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

Citation for published version:

Digital Object Identifier (DOI):
10.1002/pmic.201300056

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Proteomics

Publisher Rights Statement:
Available under Open Access

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Total variance should drive data handling strategies in third generation proteomic studies

Abigail G. Herrmann, James L. Searcy, Thierry Le Bihan, James McCulloch, and Ruth F. Deighton

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

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

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 in 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.
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
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.

Figure 2.
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


