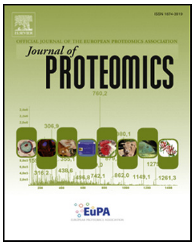
Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jprot



Application of clinical assay quality control (QC) to multivariate proteomics data: A workflow exemplified by 2-DE QC☆

David Jackson[*], *David Bramwell*

Biosignatures Ltd., Keel House, Newcastle Upon Tyne, UK

ARTICLE INFO

Article history:
Received 25 January 2013
Accepted 25 July 2013

Keywords:
Clinical proteomics
Quality control
2-DE
Control charts

ABSTRACT

Proteomics technologies can be effective for the discovery and assay of protein forms altered with disease. However, few examples of successful biomarker discovery yet exist. Critical to addressing this is the widespread implementation of appropriate QC (quality control) methodology. Such QC should combine the rigour of clinical laboratory assays with a suitable treatment of the complexity of the proteome by targeting separate assignable causes of variation. We demonstrate an approach, metric and example workflow for users to develop such targeted QC rules systematically and objectively, using a publicly available plasma DIGE data set. Hierarchical clustering analysis of standard channels is first used to discover correlated groups of features corresponding to specific assignable sources of technical variation. These effects are then quantified using a statistical distance metric, and followed on control charts. This allows measurement of process drift and the detection of runs that outlie for any given effect. A known technical issue on originally rejected gels was detected validating this approach, and relevant novel effects were also detected and classified effectively.

Our approach was effective for 2-DE QC. Whilst we demonstrated this in a retrospective DIGE experiment, the principles would apply to ongoing QC and other proteomic technologies.

Biological significance

This work asserts that properly carried out QC is essential to proteomics discovery experiments. Its significance is that it provides one possible novel framework for applying such methods, with a particular consideration of how to handle the complexity of the proteome. It not only focusses on 2DE-based methodology but also demonstrates general principles. A combination of results and discussion based upon a publicly available data set is used to illustrate the approach and allows a structured discussion of factors that experimenters may wish to bear in mind in other situations. The demonstration is on retrospective data only for reasons of scope, but the principles applied are also important for ongoing QC, and this work serves as a step towards a later demonstration of that application. This article is part of a Special Issue entitled: Standardization and Quality Control.

Abbreviations: PC, Principal Component; PCA, Principal Component Analysis; QC, quality control; VSN, variance stabilisation normalisation

☆ This article is part of a Special Issue entitled: Standardization and Quality Control.

[*] *Corresponding author.* Tel.: +44 191 6453645; fax: +44 191 2302131.
E-mail address: dave.jackson@biosignatures.com (D. Jackson).

1874-3919/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jprot.2013.07.025

1. Introduction

Proteomics has been applied in the effort to discover novel clinical biomarkers. However, initially seemingly promising high-profile proteomics studies (e.g. [1]) fell prey to concerns including but not limited to statistical, data processing and sample processing issues [2–5], leading to adverse attention directed at the field. Leading groups have been addressing this, with a range of global standardization efforts underway evinced in key publications (e.g. [6–11]), and an increased emphasis on the importance of biomarker validation and clinical translation [12–14]. Promisingly, this trend has culminated in the FDA certification of biomarkers derived from the same technique that initially caused the greatest controversy, SELDI-MS, for ovarian cancer [15–17] which has been shown to be reproducible with rigorous application [18]. This demonstrates that with (and only with) appropriate attention to critical quality factors, the field can deliver successfully validated biomarkers.

One critical factor is the application of appropriate quality control (QC) metrics to proteomics data [19]. Treating proteomics discovery technology with the same rigour as is expected of downstream clinical validation efforts, and of measurements of validated clinical biomarkers, could reduce the high number of studies that fail to validate by acting as a front-line quality filter. In clinical situations, rigorous QC is commonly applied to measurements of a single analyte. Multi-rule QC approaches applied to Levey–Jennings control charts [20], such as those commonly referred to as Westgard rules [21], can reveal deviations from the normal limits of analyser variation in a relatively simple, but sensitive plot. Key metrics are monitored over time against strict rejection criteria [21–24]. In proteomics discovery experiments on the other hand, the level of QC employed is recognised to often fall far short of such a level of rigour [19]. In mitigation, proteomics discovery technologies potentially assay many thousands of species simultaneously, so rigorous QC is not without challenge, but bridging this QC gap between well-controlled single analytes in the clinic and poorly controlled multiple analytes in discovery could be crucial for improving the translational success of proteomic platforms. This could have considerable cost, efficiency and reputational benefits for the field if successfully implemented.

We have demonstrated elsewhere the application of clinical QC approaches to proteomics data, using 2-DE as an example. This was achieved by reducing the complexity of the entire 2-DE profile to a single metric, and then plotting this metric on control charts [24]. It was shown that effective QC was possible for the detection of known outlier effects by deriving a single measure from complex data and monitoring it over time analogously to clinical analyses. That work may also be consulted for a good general introduction to statistical process control in proteomics. However for the purposes of that demonstration, such measures were relatively ‘global’ e.g. the median raw feature volume. This sufficed to demonstrate that clinical QC approaches are effective on 2-DE; it also introduced, but for reasons of scope did not examine in detail, one key point which we now address. As above, there is the challenge that proteomic profiles are complex multivariate data sets (~2000 protein forms may be analysed simultaneously by 2-DE [25]) and numerous separate assignable technical effects may arise within a given profile. Hence a given technical issue may only affect a proportion of the features. A good illustration of this profile complexity is the spatial bias effect described in [26]. This suggests that to fully take proteomic complexity into account and maximise the effectiveness of QC, the approach employed must be able to resolve different technical effects within each run, evaluate them for importance, and quantitate the important effects separately.

Here, we demonstrate an approach for systematically deriving, evaluating and applying such targeted QC rules, using a publicly available 127-gel plasma DIGE data set. Our approach to systematic rule development utilises correlation analysis of feature volumes in the standard channels to create a hierarchical clustering plot covering the entire gel profile. Correlated groups of features are revealed by cutting this tree plot into separate groups, the assumption being that the groups potentially represent specific, assignable technical effects. The next step is to quantitate the variation represented by the clusters, and this is achieved by deriving a single measure for the behaviour of each cluster in each gel. The relevant features are first plotted in PCA (Principal Component Analysis) space to reduce the dimensionality of the data. The Mahalanobis D2 distance metric [27] from the centre of the distribution of the gels is then calculated for every gel for selected PCA components explaining the majority of the variance. This single metric represents a variable that can be easily plotted on a control chart, and compared to a confidence interval cut-off, to flag up outlying gels for comparison to run order and batch metadata. The rules employed and control chart limits can be refined based upon user assessment of factors such as the nature of the cluster on the gel and magnitude of the effect. By examining rules under varying hierarchical cluster division conditions until ‘false positive’ technical effect clusters begin to appear (in this case meaning effects of low magnitude or no interest), a user can be confident that all current sources of variation important to them have been accounted for. Whilst this is done retrospectively here, plotting gels in run order can usefully reveal drift during the experiment, which is critical for understanding whether this drift is a co-ordinated effect, and the method could also be applied to ongoing studies.

As demonstrated, the method relies on the presence of an internal standard to reveal the technical component of variation unaffected by biological heterogeneity. This therefore acts as a clear arbiter for technical effects. However, we also discuss the potential for the operation of the approach in the context of both technical and biological variation.

In summary, we offer a systematic and objective approach to implement QC methods that are effective on multivariate proteomic data. These methods are illustrated with reference to specific results from an existing DIGE set, but the concepts encountered could be applied more widely.

2. Materials and methods

The data set described in [28] was used for this analysis. This consisted of 121 gels of minimal CyDye™ (GE Healthcare, Amersham, UK) fluorescently-labelled material from human plasma, with 120 of the images corresponding to 60 individuals sampled twice, a week apart, and one being a single sampling. A pooled internal standard DIGE methodology was applied with Cy5 labelling of individual samples and Cy3 labelling of the pooled standard. The data set also includes six further gels that had been rejected and repeated owing to an issue with a pre-supplied buffer (causing them to over-migrate relative to the dye front with the loss of relatively low molecular weight species), taking the total to 127 image pairs. These were included in this work to represent demonstrative pre-existing quality control failures with an assignable difference. The key separation of note here is that second dimension gel runs were divided into 14 batches (13 normal batches, after the single batch of six gels that was repeated). Further detail on the data set is available in the Supplementary Methods, including the experimental design in Supplementary Methods Figure SM1.

3. Data analysis and visualisation

3.1. *Feature detection*

Aligned images for the main 121 gels were opened using SameSpots v4.5 (Nonlinear Dynamics, Newcastle upon Tyne, UK). The same reference gel as the original study, mem001_121_cy3, was selected. No mask of disinterest was applied, as the images were pre-cropped. Spot features were auto-detected based on all the images in both channels, and the pattern then edited to remove extraneous material and correct the profile where necessary. The central albumin spots, which were allowed to saturate on the scans in the original study and then removed from the image analysis (to improve the detection of lower abundance species by boosting the signal), were also removed here. The final spot pattern contained 1004 spots. This is fewer than the original study because of the use of updated software and editing criteria. The six rejected gels were added to the analysis only after the pattern detection and editing stages, which their repeats contributed to in their place. Their later addition at this stage ensured that the anomaly present in them did not affect the base pattern, which was our intention as they represented known deviation from standard behaviour. The feature-free gel image for the reference image is shown in Supplementary Methods Figure SM2 for illustration of the profile.

3.2. *Data extraction*

Raw spot feature volumes for the standard channels were extracted from the SameSpots experiment, and $\log_{10}$-transformed for analysis throughout this work unless otherwise noted. Data were analysed without normalisation or background subtraction to ensure that no information was discarded for QC purposes; normalisation, for example the DIGE scheme in the original work, leads to the smoothing out of biases [29]. This improves data precision and reproducibility; however for characterising and monitoring system technical performance, this is less appropriate. The possibility of post-analytical bias ([30]) was also largely obviated by not applying these steps. Log_{10} transformation was, however, applied to generate a normal distribution for statistical analyses.

Spots in the master pattern from the 121 gels that were not present at the base of the six over-migrating gels, or that were partially missing, required a missing value correction for their lack of features in the over-migrated region. These features were assigned a background value for volume in each relevant gel based on a projection of 3 standard deviations of $\log_{10}$ volume below the mean $\log_{10}$ volume for all spots present in that gel. This imputation method generated a complete data set for those gels, allowing us to demonstrate our automated workflow on a strong known effect that one would expect should be detected, with a real assignable cause.

3.3. *QC analysis*

Correlation analysis and selection of hierarchical groups were carried out using the *cor*, *hclust* and *cutree* commands from the default *stats* package in R [31]. All 1004 features were analysed over all 127 standard channels. The *cor* function was used to calculate the Pearson correlation of the features to every other feature. The correlation values were then transformed so that correlation and anti-correlation were treated equally. This correlation data was then given to *hclust* so as to cluster features together by similarity of behaviour. The *cutree* command was used to split the resulting clustering tree into 10 distinct groups.

PCA plots were then prepared using *PcaClassic* from the *rcorr* package [32] for each cluster of features independently. PCs (Principal Components) 1 and 2 were plotted in all cases. Normal-probability contour ellipses shown on the plots were calculated using *dataEllipse* from the *car* package. This consistent workflow was not applicable if there was only a single feature in a given cluster. This issue was avoided upon occurrence simply by adding a 'composite' feature to generate a second dimension i.e. single-feature clusters were augmented by analysing along with the median value of all of the other features. This would achieve a consistent representation, but not mask the unusual behaviour of a feature, as this would already have been revealed by its failure to cluster with any other features.

The Mahalanobis D2 distance for the gel position in PCA space was calculated to measure gel behaviour for each cluster group in a single metric for plotting on control charts [27]. In each case this was calculated using the *mahalanobis* command from the default *stats* package, and measured only for PCs 1 and 2 as per the plots. This metric was then plotted on a control chart as adapted from [20,21]; data were plotted in run order and marked by second dimension gel batch. The gel batch was used in this example workflow as this is one of several primary sources of inter-batch variation ([28,33]) and is therefore an appropriate batching division with which to demonstrate the approach. Other batch metadata are recorded in the data set used for alignment against the run order if desired [34]. The QC rule applied for this work was analogous and similar in stringency to the 1_{2s} rule described in [21], adapted for the Mahalanobis D2 distance metric. The 1_{2s} rule (i.e. investigate any one run incidence occurring more than 2 SD from the mean) would normally be a 'warning' not rejection rule; using a rule similar to this could therefore produce more 'false positives' than would normally occur in QC, but this is appropriate to demonstrate the method and a starting point for rule refinement. In this work any

Please cite this article as: Jackson D, Bramwell D, Application of clinical assay quality control (QC) to multivariate proteomics data: A workflow exemplified by 2-DE QC, J Prot (2013), http://dx.doi.org/10.1016/j.jprot.2013.07.025

gels for which the Mahalanobis D2 distance metric in PCs 1 and 2 was greater than the 95% confidence interval of the data set were regarded as QC 'flags' for investigation. These were provisionally classified as outliers (out of control) as against gels within the limits which were regarded as in control.

4. Results

The workflow is illustrated schematically with numbered steps in Fig. 1. We now show and discuss results derived from applying this approach to the example data set. The workflow steps are more important than any recommendation of the exact methodology used and the specific rules derived; we aim to reiterate the key messages that i) assignable causes should be resolved from multivariate data for technical effects, ii) an appropriate single metric should be derived to measure these, and iii) control charts are the ideal way to monitor that metric.

- Step 1. Standard gels are analysed to derive cluster groups that are correlated in technical variance, each representing a potential QC metric (Section 4.1.1).
- Step 2. Feature volume data for those cluster groups are projected into PCA space; the number of components required to explain the majority of the variance in each cluster is determined. PCA plots may be used for initial descriptive analysis of the groups (Section 4.1.2).
- Step 3. The Mahalanobis D2 distance metric is calculated for each standard in PCA space, for the determined number of components. Control charts based on this metric are generated for initial QC outlier classification, using first-pass rule settings (Section 4.1.3).
- Steps 4 and 5. QC rules are selected/confirmed for use based on examination of the initial QC results in the standards. Rule stringency may be adjusted based upon this also (Section 4.2).
- Step 6. The process may be repeated at a different clustering distance to derive more groups if desired (discussed in Section 4.2).

Whilst the rule selection is highlighted at Steps 4 and 5, in reality a user would be examining the results at all stages to contribute to this decision. The point on Fig. 1 referring to the 'bulk' cluster is explained throughout the results below, being a large correlated group containing the majority of spots which are not specifically assignable.

4.1. *QC rule generation*

4.1.1. *Step 1. Correlation analysis and technical effect discovery*
Hierarchical correlation clustering analysis was performed and the resulting tree cut at a height to produce 10 clusters. Ten clusters was an arbitrary choice to provide an initially manageable number of cluster groups for investigation whilst still subdividing the profile sufficiently to resolve separate local technical effects (Fig. 2(a)). This number could be adjusted as part of user customisation or selected on any appropriate statistical basis; the key result is that appropriate technical effects are revealed to the user. As per Step 6, we suggest that users make an assessment as to whether 'false positive' effects not of interest have begun to be selected (suggesting that 'meaningful' effect discovery is complete). Some means by which this might be decided are discussed in Section 4.2.

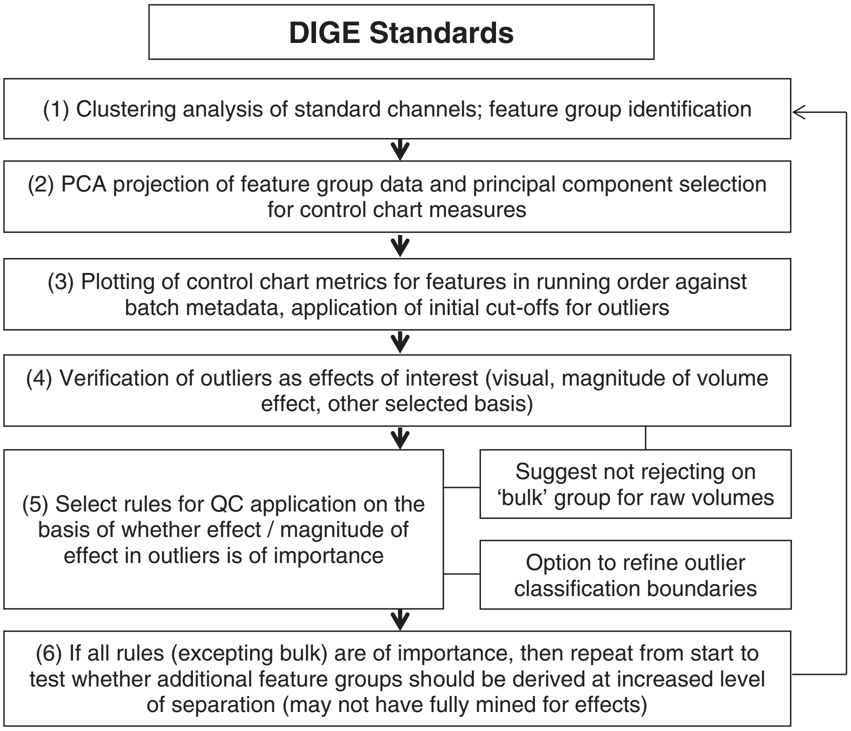

DIGE Standards
(1) Clustering analysis of standard channels; feature group identification
(2) PCA projection of feature group data and principal component selection for control chart measures
(3) Plotting of control chart metrics for features in running order against batch metadata, application of initial cut-offs for outliers
(4) Verification of outliers as effects of interest (visual, magnitude of volume effect, other selected basis)
(5) Select rules for QC application on the basis of whether effect / magnitude of effect in outliers is of importance
Suggest not rejecting on 'bulk' group for raw volumes
Option to refine outlier classification boundaries
(6) If all rules (excepting bulk) are of importance, then repeat from start to test whether additional feature groups should be derived at increased level of separation (may not have fully mined for effects)

Fig. 1 – The QC workflow for DIGE experiments as a flowchart.

The 10 clusters are shown on the protein profile in Fig. 2(b), and the number of spots in each cluster is given in Table 1. The data provided to the correlation analysis were from the standard channel only which was the same material on all gels, so any variation of note must be derived from technical effects. In a case where all features are independent and subject only to random noise variation we would not expect clear sub-trees nor spatial clustering of correlated features; hence co-localised clusters are more likely to represent correlated, concerted technical effects.

It is immediately apparent that the majority of the protein profile correlates in behaviour, in the large group comprising 775 spots, cluster 10. This region, the 'norm' or majority of the profile, will be primarily affected by gel-wide effects such as signal, scanner photomultiplier voltage, and protein loss in equilibration as we analysed raw volumes. Whilst these are indeed technical effects, it is not a cluster of technical concern specifically, as global normalisation addresses these points, and for good quality 2-DE data relatively unbiased across the profile, such a 'bulk' group would be expected when loading the same sample. Therefore we would propose that monitoring such a 'bulk' group for drift is indeed worthwhile, as systematic drift over time in the whole/most-of-gel signal would be important to track (for example, reflecting assay, equipment, or procedural drift), but not for the purpose of forming a rejection rule on the basis of it for specific gels.

Other smaller, specific clusters do display distinct behaviour, representing correlated local effects potentially of interest. Importantly these also fit with the system under study (i.e. are potentially "assignable"). Clusters 2, 3, 6, and 9 represent regions of the gel prone to interference from albumin streaking effects, both laterally and vertically, which are unsurprisingly present given that this data set comprises samples not pre-processed to remove albumin. Since these effects could occlude 'normal' protein density where present, this is a potential local QC metric for non-depleted plasma. Cluster 5 appears to represent isoelectric focussing variation with an emphasis on potential end-of-profile resolution effects (primarily at the acidic end of the gel, but also with a basic component). Clusters 4 and 8 appear to primarily denote co-ordinated variation at the top of the gel, potentially a site of greatest second dimension variability, although there are a few other specific forms that cluster with these. This could indicate second dimension effects affecting several regions of the profile simultaneously but certainly focussing on particular regions.

Cluster 7 is of key interest; this corresponds to the region affected by the known buffer issue, as will be shown below, and features in this region should have distinct highly correlated behaviour owing to their being absent in a small subset of gels. Also, one of the groups is a single feature ('cluster' 1, on the basic edge of the profile). As this spot behaves in a different manner to all others technically, this counsels caution in its quantitation (as would its location on the edge of the profile in any event). Incidentally, we also noted that this spot was relatively variable in intensity between the standard and sample channels (data not shown), possibly representing one of the rare differential labelling species that can arise in DIGE profiles.

The general trend is towards albumin-based effects and edge-of-gel effects being identified as specific small clusters. Outside of the bulk profile variation in cluster group 10 and the known buffer-repeat effects in cluster group 7, 197 (19.6% of) spots are present in any of the other clusters. 117 of these are in the clusters showing localised albumin-related effects. The remainder is largely at the edges of the gel profile.

This demonstrates well the principles behind the selection process. Clear groups of co-localised features arise from the cluster analysis. This shows that there are indeed detectable technical effects present amenable to further investigation. Had no co-localised clusters been generated, this would have suggested a lack of *assignable* specific technical variation and targeting would not be appropriate (at least at this level of

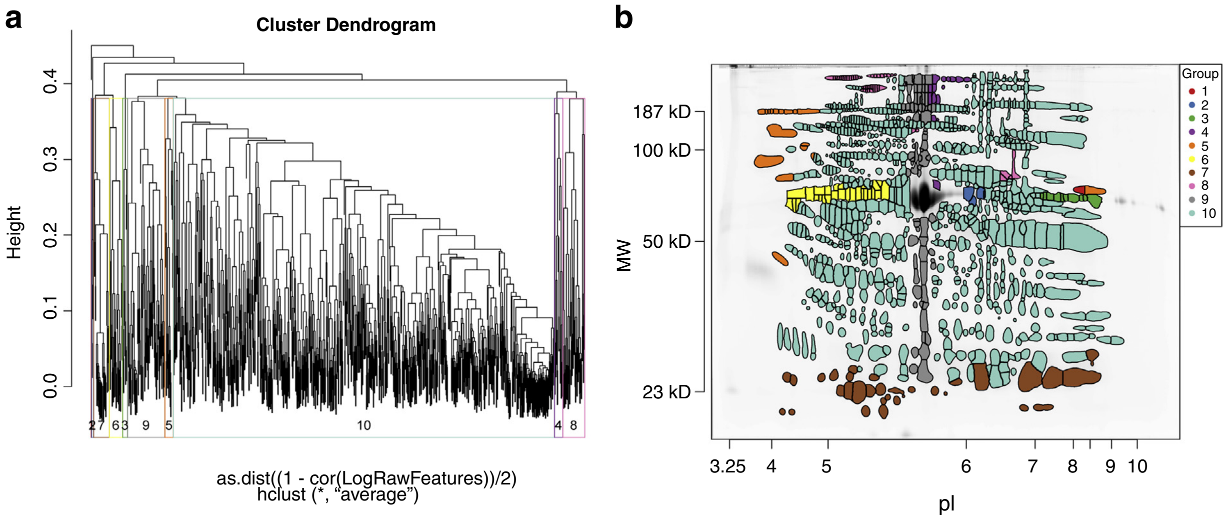


Fig. 2 – a) Cluster dendrogram representing the distance matrix of 1004 spot features as assessed by correlation over the 127 standard channels. Ten feature groups were generated, being numbered in ascending order with the number of features present in the cluster. The numbers so assigned are at the base of each group. b) The ten groups shown as coloured feature groups on the profile. Colours correspond to those in (a) and the group numbers are provided in the key.

Please cite this article as: Jackson D, Bramwell D, Application of clinical assay quality control (QC) to multivariate proteomics data: A workflow exemplified by 2-DE QC, J Prot (2013), http://dx.doi.org/10.1016/j.jprot.2013.07.025

Table 1 – Summary of information from Sections 4.1 and 4.2.

Cluster	Features	Outlying gels	Mean in [a]	Mean out [b]	Mean diff [c]	p [d]
Cluster 1	1	12	14.35138	13.45553	−0.89585	0.00033
Cluster 2	4	8	13.74491	14.67665	0.93174	0.00012
Cluster 3	10	8	12.58424	14.34819	1.76395	0.00446
Cluster 4	17	7	12.52238	12.76099	0.23861	0.16497
Cluster 5	17	5	12.81524	13.01016	0.19492	0.36978
Cluster 6	27	7	14.93233	14.92369	−0.00864	0.91447
Cluster 7	32	6	13.58477	10.31556	−3.26921	0.00004
Cluster 8	45	8	12.14622	11.84235	0.30388	−0.04845
Cluster 9	76	5	13.71401	13.70392	−0.01009	0.95054
Cluster 10	775	5	13.00420	13.10090	0.09670	0.42205
Whole Gel	1004	9	13.08108	13.02785	−0.05323	0.47127

[a] Mean $\log_{10}$ volume of the cluster features in each gel, averaged across all in-control (non-outlying) gels.
[b] Mean $\log_{10}$ volume of the cluster features in each gel, averaged across all out-of-control (outlying) gels.
[c] The difference between the two mean $\log_{10}$ values.
[d] Simple two-tailed t-test p value for the in-control versus out-of-control groups, discussed in the text.

cluster tree division). It is for a user to investigate whether such effects arise in their data set. Even if they do not, as we covered in [24], control charts can be applied to more global metrics and still detect gross issues.

Removing the six gels with the known buffer problem eliminated the relevant cluster but had little effect on the clustering results for the remainder of the profile (Supplementary Results Figure SR1); this was expected. Clustering analysis was performed on raw volume, and therefore the spots in any given cluster have no effect on the readings for other clusters, as might have occurred with the application of normalisation schemes. This demonstrates that the strong known issue and its analysis did not perturb the results presented for novel clusters. The only notable alteration was that the vertical albumin-related cluster, 9, was tipped over a splitting threshold and separated into two groups (7 and 9 in the new figure), but still arose as an effect and would be detected.

In summary, whilst there is a large proportion of the profile (spatially distributed) in which features behave similarly, particular localised technical effects can be identified that do not. These correspond to technically sensible, assignable clusters for potential QC monitoring and investigation. It remained to be determined at this stage whether any or all were issues of importance (Section 4.2) after the next step; measurement of outliers for these effects.

4.1.2. *Step 2. PCA plots for visualisation*

Plotting the clusters individually using PCA allowed a visualisation of outlying gels for each. Because of the selection of groups of features already behaving in a correlated manner in each cluster, few principal components were required to visualise the majority of the variance in each resulting data set. A mean of 63.5% of the variance was explained by the first component and 76.9% by the first two components across the clusters (Supplementary Results Table SR1). This incorporated the 'bulk group', cluster 10, for which only 44.2% of the variance was explained by the first two components. This is subject to the influence of the majority of the profile and not a targeted cluster for rule generation, hence closer to the whole-gel value of 44.5% for the first two components. Excluding this cluster, the other more targeted clusters explain on average 66.6% of the variance by the first principal component and 80.6% by the first two, with a minimum explained by PCs 1 and 2 that is as high as 63.4% for cluster 8. Therefore, the first two components were deemed sufficient for all PCA plots in this study, always explaining the majority of the variance in relevant clusters whilst also allowing visualisation in a simple two-dimensional plot. This reduction of dimensionality makes PCA effective for exploratory analysis of the data and especially so in this case where the data are already correlated. Of course, this decision represents a customisable parameter and users may wish to examine the number of components required to explain a particular proportion of the variance, for example. The numbers of components for each cluster that cumulatively explain no greater than 90% and 99% of the variance in this data set are shown illustratively in Supplementary Results Table SR1.

It should be noted that for cluster 1, the single feature, this metric was adapted as described for single features in the QC Analysis section (Section 3.3). This was intended to allow the same approach to be applied and stabilise the value relative to the bulk data, but a single feature is a special case and this is one of several ways to handle such an occurrence. This cluster only had two components by definition so they explained 100% of the variance for this cluster. This did not skew the conclusions in the preceding paragraph given the high *minimum* value explained by two components over all the relevant clusters, and only a very slight boosting effect on the mean by inclusion of this cluster; the mean variance explained by PCs 1 and 2 for clusters 2–9 was 78.1% versus the 80.6% quoted above for 1–9.

An example PCA plot is shown in Fig. 3(a) for cluster 2, which represents one of the 'novel' clusters revealed by the hierarchical clustering analysis near the main central forms of albumin; and in Fig. 4(a) for the 'known' buffer-induced gel-base issue, cluster 7. Normal-probability contour ellipses in PCs 1 and 2 are also shown here, for comparison with outliers selected in the control charts that follow in the next section, where we cover part (b) of each figure. In cluster 2, eight gels outlie beyond the 95% bound, scattered throughout the entire data set in run-order terms. In cluster 7, the six gels known to be affected are clear outliers. All but one of the remaining cluster PCA plots

Please cite this article as: Jackson D, Bramwell D, Application of clinical assay quality control (QC) to multivariate proteomics data: A workflow exemplified by 2-DE QC, J Prot (2013), http://dx.doi.org/10.1016/j.jprot.2013.07.025

are provided in the Supplementary Results (Supplementary Results SR2 to SR8, part (a)). Cluster 3 is discussed in Section 4.2 below, and its PCA is shown on Fig. 6(a) there.

For comparison to the individual plots, the equivalent for the entire gel profile is shown in Fig. 5(a); the variance associated with each component for the whole profile is also shown in Supplementary Results Table SR1, as above. As expected, outlying gels differ from those in most of the cluster sub-plots and a greater number of components are required to explain the same proportion of the variance relative to the targeted clusters, reflecting the lack of focus on specific sources of co-ordinated variation.

These plots can visually demonstrate which gels outlie. For a metadata-aligned visualisation showing drift over time, a single metric and simpler plot is needed.

4.1.3. *Step 3. Control charts*

The PCA plot is convenient whilst exploring the data initially but provides little scope for ongoing QC. We showed in [24] that control charts are an ideal representation for exploring technical issues. The key factor in a PCA plot is where a given point resides with respect to the data bounds. This suggests that a metric that effectively gives the probability of a given point (i.e. its distance from the centre of the PCA ellipse transformed into a probability) would be a useful metric to track on a control chart. Such a transform is achieved by calculating the squared Mahalanobis distance (Mahalanobis D2) for each of the points. This enables one to use the power and rules of the control chart for arbitrary subsets of features.

The Mahalanobis D2 distance for each gel in PCA space for PCs 1 and 2 was plotted on control charts for each feature cluster across all the 127 standards. The cut-off bounds in these plots correspond to the outer ellipses in the PCA plots; that is, the 95% confidence interval for the distance metric. Hence, by design, outliers were the same.

The results are shown in Figs. 3(b) and 4(b) for clusters 2 and 7. The eight cluster 2 outliers do not cluster on the chart by batch or run; their scatter is not even, with some possible loose proximity of pairs in runs close together, but there is no batch specificity, or consistency in the value of gels adjacent to an outlier. This is important to note as it indicates that one gel cannot accurately act as a reporter for adjacent gels for this effect, or for the entire second dimension batch; it is semi-random with respect to batch and gel QC must be on an individual basis. For cluster 7 in Fig. 4(b), the outliers seen in Fig. 4(a) are all within a batch based on a known single-batch effect. These two examples demonstrate the ability of control charts to reveal both sporadic and batch-associated variation.

The control chart format is also excellent for revealing subtle periodic trends over time, where incidence per batch is low (unlike cluster 7) but there is a clear temporal clustering effect. Cluster 3 (related to albumin effects on IEF) is a good example of this (Fig. 6). Fig. 6(a) shows the PCA plot for this cluster; Fig. 6(b) shows the control chart, on which it is clear that the effect is much more prevalent in later batches, without dominating any single batch. This trend may be an IPG-strip or other IEF-related variation effect as opposed to gel-related given the lateral direction of the effect on the gel. It

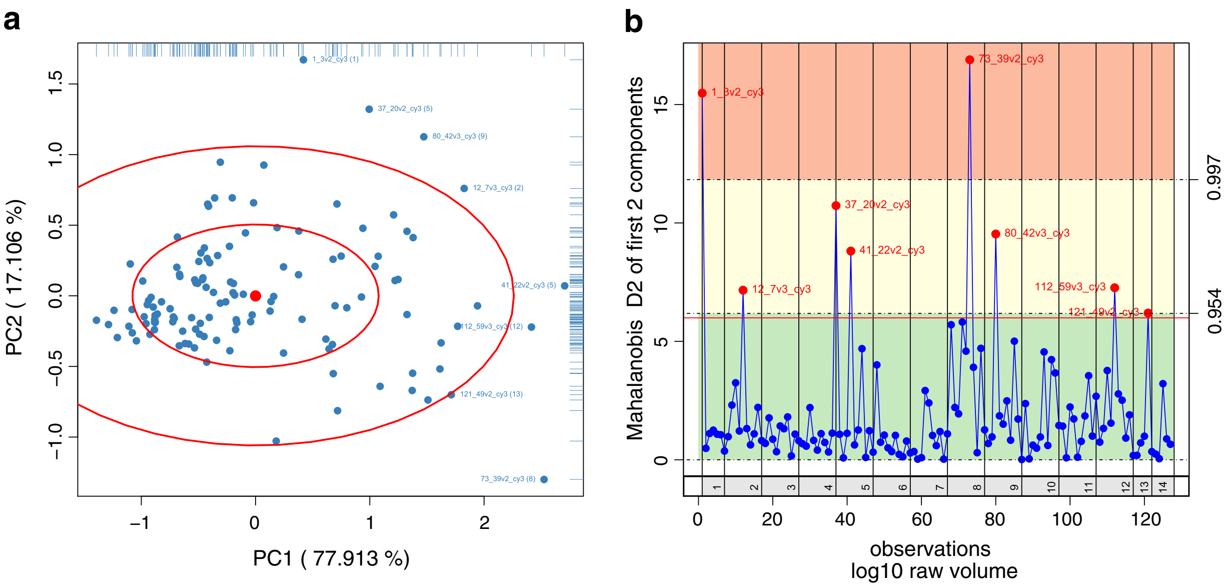


Fig. 3 – a) PCA plot of the first two principal components for cluster group 2 across all 127 standards, as an example of a 'novel' cluster in PCA space. The red ellipses represent normal probability contours (population density bounds) of 50% and 95% respectively. Gel details are also shown for standards outlying the outer ellipse, displaying run number (of 127), volunteer number, visit, and dye channel, then the 2D gel batch (of 14). b) The control chart generated from the Mahalanobis D2 distance metric for these data. Gels are plotted in batch running order with vertical lines separating the second dimension batches. The red line indicates the 95% population density bound outside which gels are flagged as outliers; such standards are also highlighted in red. For visualisation purposes, the 2 SD and 3 SD bounds of the data set are also shown in yellow and red and marked on the right-hand side.

Please cite this article as: Jackson D, Bramwell D, Application of clinical assay quality control (QC) to multivariate proteomics data: A workflow exemplified by 2-DE QC, J Prot (2013), http://dx.doi.org/10.1016/j.jprot.2013.07.025

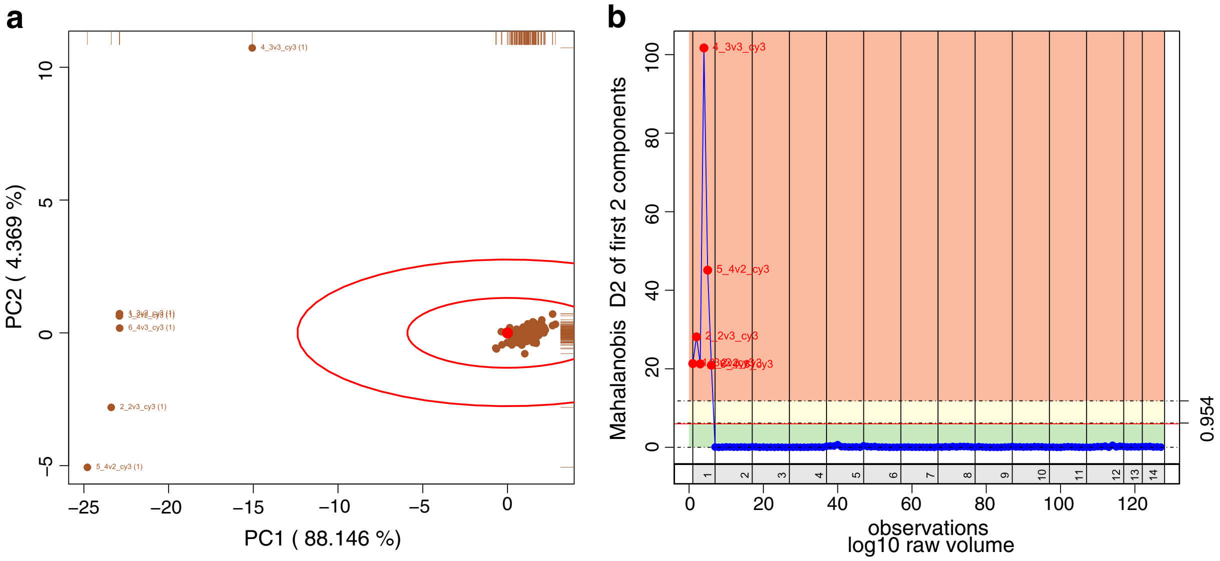


Fig. 4 – a) PCA plot of the first two principal components for cluster group 7 across all 127 standards, showing a 'known' technical effect in PCA space. The red ellipses represent normal probability contours (population density bounds) of 50% and 95% respectively. Gel details are also shown for standards outlying the outer ellipse, displaying run number (of 127), volunteer number, visit, and dye channel, then the 2D gel batch (of 14). b) The control chart generated from the Mahalanobis D2 distance metric for these data. Gels are plotted in batch running order with vertical lines separating the second dimension batches. The red line indicates the 95% population density bound outside which gels are flagged as outliers; such standards are also highlighted in red. For visualisation purposes, the 2 SD and 3 SD bounds of the data set are also shown in yellow and red and marked on the right-hand side.

is still true in this case that adjacent gels to an outlier may be perfectly 'normal' despite the temporal correlation. The benefit of the run-order control chart here is that it visually reveals such trends, and particular batches of reagents and supplies may be investigated to most effectively address an issue; this only emerges from such an ordered plot.

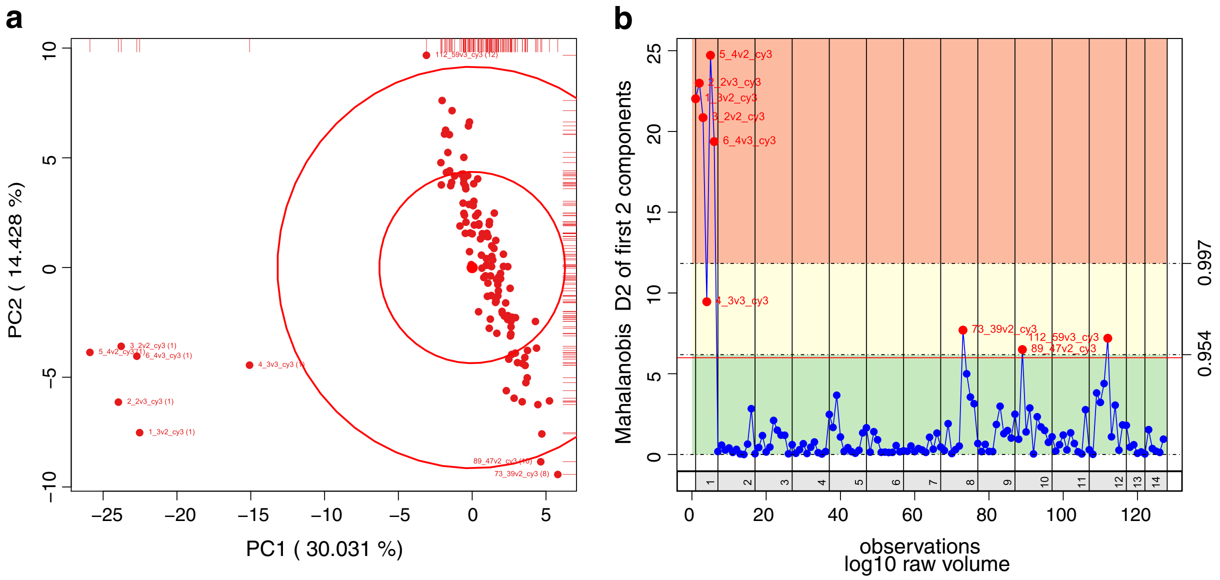


Fig. 5 – a) PCA plot of the first two principal components for all features across all 127 standards. The red ellipses represent normal probability contours (population density bounds) of 50% and 95% respectively. Gel details are also shown for standards outlying the outer ellipse, displaying run number (of 127), volunteer number, visit, and dye channel, then the 2D gel batch (of 14). b) The control chart generated from the Mahalanobis D2 distance metric for these data. Gels are plotted in batch running order with vertical lines separating the second dimension batches. The red line indicates the 95% population density bound outside which gels are flagged as outliers; such standards are also highlighted in red. For visualisation purposes, the 2 SD and 3 SD bounds of the data set are also shown in yellow and red and marked on the right-hand side.

Please cite this article as: Jackson D, Bramwell D, Application of clinical assay quality control (QC) to multivariate proteomics data: A workflow exemplified by 2-DE QC, J Prot (2013), http://dx.doi.org/10.1016/j.jprot.2013.07.025

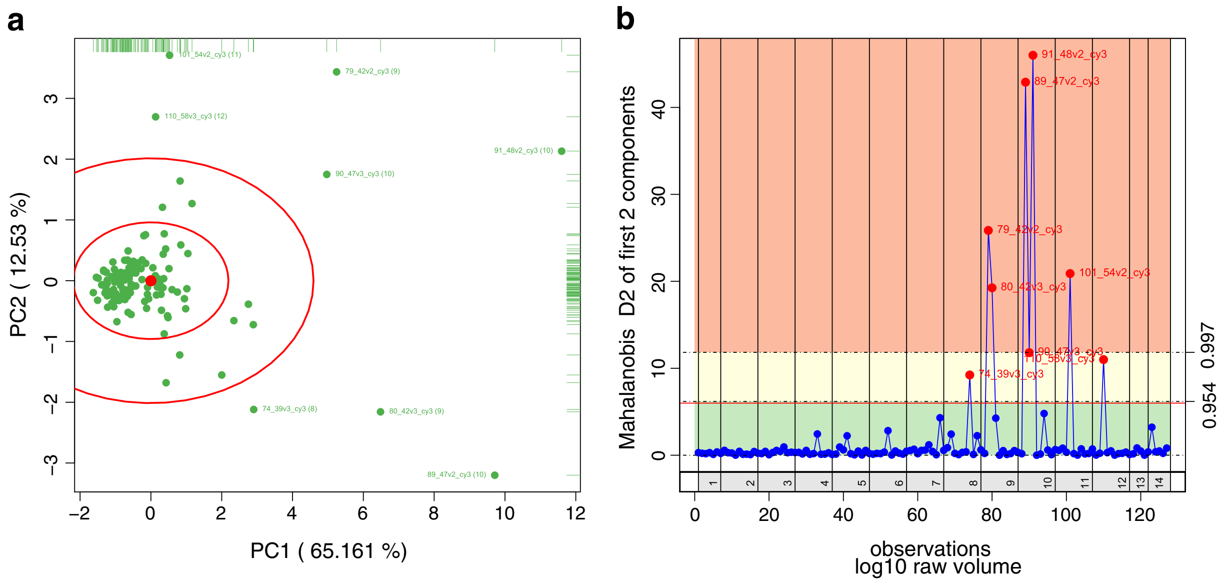


Fig. 6 – a) PCA plot of the first two principal components for cluster group 3 across all 127 standards. The red ellipses represent normal probability contours (population density bounds) of 50% and 95% respectively. Gel details are also shown for standards outlying the outer ellipse, displaying run number (of 127), volunteer number, visit, and dye channel, then the 2D gel batch (of 14). b) The control chart generated from the Mahalanobis D2 distance metric for these data. Gels are plotted in batch running order with vertical lines separating the second dimension batches. The red line indicates the 95% population density bound outside which gels are flagged as outliers; such standards are also highlighted in red. For visualisation purposes, the 2 SD and 3 SD bounds of the data set are also shown in yellow and red and marked on the right-hand side.

As a further illustration of the relative scattering of outliers, one can consider lag plots for the three examples above. These are provided in the Supplementary Results material as Supplementary Results Figures SR9 to SR11 (Figure SR9 for cluster 2, Figure SR10 for cluster 7, and Figure SR11 for cluster 3) and reiterate the trends described above. Again, this relies upon using the run order to provide temporal data.

The control charts for the seven clusters not already provided are in the Supplementary Results Material (Supplementary Results Figures SR2 to SR8, part (b) of each alongside their PCA plots). For these other 'novel effects', the same general point applies; one gel cannot be assumed to act as a reporter for quality of its batch or adjacent gels, which may have 'normal' values. Most outlier sets are scattered throughout the data set. As well as for the second dimension batch, this is also true for other batch metadata factors (data not shown, but reference may be made to the cited study metadata). Variation in the 'majority of gel' profile for cluster 10 was also noted with some outliers; this is to be expected, as previously, and is not of concern.

The total numbers of outlying gels in each cluster are summarised in Table 1. By comparison of the results, we determined that 50 non-redundant gels were outlying in at least one of the ten groups for all 127 gels; excluding the 6 'known issue' gels, 44 non-redundant gels were outliers in at least one of the groups for the main 121 standards. The fact that seven times as many gels outlie for at least one issue as, on average, outlie within any given issue (mean QC flags per cluster = 7.1) indicates the spread of these localised effects, and the importance of treating them separately; incidence patterns for outliers vary notably between clusters.

Reiterating this, the control chart for the entire gel profile is shown in Fig. 5(b). As with the PCA plot in the previous section, it is clear that the known technical issue is still detected in the 6 gels with this effect, given its sheer magnitude (see Section 4.2). However, few other gels now outlie, only 3; the specific technical effects detected in the clusters are generally lost in the whole-gel profile. Comparing the 50 gels with at least one outlying flag across the clusters to these 9 here, this again demonstrates the need to subdivide the profile.

Whilst one might prefer to use whole-profile PCA and successively zoom in on more and more eigenvectors instead of taking our approach, the difficulty in such assessment is establishing and separating assignable causes for each source of variance. By sub-selecting the profile up front to generate targeted metrics, this is made much more easily achievable. For the interested reader, this is illustrated and described further in Supplementary Results Figures SR12 to SR22, which also demonstrate that the same effects are detectable but that interpretation is not as convenient for assignable technical cause.

The key points here are that i) outliers for various technical effects are not detected by the whole-gel approach, but are detected by the hierarchical clustering approach, with effects discovered representing sensible monitoring targets; ii) whilst PCA plots are a useful visualisation, control charts are important to fully examine drift; and iii) gels in the same batch may vary notably in QC status. This concludes a demonstration of rule generation and the type of effects that may emerge; the next point to address is the selection of rules to employ for practical QC decisions (such as which gels to repeat or classify as outliers in analyses), and any prior rule customisation.

4.2. Steps 4 and 5. QC evaluation/refinement and rule selection

Whilst it is important to monitor technical drift in any rule over time to fully understand a system, in practice some effects may not be important for the study in question owing to limited magnitude, not affecting features of interest, or other reasons. Also, for rules selected as important, the cut-offs employed for, and consequences of, rule violation may vary and it is up to a user to decide these. Perhaps the simplest initial approach is to flag up a fixed proportion of gels for a given issue, e.g. any outside a 95% confidence interval as we employed for demonstrative purposes, and examine them. This examination is key; causes of monitored effects should be assignable to be successfully monitored and addressed. Here, we look at some considerations to this end illustrated in the example data set.

When monitoring a number of QC variables, the odds of a given run outlying for at least one variable rise rapidly even by chance (as we discussed in [24]); the odds per gel of no issues being flagged randomly here are simplistically 0.95^{10} assuming independence of the separated issues, i.e. as low as approximately 60%, reflected in the 50 non-redundant outliers we see in 127 gels. This means that applying arbitrary cut-offs and rejecting gels failing any rule would be over-stringent and 50/127 gels being flagged is likely to be a gross overestimate of actual ‘issues’. All that cluster discovery-based QC rules and a single distance metric can indicate is that a run outlies for a given effect; they do not (alone) tell a user whether the effect is important. Even for assignable technical variation, users should consider whether a particular cluster represents an issue of concern or not, whether outliers represent the extent of the effect a user would be concerned about, and whether the cluster represents a biologically relevant feature or could even be split or removed from the analysis. The last point may be based on location irrespective of the underlying protein identity, or entail MS-based protein identification/working from existing master maps.

One possible refinement before applying a rule could be to adjust cut-offs using the magnitude of the effect. Magnitude could also be used to evaluate whether to apply a rule at all. In Table 1, we summarise the difference between the means of outlying and in-control gels for all the cluster groups using the 95% confidence bound as the cut-off. It is evident that as our limits were defined, some effects vary between outliers and in-control gels in a manner that is considerable; greater than typical fold changes employed in proteomics studies, e.g. 2-fold. Clusters 1, 2, 3, 7, and 8 correspond to this level of change in terms of an admittedly simplistic conversion from $\log_{10}$ difference to a fold difference (>0.30103 to >2-fold). This demonstrates the potential underlying effect size of system variance, which is important, as any bias that may filter through to the final data may be critical. Users may wish to prioritise or customise thresholds based on such calculations. This also demonstrates that specific effects of notable magnitude have been effectively separated by this analysis, as against an only slight difference between outliers and in-control samples for the whole profile.

This does not guarantee that biological effects would be overwhelmed in discovery by such technical variance; DIGE or other normalisation schemes may mitigate the effect with varying degrees of success. Post-normalisation differences between outliers and in-control runs for the features in the specific cluster could also be examined where appropriate, allowing an estimation of the technical effect size that is being detected in the final data space for that rule. The feasibility of this approach would depend on the normalisation employed. We discussed the effects of normalisation on QC in more detail in [24] and would refer the reader to this publication for more background, especially on the effects of DIGE.

In the case of some ‘strong’ effects (e.g. clusters 2 and 7), a simple rise or fall in density is evident and magnitude may be a good way to assess the effect (Fig. 7 shows this for cluster 2 and Fig. 8 for cluster 7). However, magnitude is not always effective. Even where there is little change in mean volume this does not rule out a strong shift in density within the pattern; for example considering cluster 6 (Fig. 9). In this case, the most striking behaviour is a shift in density, and a user might wish to investigate this within the chain of spots as part of their rule characterisation. Cluster 6, with its shift in density, illustrates the basis for one possible rule modification. Breaking up clusters into smaller sets and plotting those separately in new rules would be an option for QC if only a part of the cluster contained analytical features which were of importance to the user and vulnerable to technical artefacts. For clusters such as 4 and 8, where there is a dominant group but also some features in a different location, there would be the option to separate the cluster spatially on this same basis. Modifications would not represent an arbitrary measurement, but rather focussing a known source of technical variation onto those regions of it that matter to a user.

Cluster 6 also illustrates the same point through a consideration of p values for the t-test between in-control and out-of-control gels, which are also shown in Table 1. Clearly, for some straightforward effects such as 2 and 7, the trend is reflected in a significant p value. However, just as with magnitude, such significance cannot be the lone measure of whether a cluster is of importance, as cluster 6 is of potential interest despite a nonsignificant p value. The p values largely corroborate magnitude-based assessment specifically, as might be expected.

Cluster 3 is a specific example of a cluster that would likely not be used for rejection. Its location is outside the main profile and its initial spot detection appeared to be largely owing to the presence of a technical effect that does not overlie ‘real’ spots. Whilst of use for alerting a user to the development of albumin-related technical artefacts, this would not necessarily indicate a direct concern for the quantitation of other relevant features in discovery.

Given that some effects were of very low magnitude and lacking any confounding sub-shifts on examination (cluster 9 being an example), and given the presence of edge-of-profile effects in several clusters where we would expect and not be as concerned by higher variability, not all the clusters and their outliers would be important to us. We would not employ all the clusters and charts as rules for QC. We would therefore also not repeat the process to derive further clusters; ten clusters were sufficient in this case to begin to reveal such

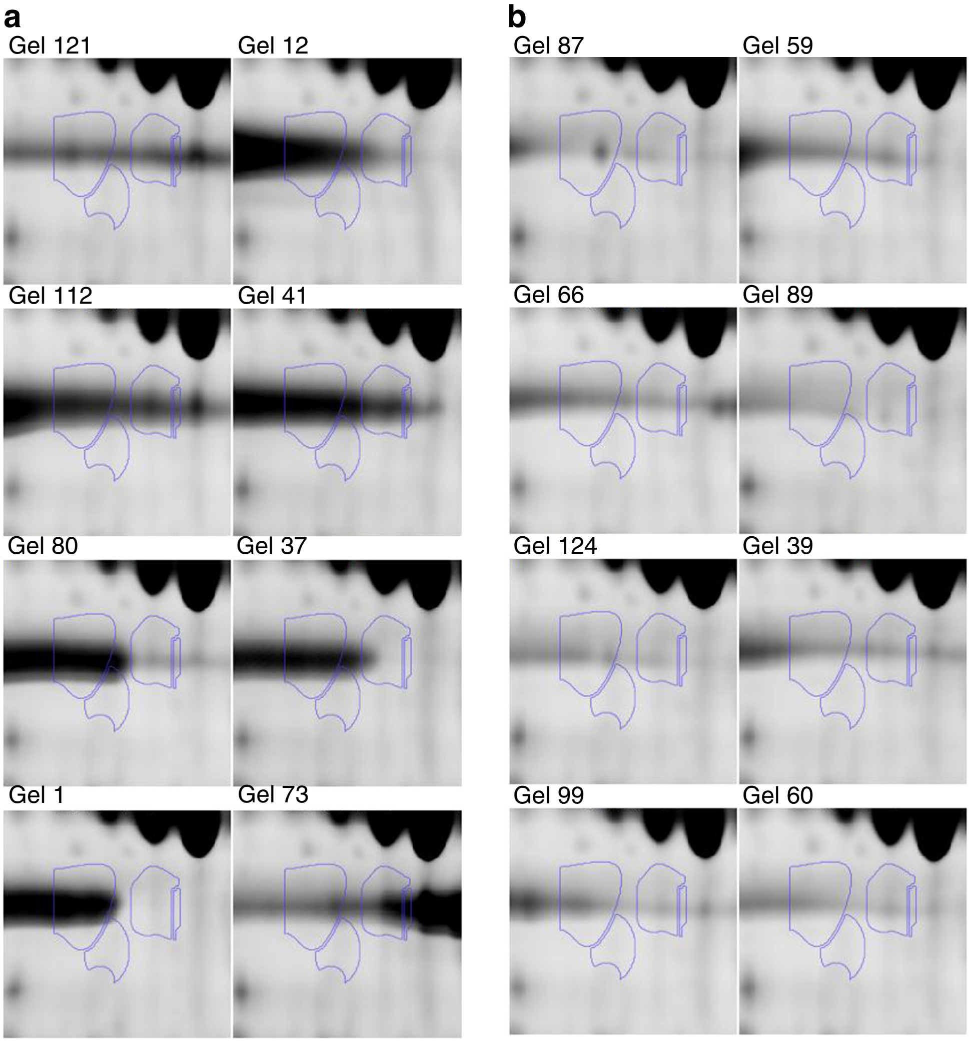


Fig. 7 – a) Montage view of the region containing cluster 2 in all eight outlying standard channels. Images are ordered left–right then top–down in order of increasing Mahalanobis D2 distance. b) Montage view of the region containing cluster 2 in eight standard channels that do not outlie (low Mahalanobis D2 distance). Images are ordered left–right then top–down in order of increasing Mahalanobis D2 distance.

'false positive' effects to us. However should further derivation be preferred then step 6 can be carried out, which simply refers to repeating the process at a different separation distance to generate a different number of groups for investigation (both original and novel groups could provide the basis for rules and it would be possible for a given feature to be part of more than one group). We would likely eliminate the feature comprising 'cluster' 1 from discovery analysis entirely, regard edge-of-gel features with caution but allow them to remain, and in the resulting pattern select to follow clusters 2, 3, 6 and 7 for QC, with cluster 3 only being monitored to indicate albumin streaking and not for rejection purposes (since it does not affect the useful spot profile and the correlation analysis suggests that it has no material impact on other spots within the gel). This would represent a focus on the notable lateral changes in density associated with albumin IEF effects (outside of the known base of gel effect). By these measures of clusters 2, 3, 6 and 7 as rules, without any refinement of the rule limits from the initial cut-off, 25 non-redundant gels would outlie, or 19 not including the known issue, and we would assess the level of effect of being an outlier on raw volume at different distance metrics to refine the limits as not all of those gels would in reality be problematic. One benefit of initially setting a relatively stringent cut-off producing potential false positives however, as done here, is that this will give a user the chance to see all their true positives (genuine out-of-control samples) in the process of refining their rules. It may be sensible to start with a high proportion of such false positives and move the limits 'upwards' to reflect a more specific test. For this reason the 95% cut-off used in this study represents a good starting point—whereas a 1_{3s} rule [21] would flag only runs outlying the central 99.7% of the population, and the effect this would have can be seen on the control charts. Outliers will always arise owing to inherent scatter in a distribution, and users must potentially accept some false positives, except in simple on/off cases.

Note that even at the stage of a 'verified' QC fail for a rule in practice, the consequences are up to the user. Whether that gel is rerun entirely or those features removed from that data

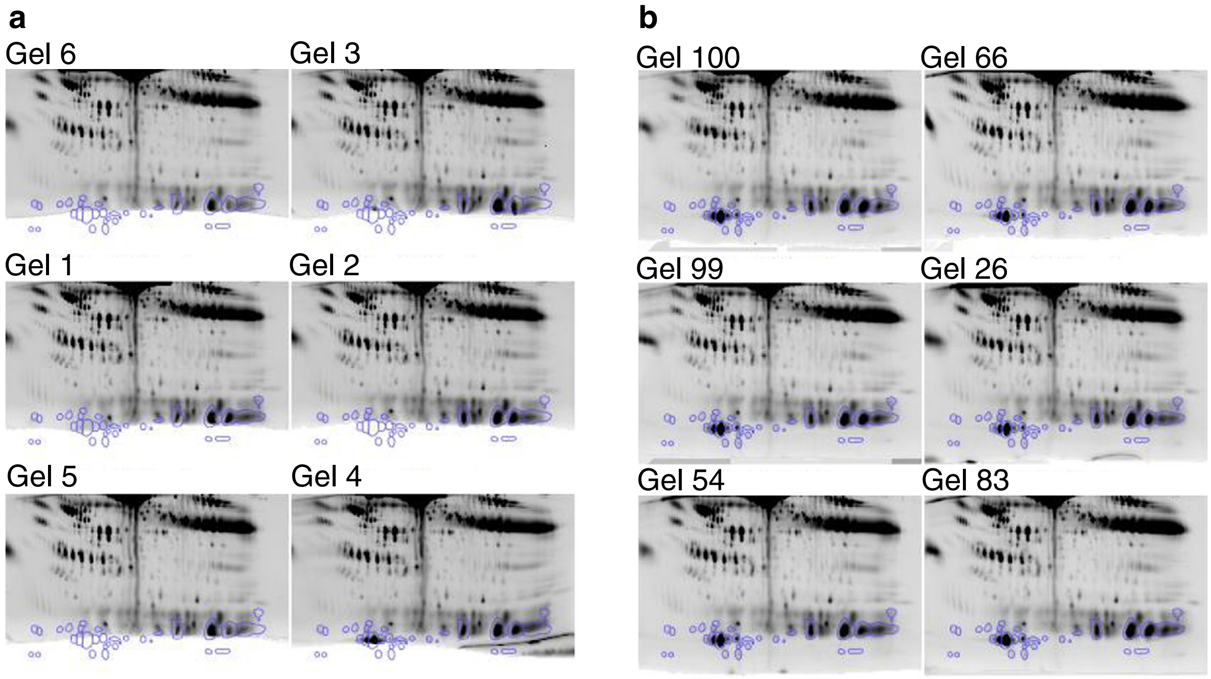


Fig. 8 – a) Montage view of the region containing cluster 7 in all six outlying standard channels. The 'missing region' is clear below the spots concerned. Images are ordered left–right then top–down in order of increasing Mahalanobis D2 distance. b) Montage view of the region containing cluster 7 in six standards that do not outlie (low Mahalanobis D2 distance). Images are ordered left–right then top–down in order of increasing Mahalanobis D2 distance.

set is a decision to be made. This will depend on the user's preferences and how downstream analyses might handle missing values. The simplest default option is to repeat the gel, but a user may decide this is not necessary based on the features affected.

Overall, these comments provide options for consideration and to complete the narrative for our example data set. None of our focus upon particular clusters in any way reduces the importance of monitoring for any co-ordinated effects in the data over time. As a core set of recommendations users should: i) track all rules whether used to pass/fail a gel or not, at least until it is clear which effects are of importance; ii) examine what a cluster and outliers for that cluster represent, to make a context-based decision on the consequences of flagging an outlier; and iii) customise rules/cut-offs if desired based on the appearance, location and magnitude of the effect. We also suggest iv) not rejecting a sample based on the 'bulk' profile comprising the majority of features (on raw volume as here, at least).

4.3. *Summary of results*

Using repeated runs of standards, one can systematically identify a moderate number of correlated feature groups, and through PCA and associated distance metrics, generate effective control charts for outlier selection and targeted QC. Rule customisation may be necessary, and users should be certain to investigate the nature of the effect behind a rule. Outlying does not directly imply failing QC. Because this QC approach must be intrinsic to the gel, DIGE is an effective means by which to QC a gel containing a sample using the repeated standard.

5. Discussion

5.1. *Overview*

We present a workflow for QC of 2-DE data illustrated in a publicly available plasma DIGE data set. Whilst we demonstrate this process for a known primary source of variation in 2-DE, the second dimension gel batch, any metadata recorded could be compared on the control chart. Also, though there is always the option to apply control charts directly to a single feature/analyte of interest as is common in clinical chemistry, our approach balances maximising the coverage obtained (to assess the whole gel) with sensitively detecting separate assignable issues. The key elements of this approach that build upon [24] are the use of hierarchical clustering (for quantitative feature selection) and the Mahalanobis D2 distance metric (for a single quantitative distance metric incorporating the effects of multiple reporters of QC).

This process applies the QC metric to every run intrinsically, which was seen to be essential in 2-DE. In other systems, there have been calls to ideally adopt intrinsic QC (e.g. in [35]) which this method addresses by design. Another implication of the fact that QC must be intrinsic is that one cannot fail an entire batch based on a given run. This does not prevent the number of individual failures being used to evaluate batch quality to monitor aggregate statistics.

Please cite this article as: Jackson D, Bramwell D, Application of clinical assay quality control (QC) to multivariate proteomics data: A workflow exemplified by 2-DE QC, J Prot (2013), http://dx.doi.org/10.1016/j.jprot.2013.07.025

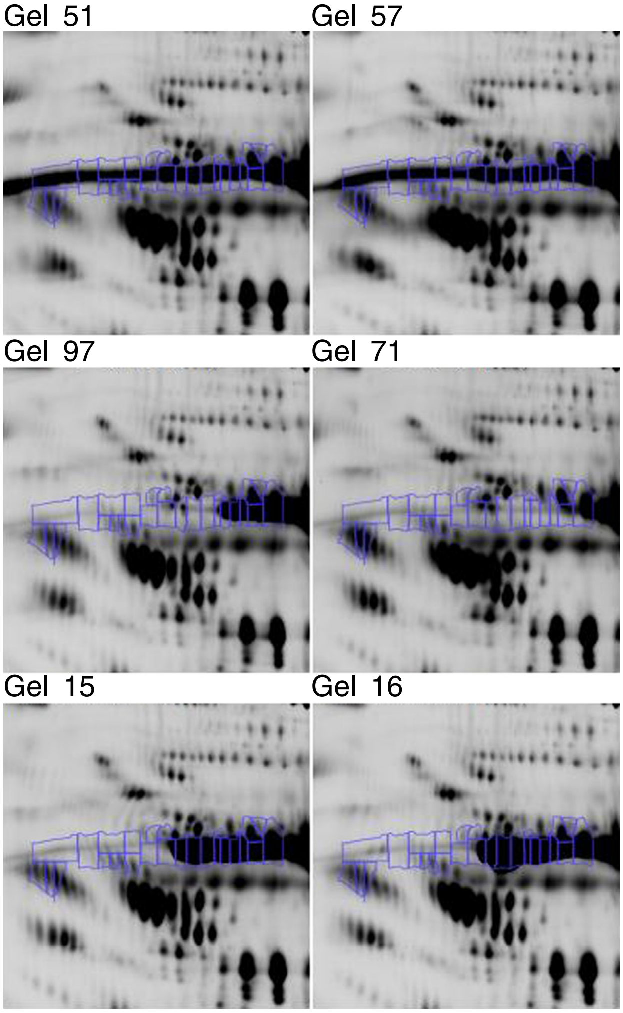


Fig. 9 – Montage view of three pairs of gels at different Mahalanobis D2 distances for cluster group 6 selected to show how the feature varies (not the full outlier set). The distances for these six gels are: 51 = 0.008, 57 = 0.022, 97 = 4.147, 71 = 4.302, 15 = 9.711 and 16 = 11.611. The 'most outlying' gels have a strong shift in density, despite a similar mean feature volume.

It should be noted that the presence of technical effect outliers as described here does not imply a lack of reproducibility in the data examined, or 2-DE in general. We intentionally analysed non-normalised data maximising the chance of detecting bias; we set distance limits lower than commonly-used single-run rejection criteria on control charts; outliers are statistically implicit and it is how and why they outlie that is key; and globally, 2-DE is known to be highly reproducible, especially as analysed by DIGE, post-normalisation (further demonstrated for these data in [24]). In this work, 50 gels are flagged by any putative rule. However, this is prior to any refinement, rule selection or consideration of importance (Section 4.2) and with ten rules a high number of runs would outlie in at least one cluster even by chance. Subselecting rules of most interest reduced this value to 25 at the over-stringent cut-offs employed which would be raised in practice. Therefore these data merely show the ability to effectively capture the most outlying gels for any given technical effect and examine them appropriately. It also allows one to make QC follow-up very targeted by usually only flagging a small region of the profile for review.

The relevant technical effects we would formally QC on occupied a small proportion of the profile, with most features instead being part of cluster 10, a group of 775 features, and several other clusters not representing variation of concern, as described in the results. Whilst we would not reject on the basis of the large cluster, it is nonetheless worth highlighting that we could QC those 775 features together with a 5% false positive rate in one rule having removed confounding effects. This means that the whole-gel approach to QC in [24] can still effectively be part of this approach, if desired, following the bulk profile once specific effects have been separated out into their own rules.

The greatest technical drift is in the albumin-affected regions directly laterally and vertically from the central albumin forms and we would focus QC primarily upon lateral albumin effects; in the original study, it was noted that retaining albumin in the sample but removing saturated albumin material from analysis might reduce the variability over the whole profile relative to introducing extra handling steps to remove albumin. Indeed, depletion strategies may offer relatively little benefit for 2-DE [36,37]. That noted, the dynamic range of the proteome is a significant challenge and albumin clearly represents a focus of technical variation in local areas; this paper does not address and makes no strong recommendation either way on depletion, save to note that the technical effects of retaining albumin were detected and quantitated by our QC, demonstrating its effectiveness.

We apply this approach to a moderately large DIGE set; whilst there is no formal minimum size of experiment for which the approach is of use, naturally it will be of most benefit where several batches of any metadata factor of interest are present to allow the narrowing down of effect progression with time. Proteomics experiments should be of such a size for powering and system characterisation concerns in any event. This also leads on to the situation of the greatest benefit for QC: ongoing monitoring.

5.2. *Ongoing QC*

The example workflow in this paper relies upon the presence of an internal standard and only covers retrospective QC of an entire experiment. However as proteomics experiments improve in terms of sample numbers, ongoing QC will (and should) become increasingly important. We advocate applying a similar approach to ongoing QC which can unlock the full potential of the control chart approach for ongoing monitoring. In brief, we suggest a DIGE pilot, with training standards for rule derivation, followed by a transfer of rule implementation to test sample channels. This transfer relies on a dye-effect correction transform of the samples into the standard PCA space. This generates rules in the sample space for ongoing QC. Running further standards over time can refine and recalibrate rules if desired. In fact, in-house data show that this is effective and those data will form the basis of a future publication. Also, rule derivation directly from samples is effective (data not shown). Limitations of scope mean that we focus here upon the message that the profile must be appropriately subdivided, quantitatively assessed and technical effects then subject to verification. We will return to this topic. However as a brief demonstration of the feasibility of relating standard channel technical effects to those in the sample channel, Supplementary Results Figure SR23 shows that for cluster 3, technical effects do correlate very well between the channels. Also in terms of submitted data supporting this, in our introductory paper on statistical process control in proteomics, we showed that control chart plots for samples alone are similar in behaviour to those for their associated standards globally [24].

5.3. *Further analytical customisation*

Further customisations might include the application of pre-filters or weighting to particular features, for example by volume, area, or signal-to-noise ratio, to improve or adapt the metrics. These would often be part of a user's initial 2-DE analysis to generate the profile.

We suggested that normalisation is generally not appropriate for QC analysis. It may sacrifice relevant information, and mask specific outlier effects. This does not preclude carrying out the process on normalised data as well as a second-line approach, or for investigating the extent of a technical effect in normalised space as discussed in Section 4.2. This is covered for the global profile in [24]. Here, the effects of applying VSN (variance stabilisation normalisation, [38,39]) to the data, in this case in $\log_2$ form as using [40], can be seen on Supplementary Results Figure SR24. The hierarchical split is much more even and clusters similar in size as inter-gel differences are smoothed out, and this would not be effective for monitoring the assignable effects we determined in this study; some order is present in the splitting, perhaps reflecting slight intensity correlations across the gel, but the specific notable cluster effects we charted do not stand out.

Our approach is feature-based. This need not be the case; pixel-based analysis approaches would also be an option with pixel intensity at a particular location providing a measure. This could avoid any biases inherent to the pattern used and spot boundaries (e.g. [41]). However, the process is effective with simpler feature-based pixel binning; feature-based analysis is also a common approach to 2-DE analysis and that most likely to be applied by the majority of users. We use SameSpots software, but this need not be the case for this approach. However, the capability to generate data sets with no missing values is required.

By dint of the approach to generating clusters, groups would be derived even if co-ordinated technical effects are not present. However, the evaluation is proof against this as such groups would be expected not to show spatial correlation. A sensible approach is to derive rules until unimportant, bulk or low-magnitude effects are observed—at this point, since the most significant multivariate effects will already have emerged, one could regard investigation as complete at that time. In this example data set, 10 groups were already sufficient to achieve this.

5.4. *Generalisability*

Variables specific to this set including the sample type and equipment used mean that our conclusions are situation-specific in terms of the groupings and rules derived. The approach as demonstrated here also relies on fluorescence-based DIGE. This is because QC must be intrinsic so a multi-channel analysis is required to accommodate the sample and the standard. However this does not limit the implementation of the approach, as a capability to carry out fluorescence imaging is common in high-quality 2-DE laboratories. Also, as per Section 5.2, we have seen and intend to show that sample-only QC is possible, which would open this type of approach to 'single-stain' analyses.

The treatment of a proteomic profile as multivariate feature sets is not limited to 2-DE. In fact, in MS-profiling studies, Mahalanobis distance metrics are used already to characterise measures including test spots on sample chips for chip rejection, comparing replicate variability as against original standard runs, and for outlier identification using the peptide profile, e.g. [35,42–44]. These provide effective QC solutions for drift measurement and our approach is related to these in the use of PCA and distance assessment, but not identical. In effect, our approach is a modification of these approaches, inserting a cluster selection step to separate the QC into sections representing different technical effects. It would be interesting to assess the effectiveness of our approach in MS data sets applied to peak features as opposed to spot features. On the assumptions that i) there are technical biases that can be detected by multivariate analysis, ii) these result in correlated effects across a sub-section of the profile, and iii) are assignable as belonging to a relevant cause, this should be feasible.

However the demonstrated approach relied upon co-migrating standard and sample to allow a standard to be used for QC but the sample to be present in the same run. This would most closely mirror MS-based approaches where an intrinsic standard for each feature can be present, such as isotopic or other labelling approaches. As per Section 5.2, we will go on to show that for gels effective QC can be carried out on samples directly on an ongoing basis, which would mirror MS-based methods without an intrinsic standard. Interestingly, in MS-based profiling Mahalanobis-based metrics are currently applied to individual runs, suggesting this is feasible [43,44]. The key element of our approach, resolving separate assignable effects, should also be possible. For example, for LC–MS, ion intensity map analysis may localise issues to particular retention times and m/z ranges, e.g. [45]. Commonality of effects in other areas such as adduct analysis may also be of interest. One would also need to assess the best data processing approach(es) for MS-based application of our methods as these may vary from gel-based data, and the best manner of rule assessment.

For some recent examples of existing variability assessment and QC in MS approaches, the reader is referred to the citations in the paragraph above, along with [11,46].

6. Conclusions

Effective proteomics QC requires both rigour and correct handling of the complexity of the proteome. We show a method for achieving this in practice using 2-DE as a specific example. This method is one way to achieve the goal, but the critical element is that study design incorporates appropriate

QC of at least this level of detail. There is no reason why development of these processes should constitute an unreasonable time investment, and their importance cannot be underestimated if proteomics is to deliver its potential. Further work will focus on applying this approach on an ongoing basis for ongoing non-DIGE studies.

Disclosure statement

Both authors are employees of Biosignatures Ltd, which was formed by a demerger from Nonlinear Dynamics; the two companies share the same CEO, Will Dracup.

Acknowledgements

The authors would like to acknowledge colleagues at Biosignatures: Dr Ben Chaffey for helpful technical discussions and Mark Lambert for generation of a useful montage visualisation application which assisted figure preparation. The authors would also like to acknowledge Dr Robert Tonge, now at Waters Corporation, for both supervising the original study from which the data set is derived, and many valuable discussions on approaches to ensuring the quality of proteomics data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jprot.2013.07.025.

REFERENCES

[1] Petricoin E, Ardekani A, Hitt B, Levine P, Fusaro V, Steinberg S, et al. Use of proteomic patterns in serum to identify ovarian cancer. Lancet 2002;359(9306):572–7.

[2] Baggerly K, Morris J, Coombes K. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. Bioinformatics 2004;20(5):777–85.

[3] Baggerly K, Morris J, Edmonson S, Coombes K. Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer. J Natl Cancer Inst 2005;97(4):307–9.

[4] Ransohoff D. Lessons from controversy: ovarian cancer screening and serum proteomics. J Natl Cancer Inst 2005;97(4):315–9.

[5] Diamandis E. Cancer biomarkers: can we turn recent failures into success? J Natl Cancer Inst 2010;102(19):1462–7.

[6] Rai A, Gelfand C, Haywood B, Warunek D, Yi J, Schuchard M, et al. HUPO Plasma Proteome Project specimen collection and handling: towards the standardization of parameters for plasma proteome samples. Proteomics 2005;5(13):3262–77.

[7] Mischak H, Apweiler R, Banks R, Conaway M, Coon J, Dominiczak A, et al. Clinical proteomics: a need to define the field and to begin to set adequate standards. Proteomics Clin Appl 2007;1(2):148–56.

[8] Gibson F, Anderson L, Babnigg G, Baker M, Berth M, Binz P, et al. Guidelines for reporting the use of gel electrophoresis in proteomics. Nat Biotechnol 2008;26(8):863–4.

[9] Mann M. Comparative analysis to guide quality improvements in proteomics. Nat Methods 2009;6(10):717–9.

[10] Mischak H, Allmaier G, Apweiler R, Attwood T, Baumann M, Benigni A, et al. Recommendations for biomarker identification and qualification in clinical proteomics. Sci Transl Med 2010;2(46) [46 ps42].

[11] Tabb D. Quality assessment for clinical proteomics. Clin Biochem 2013;46:411–20.

[12] Paulovich A, Whiteaker J, Hoofnagle A, Wang P. The interface between biomarker discovery and clinical validation: the tar pit of the protein biomarker pipeline. Proteomics Clin Appl 2008;2(10–11):1386–402.

[13] Ioannidis J. A roadmap for successful applications of clinical proteomics. Proteomics Clin Appl 2011;5(5–6):241–7.

[14] Mischak H, Ioannidis J, Argiles A, Attwood T, Bongcam-Rudloff E, Broenstrup M, et al. Implementation of proteomic biomarkers: making it work. Eur J Clin Invest 2012;42(9):1027–36.

[15] Fung E. A recipe for proteomics diagnostic test development: the OVA1 test, from biomarker discovery to FDA clearance. Clin Chem 2010;56(2):327–9.

[16] Zhang Z, Chan D. The road from discovery to clinical diagnostics: lessons learned from the first FDA-cleared in vitro diagnostic multivariate index assay of proteomic biomarkers. Cancer Epidemiol Biomarkers Prev 2010;19(12): 2995–9.

[17] Ueland F, Desimone C, Seamon L, Miller R, Goodrich S, Podzielinski I, et al. Effectiveness of a multivariate index assay in the preoperative assessment of ovarian tumors. Obstet Gynecol 2011;117(6):1289–97.

[18] Diao L, Clarke C, Coombes K, Hamilton S, Roth J, Mao L, et al. Reproducibility of SELDI spectra across time and laboratories. Cancer Inform 2011;10:45–64.

[19] Martens L. A report on the ESF workshop on quality control in proteomics. Mol Biosyst 2010;6(6):935–8.

[20] Levey S, Jennings E. The use of control charts in the clinical laboratory. Am J Clin Pathol 1950;20(11):1059–66.

[21] Westgard J, Barry P, Hunt M, Groth T. A multi-rule Shewhart chart for quality control in clinical chemistry. Clin Chem 1981;27(3):493–501.

[22] Westgard J, Westgard S. The quality of laboratory testing today. Am J Clin Pathol 2006;125(3):343–54.

[23] Westgard J. Historical perspective on laboratory QC: where we've been and where we're going! Accessed 27/09/2012; URL http://www.westgard.com/history-and-future-of-qc.htm; 2011.

[24] Bramwell D. An introduction to statistical process control in research proteomics. J Proteomics 2013. http://dx.doi.org/10.1016/j.jprot.2013.06.010 [Accessed 19/07/2013].

[25] Görg A, Weiss W, Dunn M. Current two-dimensional electrophoresis technology for proteomics. Proteomics 2004;4(12):3665–85.

[26] Gustafsson J, Ceasar R, Glasbey C, Blomberg A, Rudemo M. Statistical exploration of variation in quantitative two-dimensional gel electrophoresis data. Proteomics 2004;4(12):3791–9.

[27] Mahalanobis P. On the generalized distance in statistics. Proc Natl Inst Sci India 1936;2(1):49–55.

[28] Jackson D, Herath A, Swinton J, Bramwell D, Chopra R, Hughes A, et al. Considerations for powering a clinical proteomics study: normal variability in the human plasma proteome. Proteomics Clin Appl 2009;3(3):394–407.

[29] Alban A, David S, Bjorkesten L, Andersson C, Sloge E, Lewis S, et al. A novel experimental design for comparative two-dimensional gel analysis: two-dimensional difference gel electrophoresis incorporating a pooled internal standard. Proteomics 2003;3(1):36–44.

[30] Wheelock Å, Buckpitt A. Software-induced variance in two-dimensional gel electrophoresis image analysis. Electrophoresis 2005;26(23):4508–20.

[31] R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing3-900051-07-0; 2012 [accessed 28/12/2012; URL http://www.R-project.org/].

[32] Todorov V, Filzmoser P. An object-oriented framework for robust multivariate analysis. J Stat Softw 2009;32(3):1–47 [URL http://www.jstatsoft.org/v32/i03/].

[33] Valcu C, Valcu M. Reproducibility of two-dimensional gel electrophoresis at different replication levels. J Proteome Res 2007;6(12):4677–83.

[34] Jackson D, Herath A, Swinton J, Bramwell D, Chopra R, Hughes A, et al. Supplementary information. Considerations for powering a clinical proteomics study: normal variability in the human plasma proteome . Accessed 28/12/2012; URL ; 2009 http://onlinelibrary.wiley.com; 2009. http://dx.doi.org/10.1002/prca.200800066/suppinfo.

[35] Coombes K, Fritsche Jr H, Clarke C, Chen J, Baggerly K, Morris J, et al. Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. Clin Chem 2003;49(10):1615–23.

[36] Vasudev N, Ferguson R, Cairns D, Stanley A, Selby P, Banks R. Serum biomarker discovery in renal cancer using 2-DE and prefractionation by immunodepletion and isoelectric focusing; increasing coverage or more of the same? Proteomics 2008;8(23–24):5074–85.

[37] Smith M, Wood S, Zougman A, Ho J, Peng J, Jackson D, et al. A systematic analysis of the effects of increasing degrees of serum immunodepletion in terms of depth of coverage and other key aspects in top–down and bottom–up proteomic analyses. Proteomics 2011;11(11):2222–35.

[38] Huber W, Von Heydebreck A, Sültmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 2002;18(Suppl. 1):S96–S104.

[39] Kreil D, Karp N, Lilley K. DNA microarray normalization methods can remove bias from differential protein expression analysis of 2D difference gel electrophoresis results. Bioinformatics 2004;20(13):2026–34.

[40] Huber W. Variance stabilization and calibration for microarray data. R package version 3.24.0. Accessed 28/12/2012; URL http://www.bioconductor.org/packages/2.10/bioc/manuals/vsn/man/vsn.pdf; 2012.

[41] Silva E, O'Gorman M, Becker S, Auer G, Eklund A, Grunewald J, et al. In the eye of the beholder: does the master see the SameSpots as the novice? J Proteome Res 2010;9(3):1522–32.

[42] Cairns D, Perkins D, Stanley A, Thompson D, Barrett J, Selby P, et al. Integrated multi-level quality control for proteomic profiling studies using mass spectrometry. BMC Bioinformatics 2008;9(519).

[43] Schulz-Trieglaff O, Machtejevas E, Reinert K, Schlüter H, Thiemann J, Unger K. Statistical quality assessment and outlier detection for liquid chromatography–mass spectrometry experiments. BioData Min 2009;2(4).

[44] Matzke M, Waters K, Metz T, Jacobs J, Sims A, Baric R, et al. Improved quality control processing of peptide-centric LC–MS proteomics data. Bioinformatics 2011;27(20):2866–72.

[45] Ross M. 5 sample running problems highlighted by ion intensity maps. Accessed 28/12/2012; URL http://blog.nonlinear.com/2011/02/14/ion-intensity-maps/; 2011.

[46] Russell M, Lilley K. Pipeline to assess the greatest source of technical variance in quantitative proteomics using metabolic labelling. J Proteomics 2012;77:441–54.

Please cite this article as: Jackson D, Bramwell D, Application of clinical assay quality control (QC) to multivariate proteomics data: A workflow exemplified by 2-DE QC, J Prot (2013), http://dx.doi.org/10.1016/j.jprot.2013.07.025