which it's derived[14]. To date, fragment ion approaches have not been applied in a label-free manner or used in large-scale shotgun proteomics analysis. Here, we explore their utility as an abundance feature.

We previously discovered that multiple MS measurements of a sample are required for large-scale shotgun proteomic platforms to achieve statistically significant comprehensiveness in protein identifications[1] (supplementary notes). This is critical for biomarker discovery where proteins differentially expressed between normal and disease samples can only be meaningfully compared if samples are analyzed systematically and equivalently to completeness. This requires 4-8 MS measurements of each distinct sample[1, 2, 15]. Unfortunately, replicate data contains inherent biases and variations so that MS signals are frequently corrupted by systematic or even apparently random changes (supplementary notes).

We set out to develop and test various methods to quantify, normalize and compare complex label-free proteomic data. We concurrently developed and tested various methods to normalize these features to control for measurement biases and variations. We sought MS features of abundance recorded in all datasets that can be easily extracted, and thus can be universally mined. These include spectral count (SC, number of ms/ms spectra per peptide) and unique peptide number (PN). We also include fragment ion (ms/ms) intensities as a new feature easily extracted from typical MS data and, to our knowledge, not incorporated previously into unlabelled, normalized quantification.

The spectral index (SI) is the cumulative fragment ion intensity for each significantly identified peptide (including all its spectra) giving rise to a protein and is defined as:

$$\text{SI} = \sum_{k=1}^{pn} \left( \sum_{j=1}^{sc} ij \right) k \quad \text{eq. 1}$$

*sc*: spectral count for the peptide *k*,

*i*: fragment ion intensity of peptide *k*,

*j: j*th spectral count of *sc* total spectral counts for peptide *k*

*pn*: number of peptides identified for that protein,

Therefore, this equation inherently incorporates fragment ion intensity values with SC and PN for each protein.

To test the reproducibility of the raw MS abundance features, we graphed the mean diamonds and confidence circles (see online methods) of multiple MS measurements of the same liver endothelial plasma membrane sample, with the null hypothesis that all replicates are equal. The mean PN, SC, and SI across datasets were not sufficiently reproducible and showed significant differences (Figure 1a-c), easily visualized by the non-overlapping mean diamonds and confirmed by ANOVA (online methods). Thus, normalization is required to enable meaningful quantitative comparison within and between samples.

We began with a simplistic approach to normalize the MS datasets by using the "housekeeping" protein actin. The SI of each protein was divided by the SI of actin in each MS measurement to yield $SI_{act}$ (eq. 2),

$$SI_{act} = [\sum_{k=1}^{pn} (\sum_{j=1}^{sc} i_j)k] / \sum_{j=1}^{act} i_j \quad \text{eq. 2}$$

This normalization approach was applied to the 5,923 proteins identified in common across all the liver replicate MS measurements. A significant difference was still detected between the replicates (Fig. 1d). The results were similar when we replaced SI with SC or PN, or when we tested different standards or tissue samples (data not shown).

Next we utilized mean SI values to normalize the data. The Mean Protein Intensity (MPI) was calculated by dividing the total SI for all identified proteins $(\sum_{j=1}^{n} SI_j)$ by the total number of proteins identified, n. The SI of each protein was subsequently normalized by MPI:

$$SI_{MPI} = SI / [(\sum_{j=1}^{n} SI_j)/n] \quad \text{eq. 3}$$

Similarly, we incorporated the total SC to generate another Mean Intensity normalization (MI) method (eq. 4). MI was calculated by dividing the total SI for the dataset $(\sum_{j=1}^{n} SI_j)$ by the total SC of the dataset $(\sum_{j=1}^{n} SC_j)$. The SI of each protein was subsequently normalized by MI to yield $SI_{MI}$ for each protein:

$$SI_{MI} = SI / [(\sum_{j=1}^{n} SI_j)/(\sum_{j=1}^{n} SC_j)] \quad \text{eq. 4}$$

However, the replicates were still significantly different (Fig. 1e, f). Equivalent normalization of the SC and PN datasets were even less effective (data not shown).

Next, we incorporated PN, SC and ms/ms intensities of each peptide into subsequent normalizations, because the SI values are dependent on these features. Each individual SI was normalized by either the total PN (PNt) for the protein, p (eq.5), total SC, TSC (eq.6), or global/total intensity, GI (eq. 7):

$$\mathrm{SI}_P = \mathrm{SI} / (\sum_{j=1}^{PNt} p_j) \quad \text{eq. 5}$$

$$\mathrm{SI}_{\mathrm{TSC}} = \mathrm{SI} / \sum_{j=1}^{n} SC_j \quad \text{eq. 6}$$

$$\mathrm{SI}_{\mathrm{GI}} = \mathrm{SI} / \sum_{j=1}^{n} SI_j \quad \text{eq. 7}$$

The replicate $\mathrm{SI}_P$ and $\mathrm{SI}_{\mathrm{TSC}}$ datasets were still significantly different (Fig. 1g, h). $\mathrm{SI}_{\mathrm{GI}}$ showed no significant difference between the datasets (Fig. 1i), indicating it succeeded in normalizing the replicate datasets.

Although the $\mathrm{SI}_{\mathrm{GI}}$ method provided a dramatic improvement, we aimed for further enhancement. Large proteins can contribute more peptides than smaller ones, thus their abundance may be overestimated[16, 17]. To correct for protein length (number of amino acids), we first normalized SI by protein length ($\mathrm{SI}_L$) (eq. 8, Supplementary Table 1), resulting in improvement over SI alone (Fig. 1j), but significant differences were still evident between the samples. As $\mathrm{SI}_{\mathrm{GI}}$ successfully normalized different samples, we incorporated protein length into this method, resulting in normalized SI ($\mathrm{SI}_N$):

$$\mathrm{SI}_N = \mathrm{SI}_{\mathrm{GI}} / L \quad \text{eq. 9}$$

No significant difference could be detected between the $\mathrm{SI}_N$ normalized datasets (Fig. 1k) and $\mathrm{SI}_N$ was superior to $\mathrm{SI}_{\mathrm{GI}}$.

We applied the $\mathrm{SI}_N$ normalization similarly to the SC datasets, by substituting SI with SC to yield $\mathrm{SC}_N$.

$$SC_N = SC / [\sum_{j=1}^{n} SC_j / L] \quad \text{eq. 10}$$

$\mathrm{SC}_N$ failed to adequately reduce the variation between the datasets (Fig. 1l), showing the superiority of SI over SC. As expected, PN was even worse (data not shown).

Next, we compared $\mathrm{SI}_N$ to two published SC methods, NSAF[18] and Rsc[19]. NSAF failed to adequately normalize our replicate liver datasets (Fig. 1m). Substituting SI for SC in the NSAF approach proved much better but there was still a significant difference between the

replicates (data not shown). Rsc proved better than NSAF in reducing the variation but was inferior to the $SI_N$ method, (Fig 1n). When SC was replaced by SI in the Rsc equation ($R_{SI}$), SI yet again outperformed SC, demonstrating the substantial improvement that can be gained by using SI over SC, regardless of the normalization approach (Fig. 1n).

To validate $SI_N$ as an abundance feature, we performed spiking experiments with a mixture of 19 protein standards across a wide dynamic range (0.5-50,000fmol), with either BSA (2.65μg) or a complex plasma membrane fraction (40μg) as background (Supplementary Methods). The $SI_N$ for each of the standard proteins was calculated and plotted as a function of protein load. $R^2 = 0.9239$ (Fig. 2a, Supplementary Fig. 2a). The slope of the regression line is 1.223 (95% CI of 1.101 to 1.345), meaning the magnitude of $SI_N$ for any given change in protein abundance can be calculated (Supplementary Data).

To determine how $SI_N$ compares to the most commonly used abundance features, namely SC and AUC, we analyzed the ability of $SI_N$, SC and various AUC methods[19, 20], to accurately predict the amount of each protein in "the standard protein mixture"[21] (Supplementary Methods, Notes), a published dataset generated by ISB. For 13 of the 16 proteins in the standard mixture, no significant difference could be determined between the $SI_N$ predicted amount and the actual amount (Fig. 2 b, c, e). For the remaining 3 proteins, $SI_N$ came closest to predicting the actual protein amount 2 out of 3 times (Fig. 2d). Thus, $SI_N$ accurately predicted protein amount 81.25% and was the best method at predicting protein amount 93.75% of the time.

The next best method was total AUC, which was accurate 50% of the time (56.25%, including the one time AUC outperformed $SI_N$ when none of the methods were significantly accurate). The other AUC methods[19, 20] faired just as poorly as the "raw" AUC method in accurately determining the protein amount (18.75%, 43.75%, respectively). SC predicted correct protein abundance for the proteins only 37.5% of the time.

To determine whether $SI_N$ could control for variation in sample load, we compared two different MS datasets taken from the same sample, but analyzed different protein amounts. Proper normalization should scale the individual 40 and 150μg samples, facilitating a direct comparison based on relative abundance. Before normalization, the mean SI values between the 40 and 150μg samples were clearly and significantly different (Fig. 3a). First, we corrected the SI values for the 40μg sample by the dilution factor, SI40*150/40 (Fig. 3b). However, the samples were still significantly different, (t-ratio 24.459). The Rsc method was more effective than the NSAF method but significant variation was still apparent (Fig. 3d, e). $SI_N$ was the only method tested that eliminated the variation introduced by dissimilar or even unknown protein loads, and thereby facilitated comparisons of proteins between these samples (Fig. 3c). When the t-ratios obtained from testing $SI_N$ and Rsc are plotted at incremental peptide cut-off levels (significantly identified peptides), $SI_N$ consistently outperforms Rsc as each t-ratio falls below the significance level of 2, meaning that no significant difference can be found between the $SI_N$ normalized datasets (Fig. 3f).

Next, we used $SI_N$ values to estimate protein amounts (online methods). Linear regression analysis of the ng amounts for each of the 2,660 common proteins from the 40 and 150ug

samples fit a straight line ($R^2$>0.94) with a slope of 3.72, in excellent agreement with the 150/40 ratio of 3.75 (Fig. 3g). This test provides additional strong statistical validation for the applicability of the $SI_N$ method to thousands of proteins for large-scale quantification of protein expression.

As a more stringent test, we determined whether the $SI_N$ method could facilitate unsupervised hiercherical clustering of biologically distinct datasets to identify correlated expression patterns. Using endothelial cell plasma membranes isolated from kidney and heart, each with five replicate MS measurements, we performed two-way unsupervised clustering (not imparting any prior knowledge onto the dataset) on the complete datasets using $SI_N$ values for all commonly identified proteins (across all datasets), which are represented by rows and tissues by columns (Fig. 3h). The clustering algorithm successfully clustered replicates according to tissue type, whereas distinct samples (heart vs. kidney) visually separated. This confirmed that $SI_N$ is quite successful in maintaining sameness of replicates while exposing the differences of distinct biological samples.

Next, we compared intensities obtained from a Coomassie-stained SDS-PAGE gel (Fig. 4a) to the SI abundance values generated by LC-MS/MS analysis. Each gel slice was treated as a distinct sample with an MS and densitometric measurement to determine relative abundance. The two profiles overlapped substantially for most of the gel (Fig. 4b). This strong correlation verifies further the utility of SI as a quantitative tool. Notably, MS also detected proteins in gel regions containing high molecular weight proteins with little Coomassie staining, consistent with this stain's well-known inability to stain some high molecular weight proteins. The sensitivity and utility of the SI method is readily apparent.

To determine whether we could detect quantitative differences for individual proteins expressed in two distinct samples, we applied the $SI_N$ method to datasets from total lung homogenates (H) versus lung endothelial plasma membrane (P) subfractions. Because the P fraction is physically derived from H, the protein composition of P is a subset of the proteins present in H. To determine those proteins that are enriched in P compared to H, we used $SI_N$ and western blotting to generate P/H expression ratios for individual proteins. Statistical comparison of the P/H ratios for 64 proteins, ranging from low to high abundance, produced a Spearman's Rho correlation coefficient of 0.86, indicating an excellent positive correlation between the two methods (Fig. 4c). As the Bland Altman plot [22] showed a mean difference of 0.09, and the data points fall within 2 s.d. of the mean (95% CI for the difference between the two methods), there is very little evidence to indicate that the quantitative methods are significantly different (Fig. 4d).

In this study, we have developed and tested a number of methods to quantify and normalize complex proteomics data obtained from a variety of MS methodologies. We performed extensive statistical testing and validation of our methods, systematically proving their utility through a wide variety of tests and demonstrating their application to diverse MS datasets. With the $SI_N$ method, we have successfully compensated for experimental and random bias and noise, thus showing that protein abundances, as reflected by mean $SI_N$ from replicate samples, are essentially the same. Conversely, the analysis of biologically distinct

samples, with noise and bias controlled by proper normalization, enables meaningful direct quantitative comparisons reflecting their true biological diversity.

We aimed to generate an abundance index with the convenience of SC, but greater confidence at low peptide numbers without the added complexity of peak area or AUC measurements. We tested the benefit of combining PN, SC and ms/ms ion intensities into one metric, (SI) as opposed to using these features in isolation. This approach proved statistically robust, more so than SC or AUC (Fig. 1, Fig. 2 and Supplementary Fig. 5&6) and obviates the need for spiked samples.

Using the fragment ion intensity, specifically only those intensities that match the peptide of interest as they are inherently more reflective of the precursor, may facilitate more accurate measurements as there is less chance of including the signal of co-eluting precursors or background noise (Supplementary Discussion online). Also, for $SI_N$, it's the peak height of the ms/ms fragments that are summed where as for the AUC, the precursor ions are integrated thus providing more scope for error with the integration process. This is particularly important as most mass spectrometers operate in conjunction with liquid chromatography (LC) systems, thereby producing MS scans permeated by chemical noise. In addition, the chromatogram becomes noisier as the sample complexity increases. For example, even our typical 36hr LC-LC-MS/MS runs for complex samples still produce over-lapping chromatograms due to co-elution peaks, making AUC quantification troublesome. Most groups use LC or LC-LC setups with much shorter elution times. This will continue to be an issue, even in the advent of high-resolution instruments, as improvements in MS resolution can only really be appreciated when the chromatographic resolution improves in tandem, which has yet to be fully realized. This is not an issue for the $SI_N$ calculation as the ms/ms spectra is inherently less complex and no integration is preformed.

$SI_N$, above all other methods tested, could accurately determine the correct amount of each protein standard in a mixture (Fig. 2 b-d, e). Despite giving the AUC methods the best possible advantage (Supplementary Notes), $SI_N$ consistently outperformed them in determining protein abundance (Supplementary Fig. 5&6). SC performed as modestly as the AUC methods. $SI_N$ could accurately determine the relative abundance for thousands of proteins in complex samples, without the need for spiking of protein standards (Fig. 3g, Fig. 4). $SI_N$ also facilitated the identification of a subset of proteins enriched in P relative to H. We showed outstanding correlation between the $SI_N$ and western blot ratios (Fig. 4c&d), validating this enrichment.

As abundance features, SI, SC, PN, and AUC are not reproducible across replicate datasets (Fig. 1 a-c, Supplementary Fig. 4a). Methods of normalizing complex LC-MS/MS data are only just emerging[19], [23]-[27], but no comparison is generally shown between the pre- and post-normalized data. Stringent validation has been rather sporadic and the results and efficacy have been variable[24], [28]-[30]. Therefore, we undertook a systematic and logical approach to normalize our datasets. Of the methods tested, only $SI_{GI}$ and $SI_N$ were successful in removing the variability between replicates (Fig. 1i). For $SI_N$, we added a protein size parameter to the $SI_{GI}$ calculation, resulting in a clear benefit over $SI_{GI}$ alone (Fig. 1i, k). Even when we substituted SI for SC, PN or AUC in the normalization methods,

SI consistently outperformed all other features, regardless of the sample, measurement or normalization approached applied (Fig. 1k, l, o & Supplementary Fig. 4c). Even the raw SI values show less variation between the replicates than the $SI_N$ "normalized" AUC values (Supplementary Fig. 4).

$SI_N$ does not over-normalize, but rather can reduce replicate variability to maintain sameness in datasets from a single sample while maintaining quantitative differences between distinct samples (Fig. 3h). $SI_N$ can also successfully normalize datasets with different loading amounts (Fig. 3c), thus, the dilution factor between samples can be accurately derived from the data simply by calculating the slope of the regression line (Fig. 3g). $SI_N$ can also control for the variation introduced by different MS methodological analysis of the same sample (Supplementary Fig. 3) to facilitate comparison and quantification across all datasets (Supplementary Fig. 7). This may have significant implications for the comparison of datasets acquired in different labs.

In summary, combining and normalizing several MS abundance features, including for the first time, fragment ion intensities, represents a novel approach to MS quantification with broad utility. When we compared our new methods to each other and to previously reported methods[19], [20], the best method was $SI_N$, which was developed through logical systematic application of enabling parameters which, when combined, produced improved normalization and quantification. This method now allows the quantitative comparison of biologically distinct datasets with high confidence and relative ease, and should therefore greatly facilitate the use of label-free quantitative proteomic approaches for differential protein expression analysis. This method is all the more valuable in the era of systems biology and biomarker discovery where distinct samples must be analyzed both quantitatively and comprehensively through the replicate MS measurements necessary to gain confidence in determining expression differences and novel biomarkers.

## Methods

### Mass spectrometry analysis

Proteins were pre-fractionated on SDS-PAGE gels prior to 2D-LC-MS/MS and Reverse Phase-MS/MS. For RP-MS/MS using either an LCQ or LTQ mass spectrometer, digested peptides were extracted from each gel slice and lyophilized. For 2D-LC-MS/MS using an LCQ, peptides extracted from each gel slice were first pooled into 7 groups then lyophilized[2]. Data acquisition from both the LCQ and LTQ was carried out in data-dependent mode. Full MS scan were recorded on the eluting peptides over the 400-1400 m/z range with one MS scan followed by three MS/MS scans of the most abundant ions. A dynamic exclusion was applied for Repeat Count of 2, a Repeat Duration of 0.5 minute, and an Exclusion Duration of 3 min. A dynamic exclusion window was applied for duration of 10 minutes for 2D-LC-MS/MS.

### Database search

The acquired MS/MS spectra were converted into mass lists using the Extract_msn program from Xcalibur and searched against a protein database containing human, rat and mouse

sequences using the Sequest program in Bioworks™ 3.1 for Linux (Thermo Fisher Scientific, Inc., Waltham, MA, USA). The searches were performed allowing for tryptic peptides only with peptide mass tolerance of 1.5 Da for LTQ data, 2.0 Da for LCQ data. Accepted peptide identification was based on a minimum   Cn score of 0.1; minimum cross correlation score of 1.8(z=1), 2.5(z=2), 3.5(z=3). The false positive identification rates were   1% (see supplementary methods). The positive protein identification results were extracted from Sequest.out files, filtered and grouped with DTASelect software using the above criteria. Proteins were identified based on 2 unique significantly identified peptides.

Fragment ion intensity (intensity of the ions in the ms/ms spectrum that are assigned to a given peptide), peptide number (number of unique peptides identifying a protein) and spectral counts (number of ms/ms spectra assigned to a particular peptide) were extracted from the DTAselect output files using a script written in-house (see supplementary methods). For the purpose of this manuscript, fragment ion intensity is defined as the total intensity of all detected *b* and *y* fragment ions (ms/ms spectra) for a specific peptide. The fragment ion intensity of each peptide that passes the threshold for identification that gives rise to a significantly identified protein (see above) is summed. The combination of these summed fragment ion intensities from all ms/ms spectra and peptides relating to a given protein is combined and is referred to as the spectral index (SI) for that protein.

### NSAF and Rsc

NSAF is described by Zybailov *et al*[18] as

$$(\text{NSAF})_k = (\text{Spc}/L)_k / \sum_{i=1} (Spc/L)i$$

where Spc is the spectral count for protein k and L is the length of protein k.

The Rsc is described by Old *et al*[19] as:

$$\text{Rsc} = \log_2 \left[ (n_2 + f)/(n_1 + f) \right] + \log_2 \left[ (t_1 n_{-1} + f)/(t_2 - n_2 + f) \right]$$

where, for each protein, RSC is the $\log_2$ ratio of abundance between Samples 1 and 2; *n*1 and *n*2 are spectral counts for the protein in Samples 1 and 2, respectively; *t*1 and *t*2 are total numbers of spectra over all proteins in the two samples; and *f* is a correction factor set to 0.5 by Beissbarth *et al*[33] and varied in the study by Old *et al*. We used the 1.25 correction factor as per Old *et al*[19].

### Protein abundance calculation

SIN were converted to estimated nanogram amounts, by including the initial sample load in the final calculation using the following equation:

$$RPQ_p = \frac{SI_N}{\sum_{i=1}^{j}(SI_N)_j} * Q * 1000$$

where $j$ = number of all proteins identified with 2 unique peptides, and the subscript $i$ refers to the $i$th protein of $j$ total proteins, and Q is the amount of sample (μg) used in a given measurement.

**Statistics**

JMP IN 5.1 (SAS Institute) was used for all statistical analysis. T-tests and ANOVAs are common statistical tests used for determining difference between sample means but require data to be normally distributed to achieve analytical rigor. Our raw SC, PN and SI datasets were not normally distributed (Supplementary Fig. 1) as measured by the skewness and kurtosis of the frequency distribution. To maintain statistical rigor and to avoid inflated variance, we performed a $\log_{10}$ transformation of our datasets which produce a reasonable normality as determined from the histogram and Q-Q plots (Supplementary Fig. 1). Thus, for comparative statistical analysis, we similarly transformed all the datasets after performing the normalizations described below. It should be noted that equivalent results to those described below were obtained with non-parametric analyses (data not shown).

To visualize normalized datasets, we graphed the mean (center line) and 95% confidence intervals (CI), indicated as diamonds on the graphs, of normalized spectral indexes. If the CIs shown by the mean intervals do not overlap, the groups are significantly different. The reverse is not necessarily true and significance is determined from the summary statistics associated with the analysis (see below). The confidence circles are another way of visualising the diamonds and aids in determining CI overlap.

To determine whether there was any evidence that the replicate values were significantly different before and after application of the normalization methods, we applied a t-statistic (2 replicates) or ANOVA, one-way (>2 replicates) to look for differences in normalized mean abundance features. For the statistical analysis, we used only the proteins that were identified in common across all replicate datasets for a particular comparison. Our null hypothesis was that both (2 replicates) or all (>2 replicates) samples were equal. For the t-statistic (2 replicates), the normalized values were deemed significantly different if a large t-ratio (as determined from the t-tables) and a small P-value (p<0.05) were produced from the t-statistic. We use a t-ratio <2 in absolute value for significance as it approximates the 0.05 significance level. For analysis of difference in mean intensities between multiple replicate (>2) samples, analysis of variance (ANOVA, one-way) was performed. Our null hypothesis was that all replicate samples were equal. If our null hypothesis was true then we expect the F-ratio to be ∼1. (Informally, the smaller the F statistic [equivalently, the larger the p-value], the closer the agreement across the replicates). Our significance level was p<0.05. If there is no statistically significant difference between the replicates (as indicated by F-ratio ∼1) we conclude that the normalization method succeeded in controlling for the variation between the replicate datasets.

**Unsupervised hierarchical clustering**—Cluster analysis was performed on a dataset from 5 replicate MS measurements of endothelial cell plasma membranes isolated from kidney and heart samples using JMP 5.1, and using Wards hierarchical method[34]. Ward's method is a hierarchical method designed to optimize the minimum variance within clusters (minimizes within-group dispersions). The SIN values for each protein was normalized across each row (all 10 samples) using the following standard approach: $(SI_N - (mean\ SI_N)_{row}/(standard\ deviation)_{row})$.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

## References

1. Durr E, et al. Direct proteomic mapping of the lung microvascular endothelial cell surface in vivo and in cell culture. Nat Biotechnol. 2004; 22:985–992. [PubMed: 15258593]

2. Li Y, et al. Enhancing identifications of lipid-embedded proteins by mass spectrometry for improved mapping of endothelial plasma membranes in vivo. Mol Cell Proteomics. 2009; 8:1219–1235. [PubMed: 19155209]

3. Oh P, et al. Subtractive proteomic mapping of the endothelial surface in lung and solid tumours for tissue-specific therapy. Nature. 2004; 429:629–635. [PubMed: 15190345]

4. Wong JW, Sullivan MJ, Cagney G. Computational methods for the comparative quantification of proteins in label-free LCn-MS experiments. Brief Bioinform. 2008; 9:156–165. [PubMed: 17905794]

5. Oh P, et al. Live dynamic imaging of caveolae pumping targeted antibody rapidly and specifically across endothelium in the lung. Nat Biotechnol. 2007; 25:327–337. [PubMed: 17334358]

6. Shiio Y, et al. Quantitative proteomic analysis of Myc oncoprotein function. Embo J. 2002; 21:5088–5096. [PubMed: 12356725]

7. Shiio Y, et al. Quantitative proteomic analysis of myc-induced apoptosis: a direct role for Myc induction of the mitochondrial chloride ion channel, mtCLIC/CLIC4. J Biol Chem. 2006; 281:2750–2756. [PubMed: 16316993]

8. Chiang MC, et al. Systematic uncovering of multiple pathways underlying the pathology of Huntington disease by an acid-cleavable isotope-coded affinity tag approach. Mol Cell Proteomics. 2007; 6:781–797. [PubMed: 17272267]

9. Service RF. Proteomics. Proteomics ponders prime time. Science. 2008; 321:1758–1761. [PubMed: 18818332]

10. Service RF. Proteomics. Will biomarkers take off at last? Science. 2008; 321:1760. [PubMed: 18818333]

11. Kolodziej EP, Gray JL, Sedlak DL. Quantification of steroid hormones with pheromonal properties in municipal wastewater effluent. Environ Toxicol Chem. 2003; 22:2622–2629. [PubMed: 14587901]

12. Wolf-Yadlin A, Hautaniemi S, Lauffenburger DA, White FM. Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. Proc Natl Acad Sci U S A. 2007; 104:5860–5865. [PubMed: 17389395]

13. Kuhn E, et al. Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and 13C-labeled peptide standards. Proteomics. 2004; 4:1175–1186. [PubMed: 15048997]

14. Ross PL, et al. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol Cell Proteomics. 2004; 3:1154–1169. [PubMed: 15385600]

15. Koziol JA, Feng AC, Schnitzer JE. Application of capture-recapture models to estimation of protein count in MudPIT experiments. Anal Chem. 2006; 78:3203–3207. [PubMed: 16643015]

16. Ishihama Y, et al. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. Mol Cell Proteomics. 2005; 4:1265–1272. [PubMed: 15958392]

17. Rappsilber J, Ryder U, Lamond AI, Mann M. Large-scale proteomic analysis of the human spliceosome. Genome Res. 2002; 12:1231–1245. [PubMed: 12176931]

18. Zybailov B, et al. Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. J Proteome Res. 2006; 5:2339–2347. [PubMed: 16944946]

19. Old WM, et al. Comparison of label-free methods for quantifying human proteins by shotgun proteomics. Mol Cell Proteomics. 2005; 4:1487–1502. [PubMed: 15979981]

20. Silva JC, et al. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. Mol Cell Proteomics. 2006; 5:144–156. [PubMed: 16219938]

21. Klimek J, et al. The standard protein mix database: a diverse data set to assist in the production of improved Peptide and protein identification software tools. J Proteome Res. 2008; 7:96–103. [PubMed: 17711323]

22. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986; 1:307–310. [PubMed: 2868172]

23. Callister SJ, et al. Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. J Proteome Res. 2006; 5:277–286. [PubMed: 16457593]

24. Wang W, et al. Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. Anal Chem. 2003; 75:4818–4826. [PubMed: 14674459]

25. Lukas TJ, et al. Informatics-assisted protein profiling in a transgenic mouse model of amyotrophic lateral sclerosis. Mol Cell Proteomics. 2006; 5:1233–1244. [PubMed: 16571896]

26. Forner F, et al. Quantitative proteomic comparison of rat mitochondria from muscle, heart, and liver. Mol Cell Proteomics. 2006; 5:608–619. [PubMed: 16415296]

27. Choi H, Fermin D, Nesvizhskii AI. Significance analysis of spectral count data in label-free shotgun proteomics. Mol Cell Proteomics. 2008; 7:2373–2385. [PubMed: 18644780]

28. Baggerly KA, et al. A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. Proteomics. 2003; 3:1667–1672. [PubMed: 12973722]

29. Wagner M, Naik D, Pothen A. Protocols for disease classification from mass spectrometry data. Proteomics. 2003; 3:1692–1698. [PubMed: 12973727]

30. Anderle M, et al. Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum. Bioinformatics. 2004; 20:3575–3582. [PubMed: 15284095]

31. Kramer CY. Extension of multiple range tests to group means with unequal numbers of replications. Biometrics. 1956; 12:309–310.

32. Tukey JW. Some selected quick and easy methods of statistical analysis. Trans N Y Acad Sci. 1953; 16:88–97. [PubMed: 13169230]

33. Beissbarth T, et al. Statistical modeling of sequencing errors in SAGE libraries. Bioinformatics. 2004; 20(Suppl 1):i31–39. [PubMed: 15262778]

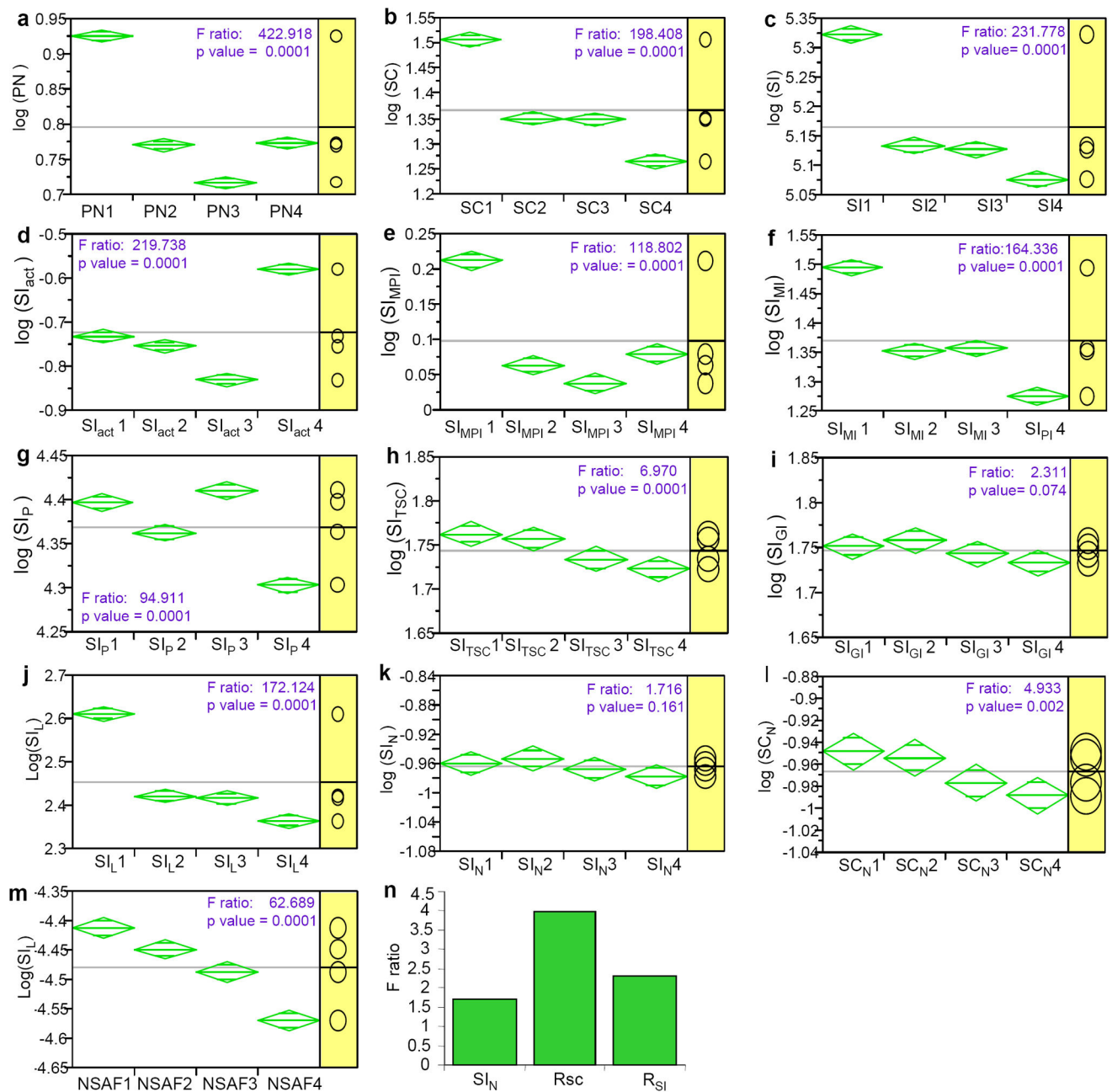34. Kendall, M. Multivariate analysis. 2nd. MacMillan; New York: 1980.

**Figure 1. Statistical analysis of replicate MS measurement variation before and after normalization**

The mean and 95% confidence interval (CI) for the abundance features, peptide number, PN (**a**), spectral count, SC (**b**), and spectral index, SI (**c**) were calculated for 4 MS replicate measurements of pooled endothelial cell plasma membrane isolated from liver and were plotted using the mean diamonds and comparison circles methods. If the CIs, as indicated by the diamonds, do not overlap, the groups are significantly different. For statistical analysis of difference in mean intensities or other features, between multiple replicate samples, analysis of variance (ANOVA, one-way) was performed. Our null hypothesis was that all

replicate samples were equal. If our null hypothesis is true then we expect the F-ratio to be ~1 (d.f.= 5919). Our significance level was $p<0.05$. The x-axis represents each of the 4 replicate datasets and the y-axis represents the log of abundance feature being examined (n=5923). The indicated normalization methods were applied separately to the SI (**d-k**) or SC (**l**) datasets and tested for differences as described above. We applied NSAF[20] and Rsc[19] methods to the replicates datasets (see online methods for equations) and tested for differences (**m,n**). Graph of the comparison of F-ratios obtained from statistical testing of $SI_N$, Rsc and $R_{SI}$, where $R_{SI}$ is the Rsc equation with SI substituted for SC (**n**).
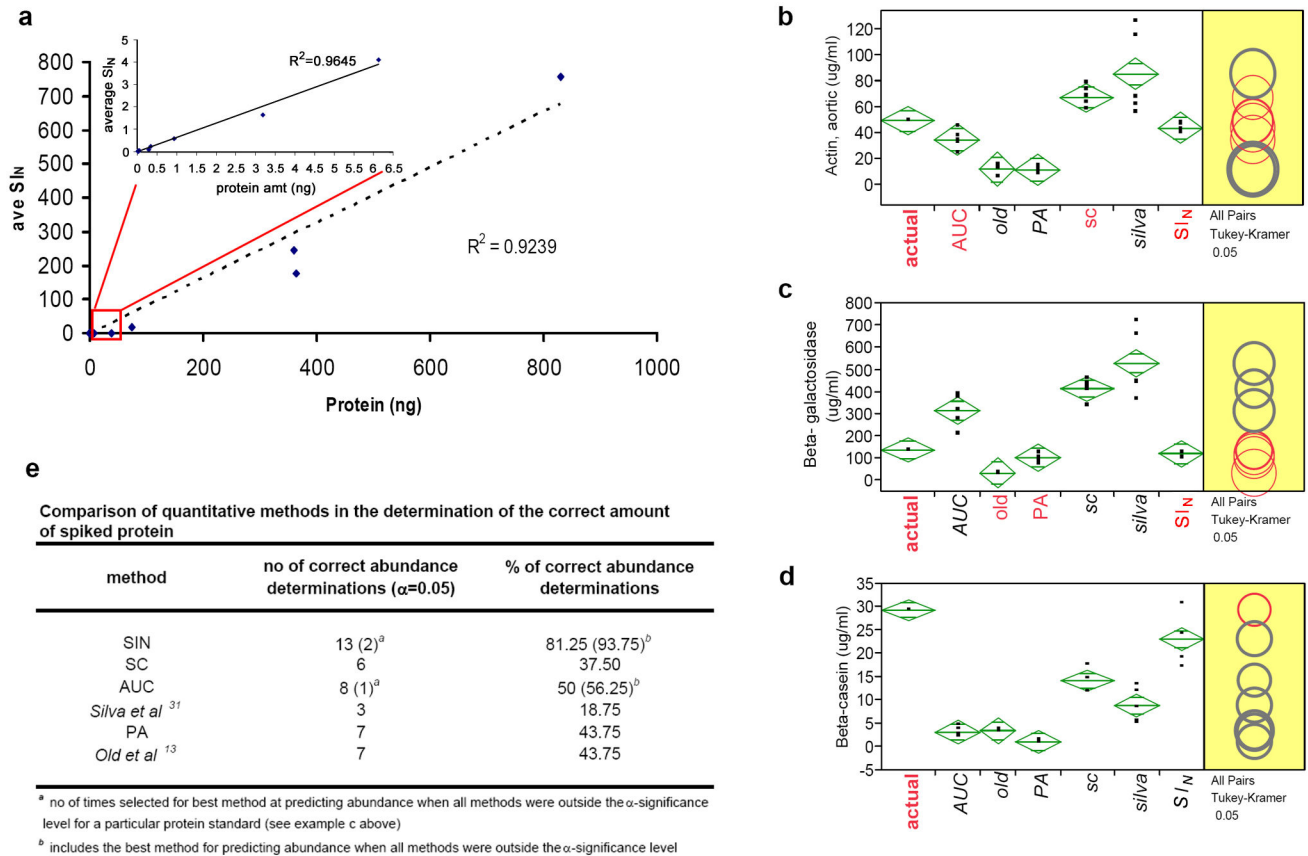
**a**



**b**



**c**



**d**



**e**

Comparison of quantitative methods in the determination of the correct amount of spiked protein

| method | no of correct abundance determinations ($\alpha$=0.05) | % of correct abundance determinations |
|---|---|---|
| SIN | 13 (2)[a] | 81.25 (93.75)[b] |
| SC | 6 | 37.50 |
| AUC | 8 (1)[a] | 50 (56.25)[b] |
| *Silva et al* [31] | 3 | 18.75 |
| PA | 7 | 43.75 |
| *Old et al* [13] | 7 | 43.75 |

[a] no of times selected for best method at predicting abundance when all methods were outside the $\alpha$-significance level for a particular protein standard (see example c above)

[b] includes the best method for predicting abundance when all methods were outside the $\alpha$-significance level

**Figure 2. Correlation of $SI_N$ with protein abundance**

(**a**) A protein standard mix spanning a wide dynamic range (0.5 – 50000fmol) was spiked into BSA, separated by SDS-PAGE, trypsin digested and analyzed by 2DLC. $SI_N$ values for each spiked protein were calculated, averaged and plotted against the amount of the protein standard added. (Note: clustering of many of the data points close to the origin due to the large range in protein abundance; this region was zoom and expanded for ease of visualization). The $R^2$ correlation was 0.9239. (**b-d**) Statistical analysis comparing the quantification of proteins across replicate measurements using 6 quantification methods (relative to known value). The mean and 95% CI for protein abundance, as determined by various relative quantitative methods, were plotted for three representative proteins from the "standard protein mixture" and compared to the actual loaded amount using ANOVA, and individual means were compared using the Tukey-Kramer HSD method [31], [32], Quantitative methods that were not significantly different from the actual protein abundance, (ANOVA, $\alpha$=0.05) are highlighted in red. (**e**) The table summarizes the statistical analysis used to compare the ability of 5 quantitative methods to accurately determine the correct amount of a protein in a standard mixture across replicate datasets. The number of correct abundance determinations was determined using ANOVA where the predicted protein amount, as determined by each method, did not deviate significantly from the mean of the actual protein amount, $\alpha$=0.05 Abbreviations; SC- spectral count, AUC- area under the curve; total AUC for all identifying peptides.
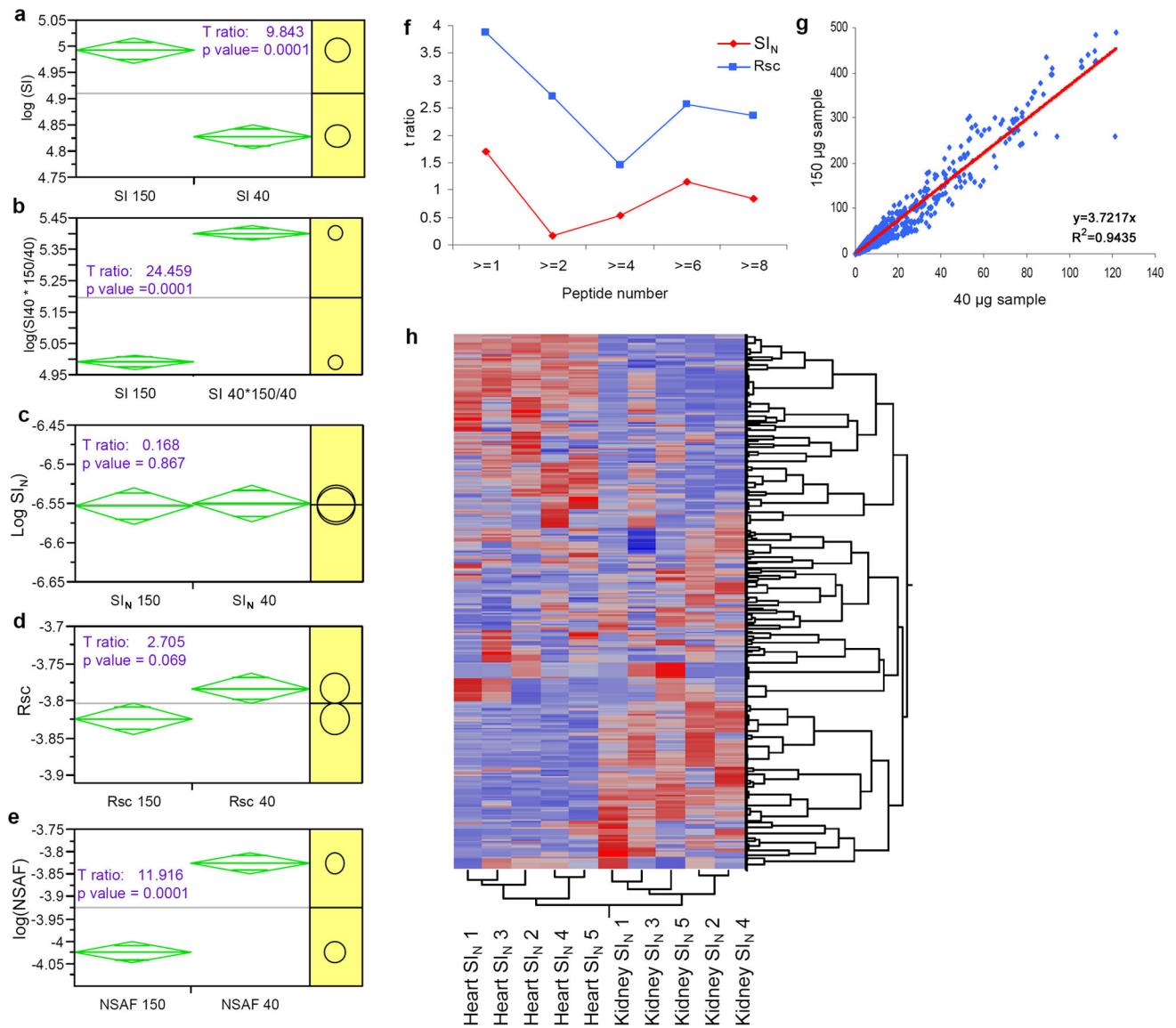
**Figure 3. Statistical analysis of normalization methods applied to variable protein load and distinct sample datasets**

The indicated normalization methods were each applied to the 40 and 150μg MS datasets from normal lung endothelial cell plasma membranes (ECPM). Mean and 95% CI for (**a**) raw SI dataset and datasets normalized by: (**b**) the dilution factor, (**c**) $SI_N$, (**d**) Rsc, (**e**) NSAF, were plotted using the mean diamonds and comparison circles. The x-axis represents the 2 different protein loads and the y-axis represents the log of the normalized abundance feature (number of common proteins, n = 2660). (**f**) T-ratios for statistical testing of $SI_N$ & Rsc are plotted as a function of peptide cut-off numbers (number of peptides/protein commonly identified between the samples). α=0.05 significance line is plotted, T-ratios above this line indicate that samples are different. (**g**) The ng converted $SI_N$ values (based on initial sample load) for 2,660 proteins common between the 40 and 150ug datasets were plotted against each other. The slope of the line is 3.72, $R^2$ = 0.94. (**h**) Two-way clustering

of ∼3,000 proteins identified in ECPM heart and kidney samples. Each column in the matrix represents a single 2D-LC-MS/MS run for either heart or kidney, based on the $SI_N$ normalized MS data. Proteins (rows) and tissues (columns) are clustered based on their similarities in protein intensity profile. Colors within the heatmap range from light blue (least prevalent) to dark red (most prevalent), illustrating the relative abundance of each protein within a particular sample.
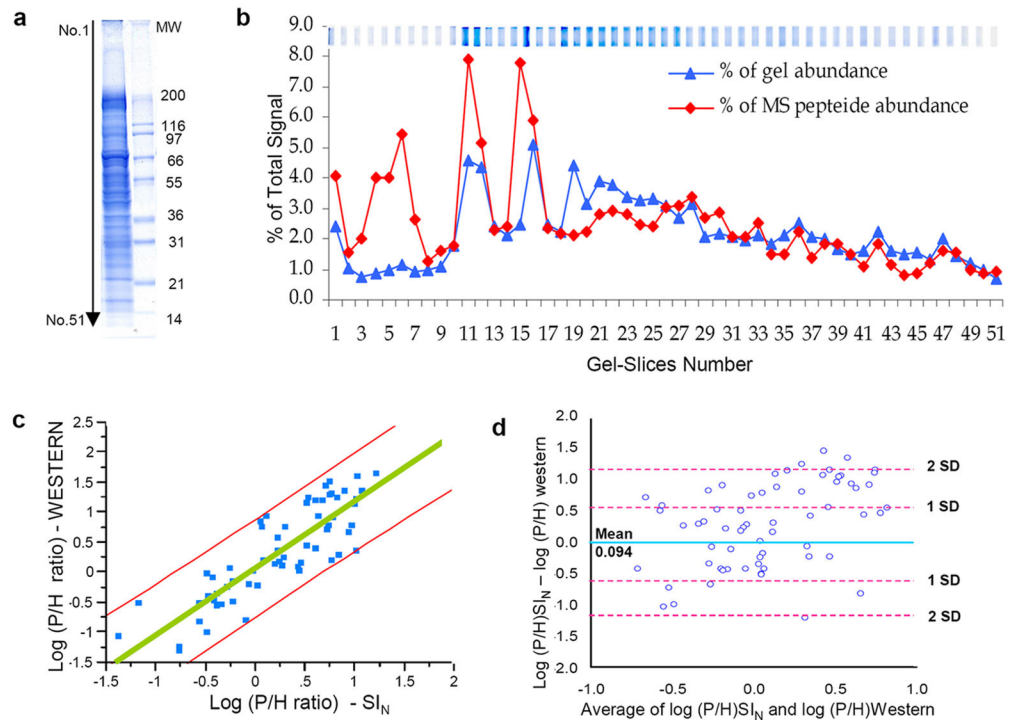
**Figure 4. Comparative analysis of proteins quantified by SDS-PAGE and MS analysis**
(**a**) Proteins in ECPM from rat lung were separated by SDS-PAGE, stained with coomassie blue and cut into 51 slices. Each gel slice was subjected to densitometry and MS analysis. (**b**) The densitometry intensities for each slice were compared to the SI on the same axis, with the x-axis being the gel slice number. (**c**) 64 proteins found in both lung ECPM (P) and the entire lung homogenate (H) were analyzed by Western blotting to quantify protein signal by densitometry. The P/H ratio for each protein from the Western analysis is plotted against its P/H ratio from the $SI_N$ values (multiple measurements). Spearman's Rho correlation between Western and $SI_N$ ratio is $\rho = 0.86$, all the points fall within 95% CI (red line). (**d**) The Bland-Altman plot for the two methods with 1 and 2 s.d. of the mean.