Total variance should drive data handling strategies in third generation proteomic studies

VIEWPOINT

Total variance should drive data handling strategies in third generation proteomic studies

Abigail G. Herrmann[1], James L. Searcy[1]*, Thierry Le Bihan[2,3], James McCulloch[1] and Ruth F. Deighton[1]*

[1] Centre for Cognitive and Neural Systems, University of Edinburgh, Edinburgh, UK
[2] SynthSys, University of Edinburgh, Edinburgh, UK
[3] Institute of Structural and Molecular Biology, University of Edinburgh, Edinburgh, UK

Received: February 6, 2013
Revised: August 2, 2013
Accepted: August 21, 2013

Quantitative proteomics is entering its "third generation," where intricate experimental designs aim to increase the spatial and temporal resolution of protein changes. This paper re-analyses multiple internally consistent proteomic datasets generated from whole cell homogenates and fractionated brain tissue samples providing a unique opportunity to explore the different factors influencing experimental outcomes. The results clearly indicate that improvements in data handling are required to compensate for the increased mean CV associated with complex study design and intricate upstream tissue processing. Furthermore, applying arbitrary inclusion thresholds such as fold change in protein abundance between groups can lead to unnecessary exclusion of important and biologically relevant data.

Keywords:
Bioinformatics / Differential protein expression / LC-MS/MS / Protein marker discovery / Proteome analysis

Proteomics is entering its "third generation," where MS is increasingly being used, not only to quantify total protein levels, but also to investigate how proteins within specific cell types and subcellular organelles respond both spatially and temporally to a host of experimental stimuli. [1]. As proteomic studies embark on more intricate designs, it is essential to re-evaluate whether the currently used data-handling strategies remain appropriate. Fundamental weaknesses and arbitrary design decisions still permeate proteomic research, despite efforts to improve the rigor of data handling [2–7]. This article compares primary datasets generated contemporaneously in our laboratory using peak intensity based LC-MS to provide a novel perspective on the suitability of various inclusion criteria and data-handling strategies in analyzing third generation proteomic data.

Correspondence: Abigail Herrmann, University of Edinburgh, 1 George Square, Edinburgh, EH8 9JZ, UK
E-mail: A.G.Herrmann@sms.ed.ac.uk
Fax: +44-131-651-1835

Abbreviations: GRP78, glucose regulated protein 78; **GRP94**, glucose regulated protein 94; **OGD**, oxygen-glucose deprivation

Many quantitative LC-MS proteomic studies use an initial inclusion criterion that proteins should be identified with two or more peptides. Though seemingly arbitrary, this inclusion criterion is important for two reasons: first, removal of proteins identified with only one peptide increases the reliability of LC-MS protein identification and helps avoid false detections. A single peptide feature may be found in several proteins or protein isoforms, therefore a truly definitive identification is less likely [8]. Second, this cut-off of two peptides for identification purposes significantly reduces the overall variance within the dataset, defined as the mean of the coefficient of variances for all proteins in the dataset. This reduction is variance considerably increases the power to detect subtle protein changes (Fig. 1). There is clearly a trade-off between reducing variance and the number of proteins remaining for analysis. Extending the inclusion criterion to identification of proteins with three or more peptides further reduces variance, however, also drastically reduces the number of proteins by nearly half of those originally identified by one peptide.

*These authors contributed equally to this work.

VIEWPOINT

Correspondence concerning this and other Viewpoint articles can be accessed on the journals' home page at: http://viewpoint.proteomics-journal.de

Correspondence for posting on these pages is welcome and can also be submitted at this site.

Subcellular proteomics will be a dominant theme in third generation proteomic research, yet sample fractionation can greatly impact variance within protein datasets. Sample processing techniques including the enrichment of microvessels, mitochondria [9] or white matter [10] can be used upstream of proteomic analysis to provide a more in-depth proteomic profile of how individual cell types and subcellular compartments are responding to experimental stimuli. However, increasing technicality upstream of protein detection increases the total variance of the final dataset, as demonstrated by analysis of our own proteomic data generated using a range of enrichment techniques (Fig. 2). White matter enrichment via micropunches of the corpus callosum and microvessel enrichment using density gradient centrifugation, two intricate upstream tissue handling techniques, induce a 7 and 15% increase in total variance in control tissue, respectively, compared to whole brain homogenates. We hypothesise that this increase in variance might be linked to varying degrees of protein degradation occurring when samples are handled at room temperature for extended periods of time. Upstream tissue processing enriches samples with targeted proteins, improving the spatial resolution of detected protein changes. However, the associated increases in variance make detection of subtle protein changes more difficult.

The magnitude of the change in protein abundance (fold change) is a popular but arbitrary inclusion criterion often used to dissect proteomic data. Analysis of our in vitro human cell line data shows that employing an arbitrary fold change value as a data dissection tool can exclude important proteins from the final analysis. This in vitro study investigated the effects of a global metabolic challenge on mitochondrial function and cellular proteomics. A total of 958 proteins were identified with two or more peptides ($n = 6$/group). A stringent a priori inclusion criterion of a $p < 0.01$ was set for a protein change to be deemed significant, resulting in a final protein list of 193 significantly altered proteins [11] (Fig. 3A). However, as well as a *p*-value threshold, many investigators also utilize a fold change cut-off to rapidly identify the most "important" protein changes. Datasets with a low overall variance allow for the detection of subtle protein changes, however, employing an arbitrary fold change inclusion criterion such as the popular "minimum 1.5 fold change" on these low variance datasets excludes the subtle yet significant protein changes. The fold change cut-off drastically reduces the number of proteins included in the final analysis and increases the risk of creating false negatives (Fig. 3A).

A similar analysis of the impact of arbitrary fold change cut-offs was carried out on the more variable microvessel extraction data (Fig. 3B). Due to the increased variability of these data (as shown in Fig. 2C), employing a stringent alpha value of $p < 0.01$ significantly reduces the number of proteins in the final list for analysis from 653 identified with two or more peptides to only 12. In this more variable system, imposing a 1.5 fold change cut-off has no further effect on protein number, due to a large fold change required to overcome the variance for inclusion at the set alpha level. It is therefore concluded that inclusion of a fold change data cut-off is either dangerous in the creation of false negatives (in studies with low overall variance) or irrelevant (in studies with high overall variance).

Alternatively, power calculations can be used to determine the magnitude of change required to detect a significant

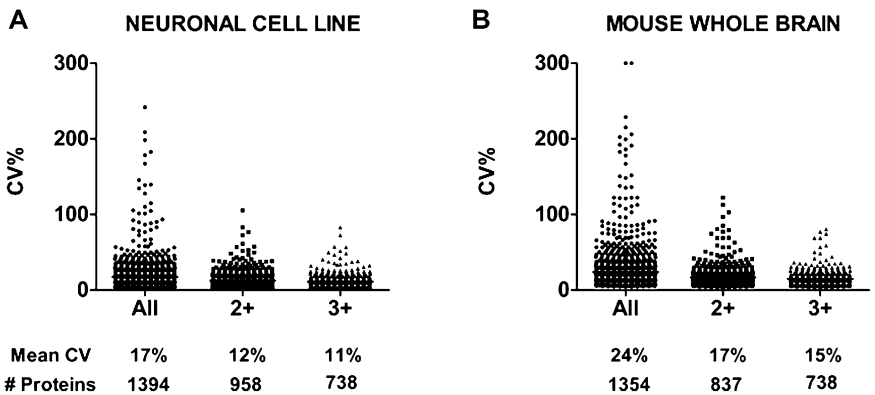

Figure 1. Variance structure is affected by the number of peptides used for protein identification. (A) Quantifying proteins with at least two peptides reduces the mean CV by 5% in a human cell line and (B) by 7% in mouse whole brain, reducing the total list of proteins by 31–38%. More stringent inclusion criteria (>3 peptides) has minimal effect on the variance but reduces the overall protein number by ~50%. A set initial inclusion criterion in proteomic data analysis should therefore be protein identification by at least two peptides. Each data point is the CV of the abundance measurements for each individual protein. The CV is calculated for individual proteins using abundance values from independent biological replicates.

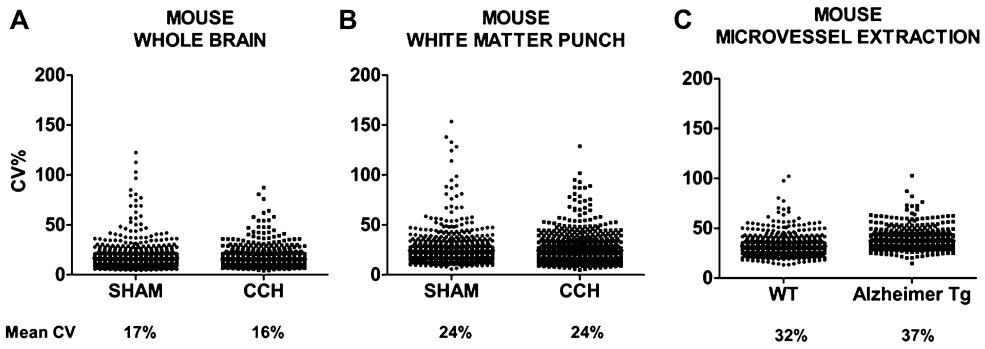

Figure 2. Complex tissue processing techniques have a direct effect on the total variance of the dataset. (A) Whole brain tissue samples show very similar overall CV in both sham and chronic cerebral hypoperfusion groups (17 and 16%, respectively). (B) White matter dissection introduces more variance into the system (24% for both sham and chronic cerebral hypoperfusion). (C) Technically demanding techniques such as microvessel dissection further increase variance (32 and 37% in wild type (WT) and transgenic (Tg) mice, respectively). Independent variables (surgery or transgene) have little effect on the variance structure of the data. The internal consistency of the three datasets controls for the technical variance introduced by the LC-MS technique, allowing the effect of tissue processing on the total variance to be assessed. Each data point represents the CV for the abundance measurement of individual proteins across independent replicates in each study, with the mean CV across all proteins shown.

difference between two populations given the technical and biological variance [5]. Used a priori, power calculations are beneficial in study design, guiding decisions regarding the number of replicates needed to obtain a set level of power [12]. However, the nature of a priori power calculations means these calculations are based on an estimate of overall biological and technical variance. Our analysis reveals that the CV is highly dependent upon the type of tissue being analyzed and the degree of upstream tissue processing involved (Fig. 2). Using a CV that is not specific to the dataset to decide detectable fold change can be problematic, and could lead to an over- or underestimation of proteins found to be differentially expressed. To ensure maximum accuracy in a priori power calculations, extensive and specific pilot data should be obtained.

The question of whether inclusion of fold change cut-offs in addition to a *p*-value cut-off adds biological value to proteomic data remains. To assess this, we identified two key proteins involved in the endoplasmic reticulum stress response: glucose regulated protein 78 (GRP78) and glucose regulated protein 94 (GRP94). In our in vitro study, experimental intervention with the metabolic challenge of oxygen-glucose deprivation (OGD) saw significant upregulation of GRP78 and GRP94 ($p < 0.01$). However, GRP78 underwent a fold change of 1.53, whereas GRP94 only had a fold change of 1.48 (Fig. 3C and D) [11]. The popular fold change cut-off of 2 would exclude both of these proteins from the analysis, and only GRP78 would be included if a fold change of 1.5 was used. The interplay between these two proteins is integral to the endoplasmic reticulum stress response; however, one or both of these proteins would be lost from the final dataset if an arbitrary fold change inclusion criterion was employed. Temporal evolution of protein level change is another important factor to be considered in understanding third generation proteomics. Data from the in vitro study demonstrate that following 6 h of OGD, small increases in protein levels of GRP78 and GRP94 predict larger increases following 18 h OGD (Fig. 3C and D). These results suggest that protein fold change should not be used as threshold for inclusion, but rather as an indicator of evolving events occurring within the cell. A protein exhibiting a small fold change at an early time point can be indicative of increasing abundance that might be detected as significant at a later time.

The ability to detect a fold change at a particular level of significance is intrinsically linked to the variance of the data, and this variance is dependent on tissue source and processing techniques (Fig. 2). It is therefore misguided to include fold change in the initial stages of data dissection. A protein reaching the threshold set by a stringent *p*-value (which in its nature incorporates the variance and the magnitude of change) should be sufficient for the initial inclusion criterion, resulting in a much reduced but relevant list of protein changes (Fig. 3).

The concept of excluding proteins based on fold change not only increases the likelihood of making type II errors, but is also fundamentally flawed given that the biological relevance of a change in protein abundance is likely to be protein specific. For example, proteins in the Bcl-2 family are important evolutionarily conserved regulators of apoptosis. However, even within this family, certain proteins are more influential than others: PUMA (p53 upregulated modulator of apoptosis) being one of the most potent [13]. Subtle changes in this protein are likely to have important cellular effects; however, may be ruled out if stringent fold change cut-offs are employed when analyzing data. The importance of subtle protein changes needs to be recognized in the analysis of large proteomic datasets to avoid the loss of valuable data through the use of inappropriate fold change cut-offs.

The issue of multiple hypothesis testing, where investigating changes in many separate proteins can lead to

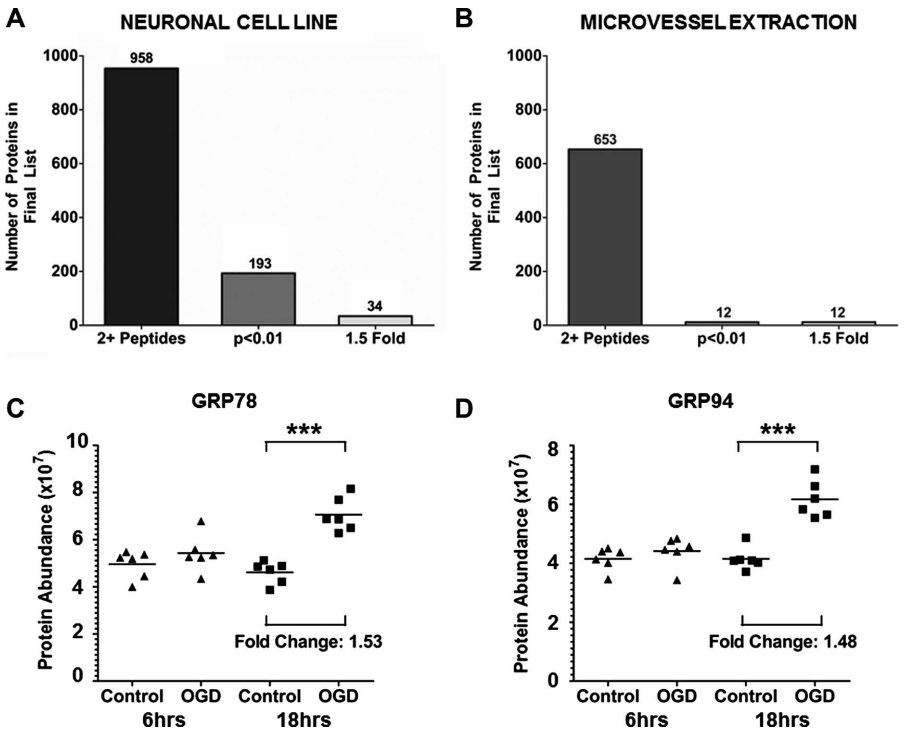

Figure 3. Arbitrary fold change cut-offs are associated with the increased likelihood false negatives. (A) Employing the common "minimum 1.5 fold change" inclusion criterion on datasets with low overall variance drastically reduces the number of proteins available for analysis. A stringent alpha value of $p < 0.01$ reduces the number of proteins in from 958 identified with two or more peptides to 193. Imposing an additional fold change cut-off of 1.5 on this reduced protein list results in a final list for analysis containing only 34 proteins. (B) Analysis of the more variable microvessel enrichment data demonstrates that a stringent alpha value of $p < 0.01$ reduces the number of proteins available for analysis from 653 identified with two or more peptides to only 12. Imposing a 1.5 fold change cut-off has no further effect on protein number, due to a large fold change required to overcome the variance for inclusion at the set alpha threshold. (C) Biologically relevant protein GRP78 is significantly increased following a severe metabolic challenge (18 h OGD), and undergoes a fold change increase of 1.53 from control to OGD samples. This protein would be included for further analysis in most proteomic studies (D) Biologically related protein GRP94 is also significantly increased following a severe metabolic challenge (18 h OGD), however, undergoes a fold change increase of 1.48. This protein would be excluded from further analysis in most proteomic studies, demonstrating the arbitrary and irrelevant nature of fold change inclusion criteria. Each data point in C and D represent an independent biological replicate ($n = 6$ for each condition).

significant results purely by chance, is an important and widely reviewed issue that is not formally dealt with in this article [4, 14–16]. However, consideration should be given to the fact that overly stringent corrections for multiple comparisons can limit the ability to glean biologically meaningful conclusions from data. Typical methods, such as the Bonferroni correction, are too stringent when studying changes in hundreds of gene or protein abundances in microarray and proteomic experiments. A less stringent method for dealing with multiple comparisons is to employ the false discovery rate, described by Benjamini and Hochberg, based on the frequency distribution of the statistically generated *p*-values [17]. It must be noted that a level of arbitrariness remains when implementing a false discovery rate. The rate of incorrectly rejecting the null hypotheses is chosen by the individual, depending on the perceived acceptability of false-positives remaining in the final dataset.

As proteomic technology advances, it is important to remember where the true power of proteomics lies: as a hypothesis generator and a tool for generating candidates of potential biomarkers and drug targets of disease. The utility of proteomics is greatest when a maximum number of proteins are identified and included for further analysis. Data processing techniques such as an initial inclusion of a protein identification threshold of two or more peptides give the researcher confidence in the protein identification. Statistical significance should then be considered as a sufficient threshold in detecting important protein changes. Pushing proteins to clear too many hurdles on their way to the final dataset increases the likelihood of omitting biologically interesting and relevant data.

AGH is supported by the MRC. This research is supported by Age UK as part of the Disconnected Mind programme, performed under the aegis of the Centre for Cognitive Aging and Cognitive Epidemiology. TLB is funded by SynthSys, a Centre for Integrative Systems Biology funded by BBSRC and EPSRC; reference BB/D019621/1. RD is funded by the Melville Trust.

The authors have declared no conflicts of interest.

References

[1] Lamond, A. I., Uhlen, M., Horning, S., Makarov, A. et al., Advancing Cell Biology Through Proteomics in Space and Time (PROSPECTS). *Mol. Cell. Proteomics* 2012, *11*, O112.017731.

[2] Podwojski, K., Stephan, C., Eisenacher, M., Important issues in planning a proteomics experiment: statistical considerations of quantitative proteomic data. *Methods Mol. Biol.* 2012, *893*, 3–21.

[3] Karp, N. A., Lilley, K. S., Design and analysis issues in quantitative proteomics studies. *Proteomics* 2007, *7*, 42–50.

[4] Diz, A. P., Carvajal-Rodriguez, A., Skibinski, D. O., Multiple hypothesis testing in proteomics: a strategy for experimental work. *Mol. Cell. Proteomics* 2011, *10*, M110 004374.

[5] Levin, Y., The role of statistical power analysis in quantitative proteomics. *Proteomics* 2011, *11*, 2565–2567.

[6] Deighton, R. F., Kerr, L. E., Short, D. M., Allerhand, M. et al., Network generation enhances interpretation of proteomic data from induced apoptosis. *Proteomics* 2010, *10*, 1307–1315.

[7] Karp, N. A., McCormick, P. S., Russell, M. R., Lilley, K. S., Experimental and statistical considerations to avoid false conclusions in proteomics studies using differential in-gel electrophoresis. *Mol. Cell. Proteomics* 2007, *6*, 1354–1364.

[8] Mallick, P., Kuster, B., Proteomics: a pragmatic perspective. *Nat. Biotechnol.* 2010, *28*, 695–709.

[9] James, R., Searcy, J. L., Le Bihan, T., Martin, S. F. et al., Proteomic analysis of mitochondria in APOE transgenic mice and in response to an ischemic challenge. *J. Cereb. Blood Flow Metab.* 2012, *32*, 164–176.

[10] Reimer, M. M., McQueen, J., Searcy, L., Scullion, G. et al., Rapid disruption of axon-glial integrity in response to mild cerebral hypoperfusion. *J. Neurosci.* 2011, *31*, 18185–18194.

[11] Herrmann, A. G., Deighton, R. F., Le Bihan, T., McCulloch, M. C. et al., Adaptive changes in the neuronal proteome: mitochondrial energy production, endoplasmic reticulum stress, and ribosomal dysfunction in the cellular response to metabolic stress. *J. Cereb. Blood Flow Metab.* 2013, *33*, 673–683.

[12] Bezeau, S., Graves, R., Statistical power and effect sizes of clinical neuropsychology research. *J. Clin. Exp. Neuropsychol.* 2001, *23*, 399–406.

[13] Yu, J., Zhang, L., PUMA, a potent killer with or without p53. *Oncogene* 2008, *27*, S71–S83.

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

[15] Almudevar, A., Multiple hypothesis testing: a methodological overview. *Methods Mol. Biol.* 2013, *972*, 37–55.

[16] Goloborodko, A. A., Mayerhofer, C., Zubarev, A. R., Tarasova, I. A. et al., Empirical approach to false discovery rate estimation in shotgun proteomics. *Rapid Commun. Mass Spectrom.* 2010, *24*, 454–462.

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

