Published in final edited form as:

Proteomics. 2011 December ; 11(24): 4736–4741. doi:10.1002/pmic.201100078.

A Statistical Selection Strategy for Normalization Procedures in LC-MS Proteomics Experiments through Dataset Dependent Ranking of Normalization Scaling Factors

Bobbie-Jo M. Webb-Robertson, Melissa M. Matzke, Jon M. Jacobs, Joel G. Pounds, and Katrina M. Waters

Pacific Northwest National Laboratory

Abstract

Quantification of LC-MS peak intensities assigned during peptide identification in a typical comparative proteomics experiment will deviate from run-to-run of the instrument due to both technical and biological variation. Thus, normalization of peak intensities across a LC-MS proteomics dataset is a fundamental step in pre-processing. However, the downstream analysis of LC-MS proteomics data can be dramatically affected by the normalization method selected. Current normalization procedures for LC-MS proteomics data are presented in the context of normalization values derived from subsets of the full collection of identified peptides. The distribution of these normalization values is unknown *a priori*. If they are not independent from the biological factors associated with the experiment the normalization process can introduce bias into the data, possibly affecting downstream statistical biomarker discovery. We present a novel approach to evaluate normalization strategies, which includes the peptide selection component associated with the derivation of normalization values. Our approach evaluates the effect of normalization on the between-group variance structure in order to identify the most appropriate normalization methods that improve the structure of the data without introducing bias into the normalized peak intensities.

Keywords

Bias; normalization; peptide filtering; statistical models; shotgun proteomics

Normalization of peak intensity values (i.e., abundance) for proteolytic peptides obtained from an LC-MS based platform is essential to proteomic data analysis to remove systematic noise caused by sample processing. Due largely to the issues associated with peptide identification, as well as potential bias introduced by biological factors [1, 2], normalization methods for proteomics data have focused on the derivation of normalization factors from a subset of the full collection of identified peptides [3–6]. It is unknown *a priori* the random nature of peptide subsets, their relationship to underlying biological factors, and ultimately the potential impact on the transformed data. To date, the selection procedure for which peptides are used to derive normalization factors (e.g., median) have been threshold based and user-defined [3, 5, 6]. However, the selection of appropriate components of normalization (peptide selection and normalization metrics) is critical to downstream statistical analyses and should be selected with care.

An unfortunate reality of high-throughput data normalization is that there is no one size fits all solution; an appropriate normalization strategy for one dataset may introduce bias into a different dataset [2, 7, 8]. In this vein we do not introduce a novel normalization method, but a systematic approach to evaluating a series of normalization strategies in a dataset dependent manner, which has not been presented previously. For simplicity we refer to the methodology as a Statistical Procedure for the Analyses of peptide abundance Normalization Strategies (SPANS). The SPANS protocol was implemented in Java and all statistical methods were performed using MatLab® 2010a; a GUI interface for the use of SPANS is available at <https://biopilot.org/docs/Software/SPANS>, result presented here are based on the defined default settings.

The normalization of proteomics data requires two components that are often not distinguished, which we identify as a normalization strategy: (1) the peptide selection approach, and (2) the mathematical function used to derive normalization values (i.e., normalization method). For comparative purposes, we identified three peptide selection methods currently used [3–6] and introduce a fourth. These are summarized as peptides Common Within a Sample (CWS), meaning that the peptide is observed in all technical replicates of a biological sample [3], peptide intensities of the top L Order Statistics (LOS) [6], peptides with a Percentage of the Peptides Present (PPP) [4, 5], and a new Rank Invariant Peptide (RIP) selection methodology. Briefly, RIP uses peptides with complete data subjected to a Kruskal-Wallis test (non-parametric one-way ANOVA): peptides that have a p-value greater than a defined threshold are retained as RIP peptides. Overall, we utilize eight combinations of these peptide selection routines for demonstration;

- 1 NONE (all peptides);
- 2 CWS;
- 3–4 LOS with L set to ~5%, and ~20%, respectively, of the total number of peptides;
- 5 PPP at 50%;
- 6 COM (complete peptides), which is PPP at 100% or equally RIP at p-value of 0, and
- 7–8 RIP at p-value thresholds of 0.1, and 0.25, respectively.

Normalization methods are generally defined by the metrics used for location and/or scale shifts. We focus on five common normalization approaches; however, the methodology could be simply extended to include any number of normalization methods. Normalization methods considered are (1) mean centering, (2) median centering, (3) z-score transformation, (4) median-absolute deviation (MAD) scaling, and (5) regression.

Normalization of proteomics data should not introduce bias into the dataset that will affect downstream statistical analyses. A common strategy that has been used to compare normalization approaches is the pooled estimate of variance (PEV) within replicates [3, 5], however; the PEV only determines if replicates are better aligned and cannot determine if bias has been introduced. We are interested in evaluating the change in between-group variation (or ranking of the measured intensity values), which is assessed with respect to peptides that are identified with high confidence as either significant or non-significant in a quantitative manner based on the log transformed data. Indeed peptides can have a qualitative difference based on the over-expression of peptides in one biological group versus another, however; these qualitative changes will not be affected by normalization since they are based on counts [10]. Thus peptide significance for the purpose of normalization is in reference to a quantitative change of observed values. In particular, it is expected that non-significant peptides should have a randomly dispersed rank order of

intensity values across biological groups, which results in a large p-value. Similarly, peptides that are highly significant before normalization should remain so after normalization; as within-group variation decreases and the between group variation increases, the p-values will become even smaller. Since ANOVA is based on rigorous data distributions assumptions (e.g, constant variance), which often does not hold for small sample sizes, p-values are computed using a Kruskal-Wallis test. All combinations of multiple factors are input as individual groups since Kruskal-Wallis is a single factor test. The base level non-significant (NS) and significant (HS) peptides are set to be all peptides that have a p-value greater than 0.5 or less than 0.001, respectively. As our indicator of bias for peptide selection routine j and normalization method k , we compute p-values using Kruskal-Wallis on the normalized data $KW_{jk}=[p_1^{jk}, p_2^{jk}, \dots, p_p^{jk}]$ and compute the median p-value for the non-significant and significant sets respectively as,

$$NSmPV_{jk} = \text{median}_{i \in NS} (KW_i^{jk}) \text{ and, } \quad (1)$$

$$HSmPV_{jk} = \text{median}_{i \in HS} (KW_i^{jk}). \quad (2)$$

SPANS was applied to two independent LC-MS proteomics experiments; both datasets were analyzed using a LTQ-OrbitrapTM mass spectrometer (Thermo Electron Corp., Waltham, MA) with an in-house nanoelectrospray ionization LC component. Spectra were analyzed using the accurate mass and elution time (AMT) tag approach [9]. Details of the instrumentation and mass spectra data processing are given in the Supplementary material. The first dataset obtained from C57BL/6 mouse lung tissue is a 2-factor experiment; factor one included sham controls (SC) and exposure to lipopolysaccharide (LPS) and factor two included normal weight (NW) and diet induced obesity (OB). After quality control and minimum occurrence filters [10], the dataset included 61 instrument runs associated with 32 mice and 7084 peptides identified and quantified by the AMT tag process. The second dataset is associated with a comparative experiment to study the difference in peptide profiles associated with obesity and smoking in human plasma. There were a total of 28 individual plasma samples run in duplicate and after quality control and minimum occurrence filters, the dataset included 56 runs and 4687 peptides identified and quantified by the AMT tag process.

The overall process of SPANS is outlined in Figure 1. Step 0 generates a random distribution of $NSmPV$ and $HSmPV$ values to evaluate each normalization approach against. The N (number of LC-MS runs) by P (number of peptides) data matrix is normalized using a random selected normalization method where the location/scale parameters are derived from a randomly selected (without replacement) subset of that data of total size ($A < N \times P$). The normalized data resulting from this random peptide subset is then evaluated by $NSmPV$ and $HSmPV$ (Eq. 1–2). This process is repeated for I ($I=1000$), yielding a vector of 1000 $NSmPV_{Rand}$ and $HSmPV_{Rand}$ values. The Log10 and $-\text{Log10}$ transforms of these data are fit to an empirical cumulative distribution function, denoted as $F_{NSRand}(x)$ and $F_{HSRand}(x)$. These distributions are displayed in Supplemental Figures S1 and S2 for the mouse and human datasets respectively.

Step 1 of SPANS (Figure 1) evaluates the normalization values associated with each peptide selection routine for statistical relationships to unique biological groups. As seen in Figure 1, each normalization factor is defined as a vectors of the N location and scale parameters; $L = [l_1, l_2, \dots, l_N]$ and $S = [s_1, s_2, \dots, s_N]$, which is dependent upon the peptide subset. Each vector of normalization values (e.g., the median) is assessed using a Kruskal-Wallis test yielding

$p(L)$ and $p(S)$, where the p-value tests the null hypothesis that the computed normalization values have equal rank across biological groups. For the mouse and human datasets, of the 40 normalization strategies evaluated, there were 21 and 3 which had either a location or scale p-value less than 0.05. If both $p(L)$ and $p(S)$ are greater than 0.05 then the normalization approach is evaluated using Step 2, otherwise the normalization and peptide selection routine is iterated so as to not consider a normalization strategy with a high likelihood of introducing bias. Supplemental Tables S1 and S2 give $p(L)$ and $p(S)$ and the associated decision.

Step 2 computes $\log_{10}(NSmPV)$ and $-\log_{10}(HSmPV)$, for each evaluated normalization strategy. The statistical significance of the evaluated normalization approach of peptide selection j and normalization k is computed as $F_{NSRand}(\log_{10}(NSmPV_{km})) = P(X \leq \log_{10}(NSmPV_{km}))$ and $F_{HSRand}(-\log_{10}(HSmPV_{km})) = P(X \leq -\log_{10}(HSmPV_{km}))$, where a large value for either metric would indicate that the majority of the empirical CDF falls below the measured value. The overall score is then the average of these two probabilities;

$$SPANScore_{km} = 1/2(F_{NSRand}(\log_{10}(NSmPV_{km})) + F_{HSRand}(-\log_{10}(HSmPV_{km}))). \quad (3)$$

All of these values are captured in tables S1 and S2. Figure 2A and 2B gives a two-dimensional scatter plot of the comparison of $F_{NSRand}(\log_{10}(NSmPV_{km}))$ versus $F_{HSRand}(-\log_{10}(HSmPV_{km}))$ and Figures 2C–2D are the associated heatmaps of $SPANScore$. The “X” notation is normalization strategies that would have been eliminated in Step 1. In comparing Figures 2C–D the most appropriate normalization strategies are highly dataset dependent. Additionally, it is observed that simply because a normalization method is not eliminated in Step 1 it can have severe a tradeoff in respect to $NSmPV$ and $HSmPV$. Overall it is observed that non-candidates identified from Step 1 have a significantly smaller $SPANScore$ than the rest, 0.387 ± 0.276 versus 0.615 ± 0.318 and 0.171 ± 0.096 versus 0.458 ± 0.306 , respectively for the mouse lung and human plasma studies.

To further illustrate issues with improper normalization, Figure 3 gives an example of the bias introduced into a dataset when normalization factors are driven by variance introduced by biological factors and not simply by extraneous sample preparation and instrument-based variability. In this example the log transformed peptide intensities from the mouse lung experiment is given in the top two panels (Figure 3A–B). There is no statistical difference between treatment groups based on this pre-normalization level information. Figure 3C is the boxplot of this peptide after median centering using PPP-50, which was found to have a significant relationship to one or more of the biological factors (Figure 2B). The y-axis is on the same scale and thus it is evident that the within-group variation has improved between biological replicates, however; the between group variance for the LPS group has now deviated after normalization to express a significant difference by LPS exposure, which is not supported by the pre-normalization data. For comparison purposes, RIP-0.25 did not have a significant difference in the median by biological factor, which is represented in the boxplot (Figure 3D). Again, the within-group variance is reduced but not at the expense of the between-group variation. In this case the bias, as measured by $NSmPV$ and $HSmPV$, is minimized by the RIP peptide selection approach (Figure 2A). However; the $LOS20$ approach for mean and median centering would also be appropriate and the other strategies are clearly not. Because the RIP subset is selected based on a statistical test against experimental factors, they typically are appropriate for normalization, but there is no guarantee they will always maximize the improvement in variance structure for $NSmPV$ and $HSmPV$ (Figure 2).

Normalization in proteomics has largely focused on the minimization of variance between technical replicates or shifting the center of the data distributions into equality. Here it has been demonstrated that often this assumption does not hold. This finding is not surprising because the MS-based proteomic process identifies high-abundant proteins with high confidence and therefore experiments that introduce a strong protein response in the subjects, such as in the demonstrated mouse experiment, can have dramatically different peptide distributions based on biological variability. Thus, the selection of the subset of peptides used to derive appropriate normalization factors is critical. Since proteomic datasets can themselves appear quite different based on multiple factors (e.g., chromatography, MS instrumentation, peptide identification algorithm), a single normalization strategy, which includes both the normalization transformation as well as the peptides used to compute the normalization factors, cannot be defined to fit most purposes and thus we present a systematic approach to the evaluate individual datasets to minimize potential bias.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institutes of Health under grants 1R01GM084892 (BJWR), U54-016015 (JGP), U54-AI081680 (KMW) and contract HHSN272200800060C (KMW). Proteomics data was processed by the Instrument Development Laboratory at the Environmental Molecular Sciences Laboratory (EMSL). EMSL is a national scientific user facility supported by the Department of Energy. Pacific Northwest National Laboratory is a multi-program laboratory operated by Battelle for the U.S. Department of Energy under Contract DE-AC05-76RL01830.

Abbreviations used in Manuscript

RIP	Rank Invariant Peptide
SPANS	Statistical Procedure for the Analyses of LC-MS proteomics Normalization Strategies
CWS	Common Within a Sample
LOS	<i>L</i> Order Statistics
PPP	Percentage of the Peptides Present
MAD	median-absolute deviation
PEV	pooled estimate of variance
AMT	accurate mass and elution time
SC	sham controls
LPS	lipopolysaccharide
NW	normal weight
OB	obesity
BMI	Body mass index

References

1. Hill AA, Brown EL, Whitley MZ, Tucker-Kellogg G, et al. Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls. *Genome Biol.* 2001; 2:RESEARCH0055. [PubMed: 11790258]

2. Quackenbush J. Microarray data normalization and transformation. *Nat Genet.* 2002; 32(Suppl): 496–501. [PubMed: 12454644]
3. Callister SJ, Barry RC, Adkins JN, Johnson ET, 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]
4. Karpievitch YV, Taverner T, Adkins JN, Callister SJ, et al. Normalization of peak intensities in bottom-up MS-based proteomics using singular value decomposition. *Bioinformatics.* 2009; 25:2573–2580. [PubMed: 19602524]
5. Kultima K, Nilsson A, Scholz B, Rossbach UL, et al. Development and evaluation of normalization methods for label-free relative quantification of endogenous peptides. *Mol Cell Proteomics.* 2009; 8:2285–2295. [PubMed: 19596695]
6. Wang P, Tang H, Zhang H, Whiteaker J, et al. Normalization regarding non-random missing values in high-throughput mass spectrometry data. *Pac Symp Biocomput.* 2006:315–326. [PubMed: 17094249]
7. Rocke DM, Ideker T, Troyanskaya O, Quackenbush EJ, Dopazo J. Papers on normalization, variable selection, classification or clustering of microarray data. *Bioinformatics.* 2009; 25:701–702.
8. Oberg AL, Vitek O. Statistical design of quantitative mass spectrometry-based proteomic experiments. *J Proteome Res.* 2009; 8:2144–2156. [PubMed: 1922236]
9. Smith RD, Anderson GA, Lipton MS, Pasa-Tolic L, et al. An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics.* 2002; 2:513–523. [PubMed: 11987125]
10. Webb-Robertson BJ, McCue LA, Waters KM, Matzke MM, et al. Combined Statistical Analyses of Peptide Intensities and Peptide Occurrences Improves Identification of Significant Peptides from MS-Based Proteomics Data. *J Proteome Res.* 2010

Figure 1.
Flowchart of the SPANS process strategy.

Figure 2.

Graphics of normalization selection process. (A–B) Scatter plot of F_{NSRand} ($\log_{10}(NSmPV_{km})$) versus F_{HSRand} ($-\log_{10}(HSmPV_{km})$) - observations in top right indicate improvement in p-value structure. (C–D) Heatmap values associated with the $SPANScore$ (Eq. 3). “X” marks non-candidate normalizations based on statistical relationship to a biological group, the green circles indicate the normalization with the maximum $SPANScore$ and the open green circles indicate normalizations with average probability greater than 0.95.

Figure 3.

Example of the impact of normalization factors on individual peptide evidence. Peptide (VDPVNFK) shows no evidence of differential abundance for either BMI or LPS; (A) the individual $\log_{10}$ peptide intensity with no normalization and (B) the corresponding boxplot. (C) Normalization, such as PPP(50%), causes the LPS group to shift out of alignment and (D) RIP(0.25) peptide selection routine does maintains appropriate alignment of the data.