RESEARCH ARTICLE

A posteriori quality control for the curation and reuse of public proteomics data

Joseph M. Foster[1], Sven Degroeve[2,3], Laurent Gatto[4], Matthieu Visser[5], Rui Wang[1], Johannes Griss[6], Rolf Apweiler[1] and Lennart Martens[2,3]

[1] EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
[2] Department of Medical Protein Research, VIB, Ghent, Belgium
[3] Department of Biochemistry, Ghent University, Ghent, Belgium
[4] Cambridge Centre for Proteomics, Cambridge Systems Biology Centre, Department of Biochemistry, University of Cambridge, Cambridge, UK
[5] Philips Research Laboratories, Cambridge Science Park, Cambridge, UK
[6] Department of Medicine, Vienna General Hospital, Medical University Vienna, Vienna, Austria

Proteomics is a rapidly expanding field encompassing a multitude of complex techniques and data types. To date much effort has been devoted to achieving the highest possible coverage of proteomes with the aim to inform future developments in basic biology as well as in clinical settings. As a result, growing amounts of data have been deposited in publicly available proteomics databases. These data are in turn increasingly reused for orthogonal downstream purposes such as data mining and machine learning. These downstream uses however, need ways to a posteriori validate whether a particular data set is suitable for the envisioned purpose. Furthermore, the (semi-)automatic curation of repository data is dependent on analyses that can highlight misannotation and edge conditions for data sets. Such curation is an important prerequisite for efficient proteomics data reuse in the life sciences in general. We therefore present here a selection of quality control metrics and approaches for the a posteriori detection of potential issues encountered in typical proteomics data sets. We illustrate our metrics by relying on publicly available data from the Proteomics Identifications Database (PRIDE), and simultaneously show the usefulness of the large body of PRIDE data as a means to derive empirical background distributions for relevant metrics.

Received: September 21, 2010
Revised: February 7, 2011
Accepted: February 21, 2011

Keywords:
Bioinformatics / PRIDE / Quality assurance / Quality control

1 Introduction

The field of proteomics has undergone rapid expansion in recent years [1], morphing what was once the pioneering work of a few laboratories into a wide-spread discipline involving thousands of people and boasting large-scale collaborative efforts [2]. The applications of proteomics extend to a wide range of scientific problems from profiling complex mixtures [1], over studying post-translational modifications [3], to elucidating interaction partners and protein complexes [4]. Proteomics also offers a high-throughput solution for the analysis of clinically relevant samples for disease biomarker discovery [5], with quantified results much closer to the biological function they govern than similar approaches in genomics.

Because of the resulting increased popularity of proteomics as a method to analyse biological samples, publicly available data repositories such as Global Proteome Machine Database (GPMDB) [6], Peptide Atlas [7], Proteomics Identifications Database (PRIDE) [8] and Peptidome [9] have amassed large and ever-increasing collections of proteomics data. The accumulated, publicly available data sets represent a true treasure trove of information, and are therefore increasingly exploited for a variety of downstream purposes.

Correspondence: Professor Lennart Martens, Department of Medical Protein Research, Universiteit Gent – VIB A, Baertsoenkaai, 3 B-9000 Gent, Belgium
E-mail: lennart.martens@UGent.be.
Fax: +32-9-264-94-84

Abbreviations: HUPO, Human Proteome Organisation; **LSA,** latent semantic analysis; **MARS,** multiple affinity removal system; **PPP2,** Plasma Proteome Project 2; **PNNL,** Pacific Northwest National Laboratory; **QA,** quality assurance; **QC,** quality control; **SCX,** strong cation exchange; **tf-idf,** term frequency-inverse document frequency; **TMT,** tandem mass tag

Examples of such data reuse include various types of large-scale analyses, such as experimental technique profiling [10], detection of specific analytes [11], machine learning algorithms for proteotypic peptide prediction [12], optimal selected reaction monitoring transition selection [13] and spectral libraries [14]. In all of these approaches however, it is important to evaluate the suitability of a data set prior to including it in the downstream analysis. This selection process is essentially a quality control (QC) step that can take two forms: the first is a selection based on experimental metadata (e.g. data sets obtained from a given organism or a given mass spectrometer type), while the second is based on the characteristics of the data or corresponding results. It is important to note however that the filtering applied is very much dependent on the context and purpose of the downstream analysis, i.e. the decision on the QC metrics and annotation parameters to use for filtering, and the limits set for calling outliers based on these metrics, will differ from use case to use case. While QC is ideally carried out in-line during sample processing and data acquisition to immediately allow detection of unacceptable errors or artefacts, it remains equally important to have a set of QC metrics that can be applied long after the completion of the wet-lab workflow, when the acquired data has been deposited in a public repository. Indeed, given the large amount of possible uses the data can be put to further downstream, it is unlikely that all relevant metrics and data set characteristics have been explored during the original workflow. Furthermore, these metrics may in fact be preferentially judged in a relative sense rather than an absolute one, for instance when potential outliers are considered against an empirically derived background as estimated across a large number of existing data sets. While efforts aimed at end-stage QC have been undertaken previously [15, 16], these have so far been limited to the detailed analysis of a single mass spectrometer run, detecting potential problems within this single analysis across many of the various stages of sample processing and handling typically encountered during a proteomics workflow. Here however, we instead focus primarily on QC analysis methods that effectively span very many different runs or analyses. Indeed, such QC metrics are ideally matched to the task of performing QC on large amounts of heterogeneous data as obtained from public proteomics data repositories. The two types of QC analyses can thus complement each other quite nicely, with large-scale analyses as described here amenable to initial triage or relatively robust downstream analyses, while more fine-grained, per-run analyses as published previously [15, 16] could be employed on the remaining data sets when desired. Yet, metadata filtering relies on correct annotations, which certainly is not always the case. It is therefore important that automated ways of picking up seemingly inconsistent annotations are put in place, and that these are presented to a human curator for verification and possible correction. The methods employed for this semi-automated curation of public data are quite similar to the methods used for QC however, and we will therefore illustrate the usefulness of several of the metrics presented here in the overall curation process.

2 Materials and methods

For our analysis, we have relied on several publicly available data sets, all obtained from the PRIDE repository (http://www.ebi.ac.uk/pride), as well as some specific data sets in PRIDE from close data collaborators. The following sections provide details on each of the data sets employed in this study.

2.1 HUPO PPP2 data set obtained from the PRIDE repository

All peptide identifications from experiments performed by the Richard Smith Lab at Pacific Northwest National Laboratory (PNNL) submitted to PRIDE in the context of the Human Proteome Organisation's (HUPO) Plasma Proteome Project 2 (PPP2) [17, 18] under accession numbers 8172 to 8544 (inclusive), were retrieved. These 373 experiments represent the analysis of 12 human plasma samples, each subjected to some combination of multiple affinity removal system (MARS)-6 or IgY-12 depletion and cysteine or *N*-glycosylated peptide selection prior to offline strong cation exchange (SCX) chromatography followed by LC-MS (for full protocol details please see [17, 18]).

2.2 Tryptic digestion background data set from the PRIDE repository

In order to generate a reference for the efficiency of a tryptic digest, all PRIDE experiments annotated to have a single digestion step involving trypsin were extracted from PRIDE along with their peptide identifications. This resulted in 4582 experiments, of which 1695 were discarded due to no missed cleavages being detected in those experiments, indicating that the search engine was configured to not tolerate any missed cleavages. The remaining 2887 were used to estimate overall background missed cleavage frequencies and precursor ion mass distributions.

2.3 HUPO Test Sample Study data sets obtained from the PRIDE repository

All HUPO Test Sample Study [19] data sets that contained MS2 spectra were obtained from PRIDE (accession numbers 8130–8158), for comparison against the background distribution of precursor masses.

2.4 Precursor mass and MS2 mass difference data set obtained from PRIDE

In order to do a broad range of QC tests a set of experiments from PRIDE was retrieved that fulfilled the following criteria: MS2 spectra must be present, and must include both precursor ion *m/z* and charge annotation. This selection process resulted in 1438 experiments suitable for our purposes (see Supporting Information Table 1 for a complete list). Particularly fruitful analyses done on this data set included generating an empirically derived precursor mass distribution for data in PRIDE against which single experiments precursor mass distribution could be compared and analysing the frequency of mass difference between filtered MS2 peaks in all the spectra for an experiment.

2.5 MS2 *m/z* distribution data set obtained from the PRIDE repository

In collaboration with the group of Christopher Gerner at The Department of Medicine, Vienna General Hospital, Medical University Vienna all experiments submitted by this group were retrieved from PRIDE to generate a data set of closely comparable experiments where the instrumentation and general protocol employed were highly similar. These experiments were subject to a range of simple QC checks and potential faults detected.

2.6 Quantitative analysis data sets

From PRIDE a set of MS2 spectra (PRIDE accession number 12821) were analysed to ascertain the quality of isobaric labelling. The data set was derived from the analysis of a test sample, acquired by an external supplier for Philips Research on a Thermo-Finnigan Orbitrap in Pulsed Q Collision Induced Dissociation mode. The tandem mass tag (TMT)-6-plex labels from Proteome Science were used, and these were applied to six different samples of human serum. All six samples were subject to identical pre-labelling sample preparation, after which they were mixed in ratios 1:1:3:3:9:9 for TMT labels with reporter masses of 126, 127, 128, 129, 130 and 131 Da, respectively. Processing (de-isotoping and peak picking) of the raw MS data was done with MaxQuant [20] with settings as described in [20].

2.7 Latent semantic analysis of HUPO PPP2 data

The protocol followed here for latent semantic analysis (LSA) follows the approach used by Klie et al. [10]. Concretely, identified peptide sequences were extracted from each PRIDE experiment, and were used to generate a peptide versus experiment occurrence matrix. As described in Section 2.1 each experiment represents a single SCX fraction, and so the peptide versus experiment occurrence matrix is actually a peptide versus SCX fraction occurrence matrix. Term frequency-inverse document frequency (tf-idf) was subsequently applied to this matrix in order to attenuate the signal derived from high-abundance peptides, thus allowing for greater sensitivity of effects produced by less-abundant peptides. This method of weighting the frequency data is summarised by the set of formulas:

Term frequency: where $n_{i,j}$ is the number of occurrences of the peptide (t_i) in experiment d_j, and the denominator is the sum of the number of occurrences of all peptides in the experiment d_j (Eq. 1).

Inverse document frequency is a measure of the general importance of a peptide calculated by taking the logarithm of the total number of experiments |D| divided by the number of experiments containing the peptide (Eq. 2).

Term frequency inverse document frequency: the product of the two previous expressions for each peptide in each experiment (Eq. 3).

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

$$\mathrm{idf}_i = \log \frac{|\mathrm{D}|}{|\{d : t_i \in d\}|} \tag{2}$$

$$(\mathrm{tf\text{-}idf})_{i,j} = \mathrm{tf}_{i,j} \times \mathrm{idf}_i \tag{3}$$

LSA was then carried out on the tf-idf weighted occurrence matrix, resulting in a signal boost by deriving a denoised and less-sparse occurrence matrix. This process effectively allows the inference of a peptide's frequency based on the frequency of other peptides present in the experiment and the co-occurrences of those peptides in other experiments. Based on the LSA-transformed occurrence matrix, all possible experiment versus experiment distances are calculated using the cosine similarity function between each pair of experiment eigenvectors.

$$\mathrm{similarity} = \cos(\theta) = \frac{A \cdot B}{||A|| ||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

The resulting experiment versus experiment distance matrix is then sorted first by overall experimental strategy, and within each strategy by SCX fraction and the sorted result is displayed as a heat map.

3 Results

A typical proteomics workflow consists of a variety of experimental steps, and many of these can cause issues in the acquired data. We therefore present here a set of metrics that are aimed at detecting deviations in a number of key workflow steps: protein depletion, peptide selection and chromatographic separation, proteolytic digest, spectrum acquisition, contamination, sample annotation, and chemical labelling efficiency and recovery in quantification.

3.1 Depletion and separation analysis

Various types of LC form key steps in most proteomics experiments. Reproducible separations are furthermore essential for several quantitative or targeted approaches, where alignment of chromatographic profiles or correct scheduling of peptide elution times are paramount to success [21–23]. In order to asses chromatographic performance and reproducibility across many different elution runs, we analysed the HUPO PPP2 data sets [17, 18] deposited in PRIDE [24], since these contain several similar chromatographic runs, performed on different peptide subsets obtained from human plasma. Based on the identified peptides in each experiment, a peptide versus experiment occurrence matrix was then constructed. In order to accommodate the differences in experimental design that preceded the chromatography runs and that influenced the selection of proteins and peptides, LSA was employed as a

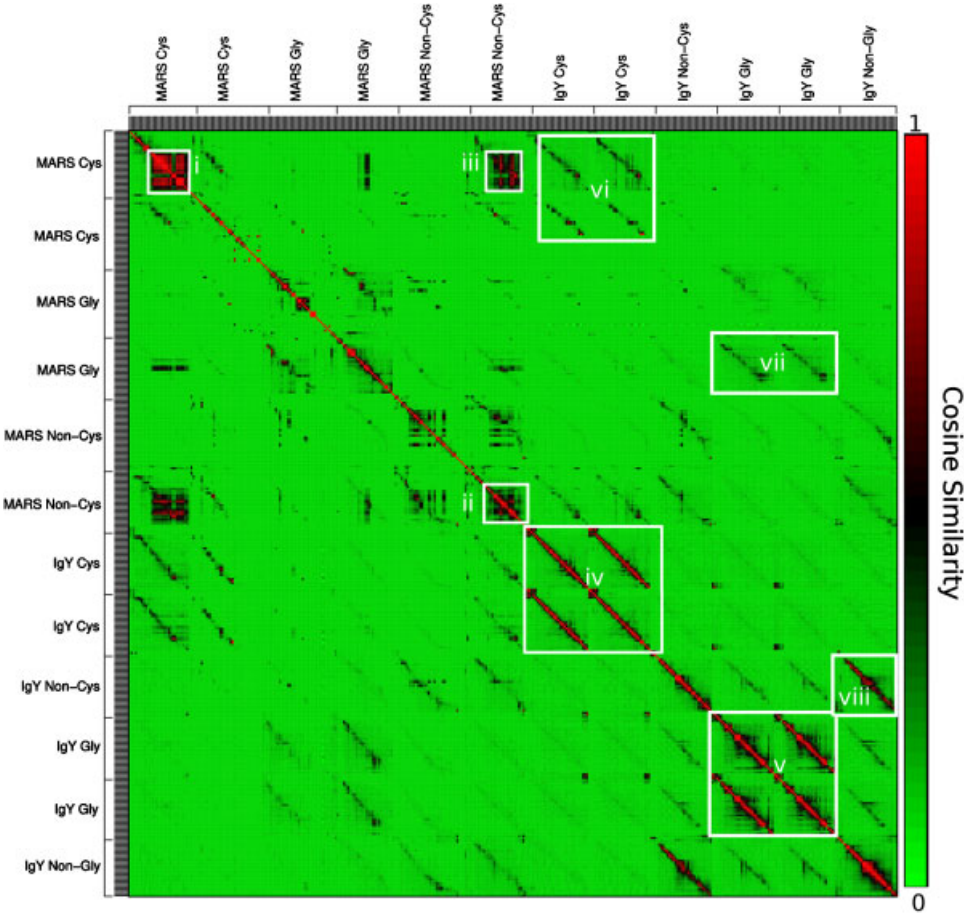

Figure 1. A cosine distance heat map of peptides identified in 373 experiments from the Richard Smith lab at PNNL. Each row and column represents a single experiment in PRIDE, which in turn corresponds to a particular SCX fraction of a sample. Twelve samples were analysed in total, encompassing combinations of MARS-6 or IgY-12 for the depletion step and cysteine or *N*-glycosylated peptides (or the inverse) for the peptide selection, these sample are ordered by protocol and SCX fraction, which are internally further ordered by SCX fraction elution order. Red represents high similarity, green high dissimilarity and black some similarity. The data was subject to tf-IDF and LSA prior to plotting in order to highlight patterns and reduce noise. Numbered highlights are discussed sequentially next. (i) Approximately one-third of the way through the SCX fractionation procedure peptides appear to be bleeding across all subsequent fractions, reducing the separation efficiency and hence the detection sensitivity of the system considerably. (ii) The effect seen in (i) is confirmed here: the separation is performing quite poorly, with bleeding evident. (iii) Additionally, the region highlighted in (ii) shows unexpected similarity between 'MARS Cys' and 'MARS non-Cys' experiments; in theory, the overlap should be extremely small due to the opposite selection procedure. (iv) Slight black blurring around the diagonal indicates peptide identification similarity between adjacent fractions; potentially an early warning sign that the SCX separation performance is starting to degrade. We do see superb reproducibility between samples that have undergone the same sample preparation protocol, however. (v) Further evidence of the points made in (iv): somewhat further increased blurring, but excellent reproducibility of identifications obtained via IgY depletion. (vi) Shows reproduciblity in identifications between different depletion methods; a good QC measure but it also indicated the depletion method does alter the peptides you detect in addition to removing highly abundant proteins. (vii) Another example of the points raised in (vi), but now for a different peptide selection technology. (viii) An unexpected similarity between 'IgY Non-Cys' and 'IgY Non-Gly' sample separation.

powerful signal booster and noise filter prior to experiment distance calculation (see Section 2 for details).

The resulting peptide-based experiment versus experiment similarity is provided graphically as a heatmap in Fig. 1. The experiments, each corresponding to an SCX fraction, are grouped by depletion technology and peptide fractionation method, and are then ordered by their elution order. Highlight (i) on Fig. 1 shows a region of high similarity between the later SCX fractions in the first 'MARS Cysteine' analysis. It appears that peptides have begun to bleed across the fractions from a certain timepoint, indicating possible problems with the gradient or the column. A highly similar smearing effect is seen in highlight (ii) for an independent sample. Furthermore, throughout the plot, various levels of such smearing (black to red blurring) can be seen along the diagonal, possibly indicating progressive evidence of SCX column degradation.

Worryingly, highlight (iii) shows that there is a clear similarity in detected peptides between the first 'MARS Cysteine' sample and the second 'MARS Non-Cysteine' sample. This contradictory similarity indicates a problem with the peptide selection protocol, as it should ideally result in two fully distinct peptide subsets. On close inspection of the offending overlapping peptides however, we found that the majority are attributed to the various isotypes of immunoglobulin, thus indicating that the selection procedure specifically fails for highly abundant proteins; a not altogether surprising finding since slight deviations from perfect peptide selection efficiency can quickly result in substantial carry-over of undesired peptides for abundant proteins. Of course, high remaining levels of immunoglobins might in turn hint at possible problems with the MARS depletion (which should include affinity binders for the depletion of albumin, transferrin, haptoglobin, IgG, IgA and α-1 antitrypsin) for these samples.

The excellent reproducibility of well-executed separations for identical depletion protocols is show in highlights (iv) and (v), with equivalent SCX fractions yielding largely the same peptide identifications (hence the strongly red off-diagonal lines). Reproducibility across different depletion methods can also be found, albeit at less intensity, in highlights (vi) and (vii) where similar peptides were consistently identified regardless of the protein depletion strategy. Unexpectedly, the peptides identified in 'IgY Non-Cysteine' and 'IgY Non *N*-Glycosylated' samples also show high similarity in highlight (viii). It is unclear why this similarity is so pronounced as it is biologically unlikely that non-cysteinyl peptides are preferentially *N*-glycosylated.

3.2 Proteolytic digestion and precursor mass analysis

Between the steps for protein depletion and peptide selection, proteins are enzymatically cleaved into peptides, typically using the endoproteinase trypsin. However, trypsin digestion is not always completely efficient [25], resulting in missed cleavages and therefore slightly longer than average tryptic peptide populations in most data sets. Two methods are suggested here for quality controlling the digest efficiency: analysis of the distribution of missed cleavages in the resulting peptides and comparison of the distribution of precursor ion masses to those empirically derived from a large set of representative experiments in PRIDE. Note that the former introduces an additional dependence on the search engine used for identification, along with its parameters, while the latter is independent of the search engine. The analyses were performed on a data set comprising 2887 experiments that specifically mentioned trypsin as the only proteolytic enzyme used during sample processing (see Section 2 for details).

In order to derive an empirical background for the frequency of occurrence of missed cleavages, the number of missed cleavages were counted for each peptide identification in an experiment, and the rate of missed cleavage for an experiment was then calculated as the ratio of missed cleavages over the total number of observed cleavages (each peptide terminus is considered a correct tryptic cleavage). The resulting distribution of missed cleavage rate, shown in Fig. 2 then provides a useful empirical background to measure individual experiments against. In Fig. 2 we see two experiments retrieved from PRIDE and their missed cleavage rates, PRIDE accession 12914 shows approximately a 100% efficient digest while experiment 12152 shows a

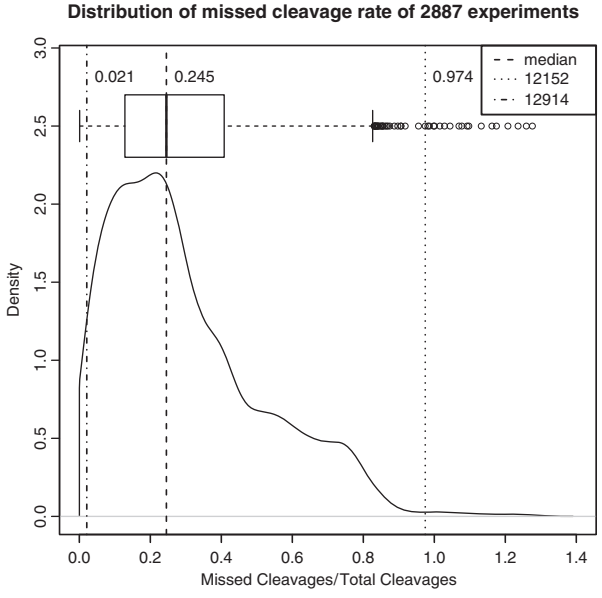

Figure 2. The distribution of missed cleavage rates calculated for 2887 experiments in PRIDE, all annotated as having a single tryptic digestion step. A boxplot reflecting this distribution is shown above the density plot, revealing quite clearly the skewness of the distribution as well as possible outliers (denoted by anything that falls outside the range of the whiskers of the boxplot). It must be noted however that many other methods of outlier detection can be applied to such data, depending on the sensitivity and specificity needs of the end user.

missed cleavage rate of near 1 and hence a 50% of all potential cleavages in the identified peptides were missed. Interestingly, closer inspection of the protocol employed for this data set indicates that free amines had been acetylated, thus excluding lysines from tryptic cleavage. As a result, the high missed cleavage rate is in fact an expected effect of the protocol, and the experiment's status as an outlier in this case indicates an interesting case for data annotation as opposed to a QC issue.

A background distribution can however also be constructed from the precursor masses obtained from the spectra in each experiment. Such a background distribution can also be used in QC as deviations in precursor masses are often indicative of issues with the data. In order to obtain the background distribution, the precursor ion masses per experiment were grouped into 60, 100 Da wide bins and the contents of each bin were then normalised by the total number of peptides. Since the bins are kept constant between the different experiments, we can then create a box plot reflecting the distribution of normalised frequencies across all experiments for each bin. To demonstrate the usefulness of this test, we compared the mass distributions from individual experiments of the HUPO Test Sample Study [19], to the empirical background. These experiments were chosen because the HUPO Test Sample Study was carried out on a single sample consisting of 20 equimolar proteins by over 20 individual laboratories worldwide, as well as a selection of instrument vendors. Furthermore, these 20 proteins were carefully chosen to reflect overall properties of the human proteome as best as possible [19]. An important detail is that individual peptide identifications are not listed for these experiments; only mass spectra and proteins are provided. As such, only an indirect measure such as the precursor mass used here can be employed for such data. Figure 3A correspondingly shows that the experimental precursor mass distribution from PRIDE experiment 8145 closely mimics the empirically derived precursor mass distribution for tryptic digests. As is clear from Fig. 3B however, the average precursor mass from PRIDE experiment 8146 is higher than the empirical distribution, hinting at specific protocols or methods that would produce such deviating data. Figure 3C on the other hand shows a precursor mass distribution from PRIDE experiment 8155 that lies to the left of the expected distri-

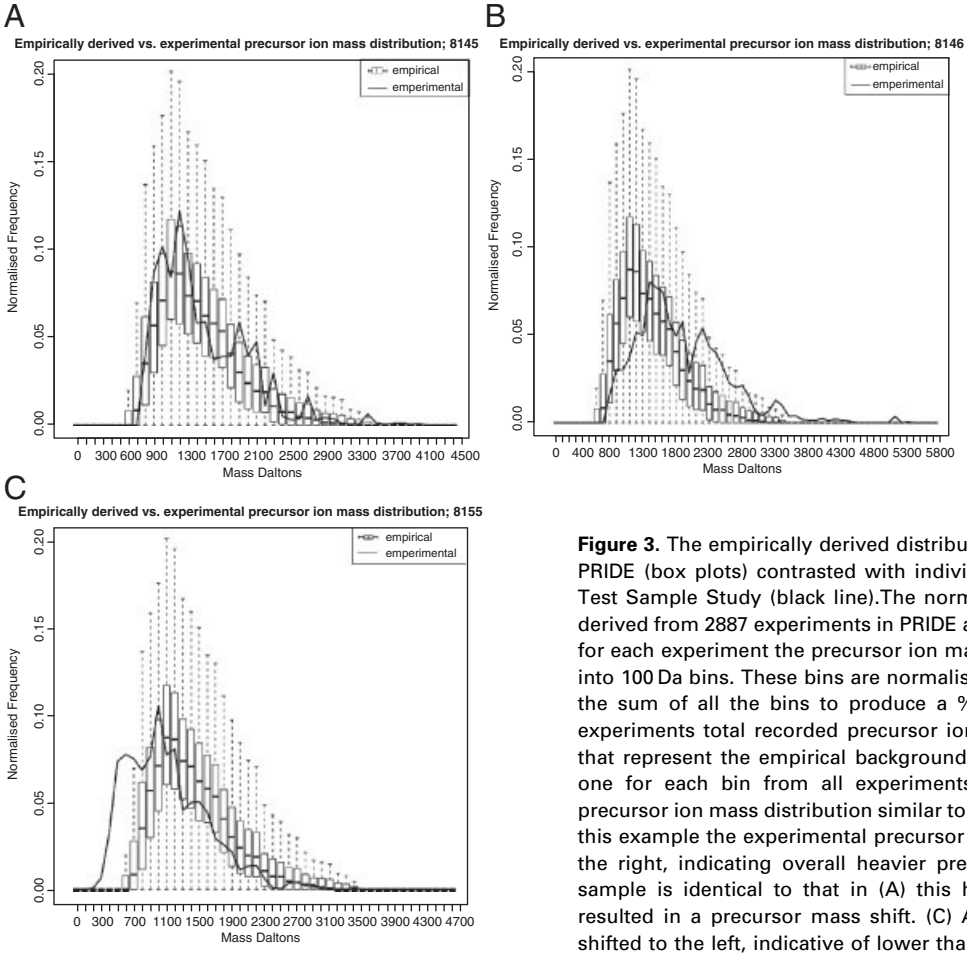

Figure 3. The empirically derived distribution of precursor ion masses from PRIDE (box plots) contrasted with individual experiments from the HUPO Test Sample Study (black line).The normalised empirical background was derived from 2887 experiments in PRIDE annotated as digested with trypsin, for each experiment the precursor ion masses were extracted and stratified into 100 Da bins. These bins are normalised by simply dividing each bin by the sum of all the bins to produce a % contribution of each bin to the experiments total recorded precursor ion mass distribution. The boxplots that represent the empirical background distribution are then constructed one for each bin from all experiments. (A) A typical experiment with precursor ion mass distribution similar to the background distribution. (B) In this example the experimental precursor ion mass distribution is shifted to the right, indicating overall heavier precursors in the sample. Since the sample is identical to that in (A) this hints at a deviating protocol that resulted in a precursor mass shift. (C) A precursor ion mass distribution shifted to the left, indicative of lower than expected precursor masses.

button, indicating that this analysis was particularly sensitive to smaller precursors. Yet, with these deviations in hand, a mechanistic explanation of course remains far off. If such a mechanistic explanation should be required, there is therefore a need for the manual inspection of the data set and/or a thorough reading of the corresponding paper(s).

3.3 MS analysis

The high sensitivity of MS renders it highly susceptible to artefactual contaminants [26] that can enter into the sample during handling, impairing the detection of bona fide peptides in the sample. Additionally, peptide fragmentation is not always optimally efficient, sometimes yielding too few or too small fragments to be useful for identification. In order to measure the amount of non-informative MS2 spectra recorded during an analysis, we examined the distribution of the mass differences between peaks in each MS2 spectrum in an experiment. In order to ensure analysis of the most significant signals, the MS2 spectra were first filtered to retain only the top 10% most intense peaks. *M/Z* difference distance matrices were computed from the filtered peak lists; these matrices were then combined resulting in a distribution of *m/z* differences and their frequency. This was then plotted as a histogram where a single bar represents a 1 *m/z* difference between two peaks, the region 40–200 *m/z* was decided to be the most useful region of *m/z* difference distribution as it encompasses the masses of all amino acids and several common contaminants. Figure 4A shows a typical high-quality example for such a distribution, with the *m/z* differences corresponding to amino acid residue masses clearly rising well above the general noise level in the histogram. Figure 4B on the other hand shows a distribution that provides a less favourable picture; the *m/z* differences corresponding to amino acid residue masses lie well within the noise, and the extremely high peak at 44 Da, corresponding to the mass of a PEG monomer building block, a common contaminant in MS. The MS2 mass difference distribution is in fact one of the best ways to quickly detect PEG contamination levels, since the actual polymers take a variety of precursor masses depending on the number of monomers they are composed of, but the MS2 spectra will always reveal the steady train of 44 Da differences. Interestingly, each experiment can now also be considered a multidimensional vector, with each dimension corresponding to a bar in the bar plot. Since amino acid occurrence rates are species-specific, we should be able to spot consistent patterns in experiments that are derived from the same species, and contrast these with the patterns obtained from experiments from another species. This is illustrated in Supporting Information Fig. 1, which depicts the two-dimensional projection of the experiment vectors through multi-dimensional scaling. It is clear that specific species do follow specific distributions, but that the separation in two dimensions lacks sufficient resolving

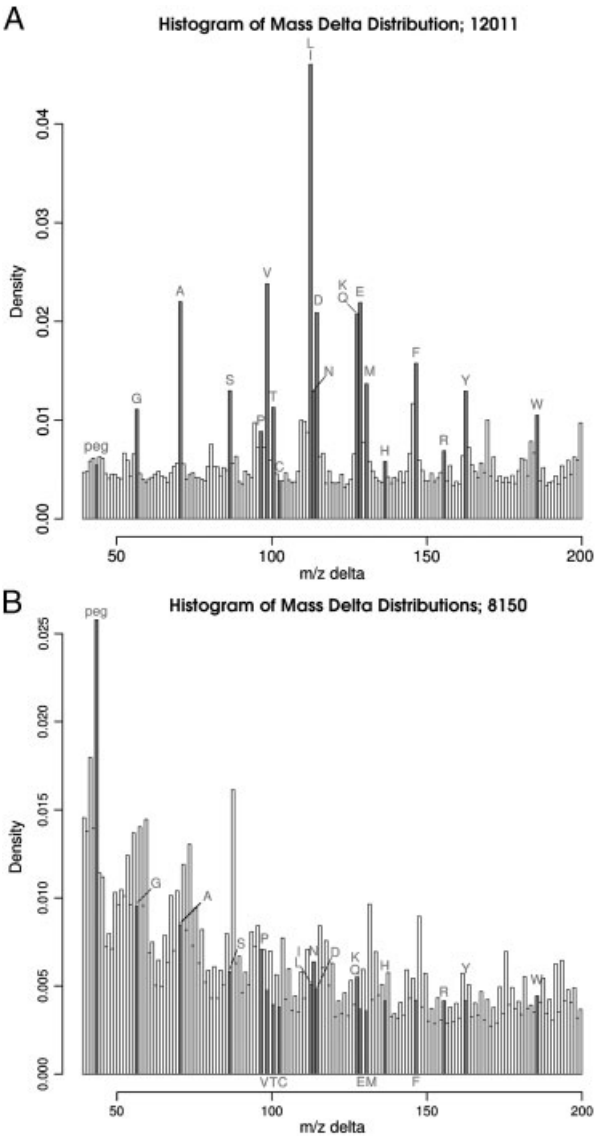

Figure 4. Histograms of the distribution of distances between the top 10% most intense peaks for each MS2 spectrum in an experiment. Note that the *Y*-axis scale is given as relative density units and that the absolute scale is therefore not comparable between plots. (A) An example of a good mass difference distribution, with the differences primarily representing the amino acid residue masses. (B) A clear example of PEG contamination in the sample, with most intense peaks in the sample derived from PEG fragmentation at the cost of actual peptide fragmentation spectra.

power to clearly distinguish the species origin of any single experiment. Regardless, such approaches can be very useful in the a posteriori annotation of experiments and in the curatorial detection of possible misannotations.

The distribution of the *m/z* of the MS2 peaks by itself can also provide useful information, somewhat similar to analysing the precursor ion mass distribution as discussed previously. Typically, the distribution of MS2 *m/z* peaks is

roughly normal, performing this analysis across the portion of PRIDE experiments that contains MS2 spectra yielded unsurprising results in the majority of cases, but two experiments submitted sequentially by a single lab (PRIDE accession numbers 8927 and 8928) prove to be outliers. Figures 5A and 5B shows the MS2 *m/z* distribution for experiments 8927 and 8927, respectively, displaying an unusual truncated bimodal distribution. While the truncation is most likely the product of mass limits imposed by the analyzer, the bimodal character is unexpected. To check this was not a legitimate feature of the submitter's data we investigated their other 150 submissions and plotted a reference distribution of MS2 *m/z* to compare against in Fig. 5C; this clearly shows that these two experiments are cause for concern. After discussion with the data submitter the raw data was analysed and the instrument records cross referenced for potential causes. The cause of this phenomenon was thus pinned down to a mis-calibration of the mass spectrometer's analyser due to a contamination of the transfer capillary, which resulted in overcharging effects.

3.4 Quantitative analysis

Quantification of proteins through peptides as surrogates has become increasingly popular over the last few years. In labelled proteomics for instance, tagging systems like iTRAQ [27] or TMTs [28] are routinely used to quantify peptides from their MS2 spectra. These tagging systems are multiplexed sets of isotope tags that are used to label all peptides generated from tryptic digestion. Since the tags are isobaric, differentially labelled versions of a peptide appear as a single precursor ion in MS mode. When labelled peptides are subjected to collision-induced dissociation, the tags release diagnostic, low-mass reporter ions that are used for quantification. With the increase in popularity of these and other quantitative methods, it is crucial to develop corresponding QC methodologies and metrics. Because the storage of quantitative proteomics data in publicly available proteomics data repositories currently lags behind their ability to store data and results from more traditional proteomics strategies, there is very little consistent data to be

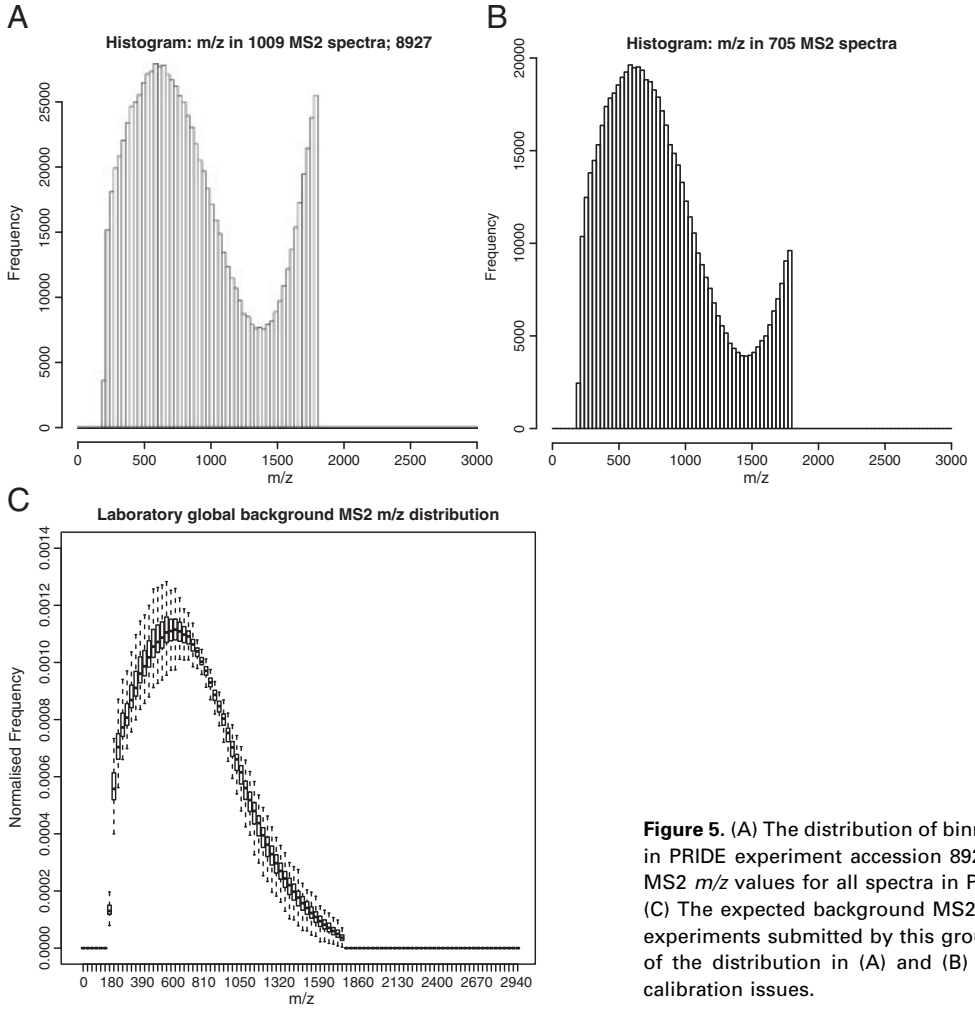

Figure 5. (A) The distribution of binned MS2 *m/z* values for all spectra in PRIDE experiment accession 8927. (B) The distribution of binned MS2 *m/z* values for all spectra in PRIDE experiment accession 8928. (C) The expected background MS2 *m/z* distribution of the other 150 experiments submitted by this group. The truncated bimodal nature of the distribution in (A) and (B) is unexpected, and indicative of calibration issues.

found in the public domain. As such, comparative QC on quantitative proteomics data remains a little further in the future, but independent QC checks can already be designed and applied with an aim to calculating them across multiple data sets to generate reference metrics when more data become available. For instance, by comparing the measured ratios with the expected ratios, the accuracy of the (relative) quantification can be determined.

Figure 6 shows an attempt to check the quality of quantification by chemical labelling on a complex sample, by comparing six identical human serum digests at known relative concentrations. The labels used in this case are TMTs. In Fig. 6A a distribution of the ratios of the peak intensities at 126 to 130 Da with respect to the peak intensity at 131 Da is shown as boxplots. On the left, data from identified spectra only was taken whereas the boxplots on the right incorporate unidentified spectra also. These are centered on $y = -2$, $y = -1$ and $y = 0$ on a log3 scale as expected from the test sample's design, indicating good accuracy, even on the very heterogeneous data provided by the sum on unidentified and identified spectra. The range of the boxplots gives the precision, in this test sample corresponding to technical reproducibility. One must however be careful to translate this 'ideal case' to the more common situation where only a few components are regulated. In that case, imperfect selection of the precursor peptide leads to overlapping TMT peaks and incorrect calculation of the regulation. Especially for low-abundance proteins, the effect will be an underestimation of the regulation. However, simultaneous selection of two or more precursor peptides is detectable from the MS2 analysis of the fragmentation spectra [29], providing a means to identify such co-fragmentation events.

In Fig. 6B we see the histogram of the number of missing values at the m/z positions of the six TMT reporter ions. The vast majority of spectra contain all six reporter ion peaks, indicating overall efficient labelling. In only 0.8% of spectra, all six peaks are missing, either because the corresponding peptide is of low concentration or because of incomplete labelling or fragmentation. It may also be that the analyte is actually not a peptide but another type of charged molecule with a modifiable free amine. This cannot be verified, as the analysis is done on all spectra, and generally we find that only a minority fraction (10–50%) of the MS2 spectra leads to an acceptable peptide identification. Limiting ourselves to spectra leading to accepted peptide identifications for this data set however leads to very similar results. Although the majority of the spectra is not identifiable in the first pass of a MASCOT search, a large fraction of these can be mapped on peptides from proteins from the first search, using MASCOT'S error-tolerant search, allowing different modifications, missed and non-tryptic cleavages (certainly likely due to the presence of proteases in serum [30]). The more recent Protein Pilot software from Life Science typically allows the identification of more than 80% of the spectra in the first pass. In Fig. 6C a histogram of the number of missing values for each of the six TMT reporter ions separately reveals that missing peaks are much more common amongst the masses 126 and 127, which are the samples with $9\times$ lower concentration, showing that abundance is an important factor in the detectability of reporter ions.

Figure 7 shows the intensity of the different TMT reporter ion peaks against the average intensity of the top 10% most intense non-TMT peaks (I_{A10}). A single dot on the charts represents a single MS2 spectrum, and axes have been drawn on log10 scale. This analysis shows whether a good balance has been found between quantification and identification. This certainly seems to be the case here, as the reporter ions have an intensity that is clearly correlated to the I_{A10} metric. It is also clear that the ratio between reporter ions to the I_{A10} is quite constant across a broad intensity range, revealing that reporters are good surrogates for peptide quantification. At the same time the 1/100 (Figs. 7A and B) to 1/10 (Figs. 7E and F) ratios indicate that peptide identification (typically based on the most intense peaks in the spectrum) has not been overshadowed by the reporter ion peaks at all, illustrating the absence of overly competitive ionisation. However, at the same time, the actual ratios of the different reporters versus the I_{A10} correlate very well with the sample composition: roughly 1/100 for 126 and 127, about 1/33 for 128 and 129, and a little under 1/10 for 130 and 131, corresponding to the 1:1:3:3:9:9 mixing ratio for these reporters, respectively. The TMT peaks are thus well suited to (relative) quantification purposes.

4 Discussion

As public data repositories are getting increasingly populated with (published) proteomics data sets, large-scale data analysis becomes an ever more powerful tool for investigating and predicting the nuances implicit in proteomics methods and results. Yet, the reliability of many downstream public data processing methods is crucially dependent on the validity of the data. Hence, it is increasingly important to have properly matched, empirically derived reference metrics available for selecting and filtering the available data sets. This rings true for both computational users as well as for database curators, since both have a vested interest in detecting, and possibly understanding, outlying data sets. The relative sparseness of the information currently deposited in public repositories compared to what is available in the lab during and immediately after sample processing and data acquisition requires the development of robust metrics that can be derived from the available data, preferentially as close to the acquisition point as possible.

Correspondingly, we have illustrated here that publicly available data, spanning many individual experiments of diverse origin, can be a posteriori examined according to several easily and reliably obtainable metrics for a typical

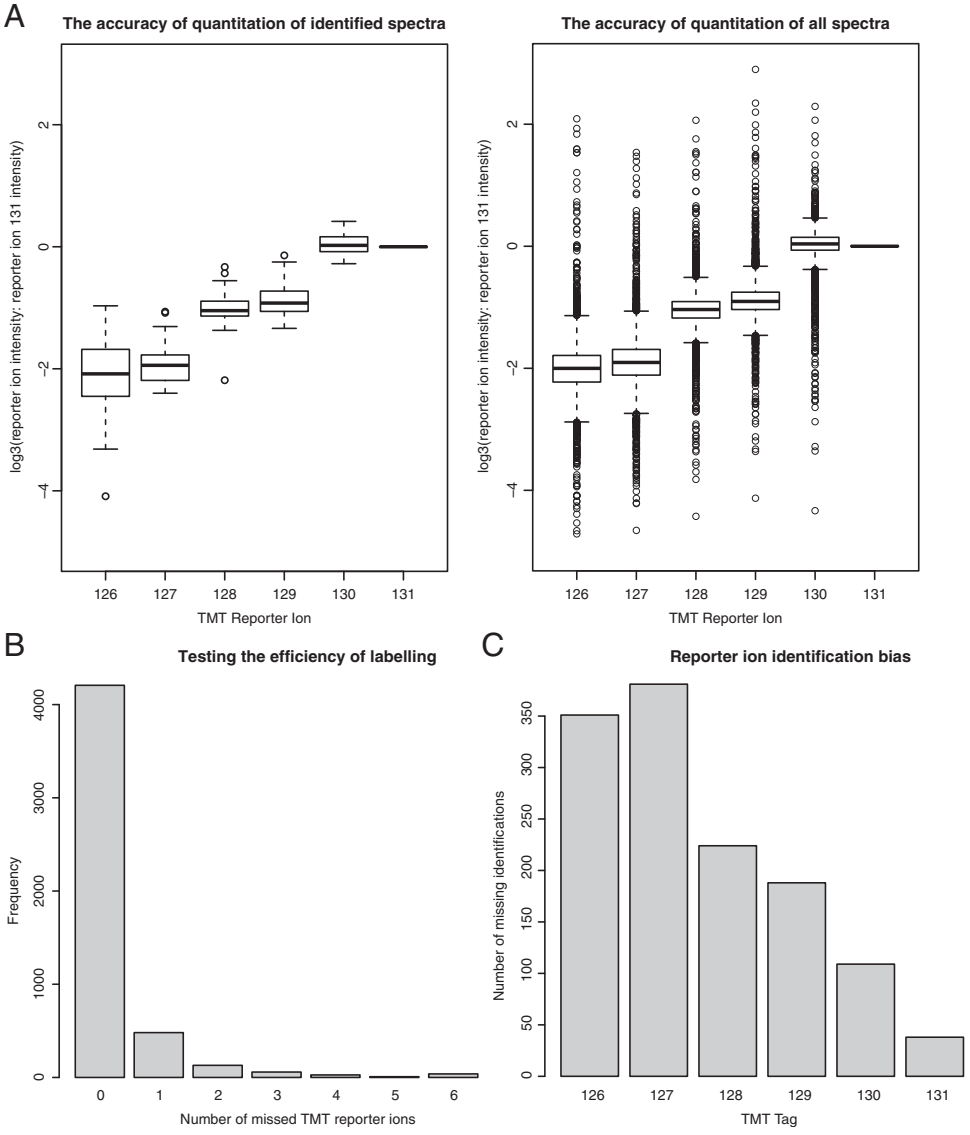

Figure 6. A selection of quantitative QC measures on a TMT data set. (A) Boxplots showing the distribution of the ratios of the peak intensities at 126, 127, 128, 129 and 130 Da with respect to the peak intensity of the 131 Da reporter ion. On the left, only data from identified spectra is shown, while the right plot contains all spectra (identified and unidentified). Interestingly, the boxplots do not change shape despite the presence of many (strong) outliers for the data from all spectra, thus validating the use of these robust metrics for highly heterogeneous data. From the sample design, ratios should be in a 1:3:9 ratio for the three groups, respectively. The distribution centres fall as expected (i.e. are accurate), with the width of the boxplot representing (in this case technical) reproducibility. (B) Histogram of the number of missing values at the positions of the six TMT reporter ions. The majority of spectra recover all six expected TMT peaks, but in some cases none are recovered at all, possibly due to poor labelling, poor fragmentation or simply low concentration peptides and ion suppression effects. (C) Histogram of the number of missing values for each of the six TMT reporter ions separately; 126 and 127 are most often missed but this is expected because the sample bearing these tags was 3 $\times$ more dilute than the one carrying the 130 and 131 labels.

proteomics workflow. This includes analyses performed within the context of a larger study, as was shown for the protein depletion, peptide selection and SCX separation procedures used in the HUPO PPP2 data set, but also extends across individual studies, where a large body of only very loosely related experimental data (e.g. selected based solely on the enzyme used for proteolytic digestion) can be used to estimate empirical background distributions complete with tolerance ranges. Furthermore, the obtained metrics can even be used in surprising ways to compare experimental metadata annotations, as was illustrated for the m/z differences between MS2 peaks. The latter point is more than a gimmick: missing or incorrect annotation constitutes a serious downstream problem for data consumers, and the ability to detect possible misannotation or to assign annotation where none is given will likely be an

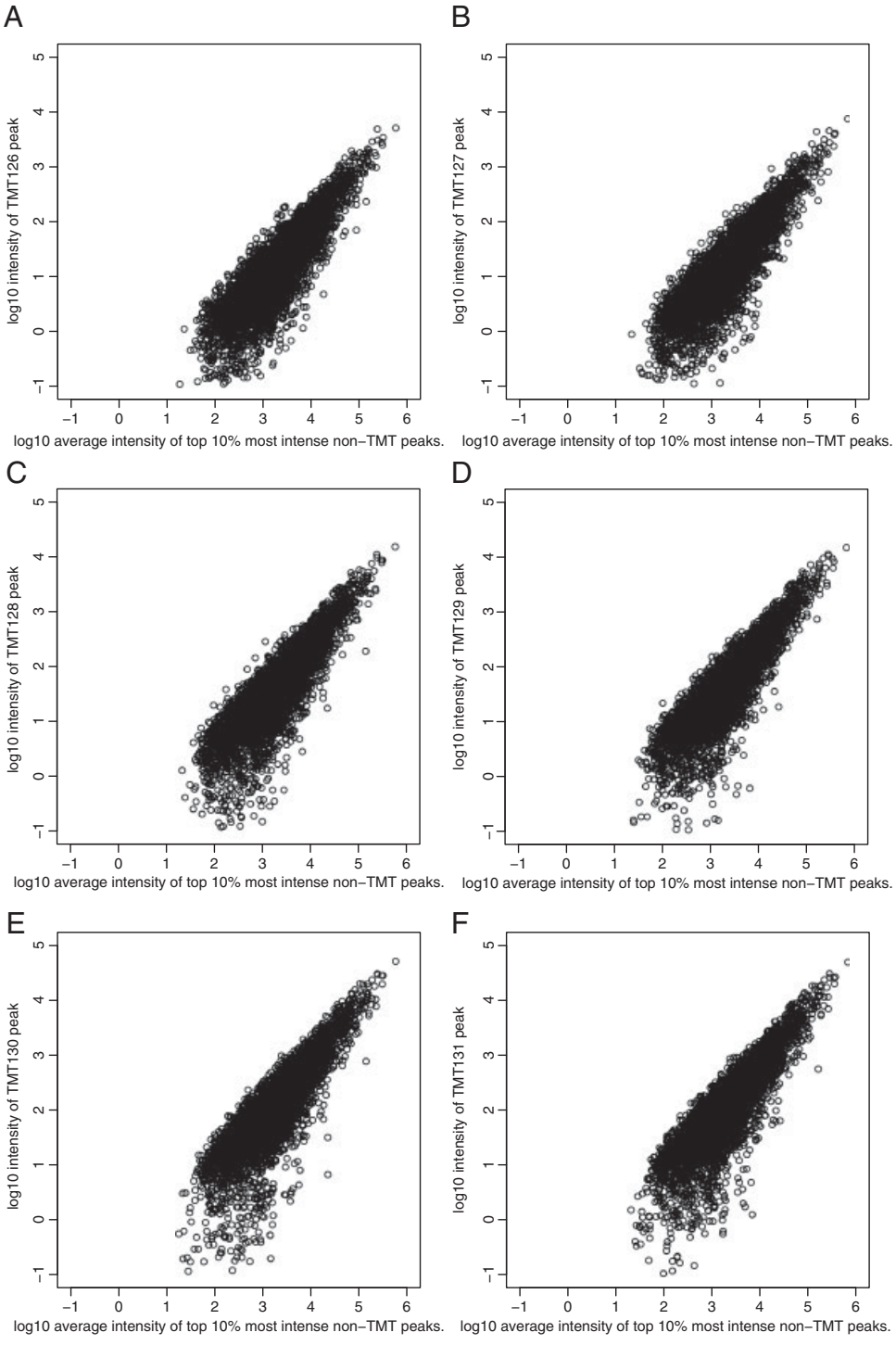

Figure 7. The intensity of the TMT reporter peaks versus the average peak intensity of the top 10% non-reporter ion peaks in each spectrum. TMT reporters shown: (A) 126 Da; (B) 127 Da; (C) 128 Da; (D) 129 Da; (E) 130 Da; (F) 131 Da. These plots relate quantification to identification, since the reporter ions have correlated intensities to the top 10% non-reporter ions that are most often used for identification purposes. The TMT reporters thus prove to be adequate estimators of quantity. Since the reporter ions are always at least an order of magnitude less intense than fragment ions, an overshadowing effect by reporter ions, reducing peptide identifications is unlikely.

important curatorial function for repositories in the years to come. The mass spectra can be mined for additional data as well: mass distributions of MS1 and MS2 ions can relate important information about biases or faults in the protocol, while the isobaric labelling of peptides for MS2-based quantification approaches can be inspected easily as well. In the latter case, the possible trade-off between quantification and identification efficiency can be monitored by comparing the reporter ion intensities against the average intensity of the top 10% most-intense non-reporter peaks.

With proteomics coming ever more into the limelight of the life sciences as a powerful and sensitive analytical platform, the need for robust QC practices is becoming ever more pressing. Such QC must in most cases also extend

beyond a single MS analysis, necessarily encompassing several runs within or even across experiments or studies. As a result, metrics need to be obtained that can function at the level of the individual run, but also across many runs. The latter can directly benefit from already acquired data for the establishment of acceptance criteria. These latter criteria are of course open to interpretation, and will depend on the downstream use case for those data. The future of QC in proteomics therefore is set to go hand in hand with that of data repositories and the standardized deposition of well-annotated data sets.

J.M.F. is supported by a BBSRC CASE studentship funded by the BBSRC and Philips. M.V. and J.M.F. thank Rolf Apweiler for his support. L.M. would like to thank Joël Vandekerckhove for his support.

The authors have declared no conflict of interest.

5 References

[1] Service, R. F., Proteomics. Proteomics ponders prime time. *Science* 2008, *321*, 1758–1761.

[2] Omenn, G. S., States, D. J., Adamski, M., Blackwell, T. W. et al., Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics* 2005, *5*, 3226–3245.

[3] Choudhary, C., Mann, M., Decoding signalling networks by mass spectrometry-based proteomics. *Nat. Rev. Mol. Cell Biol.* 2010 *11*, 427–439.

[4] Przybylski, C., Jünger, M. A., Aubertin, J., Radvanyi, F. et al., Quantitative analysis of protein complex constituents and their phosphorylation states on a LTQ-orbitrap instrument. *J. Proteome Res.* 2010, *9*, 5118–5132.

[5] Rifai, N., Gillette, M. A., Carr, S. A., Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat. Biotechnol.* 2006, *24*, 971–983.

[6] Craig, R., Cortens, J. P., Beavis, R. C., Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.* 2004, *3*, 1234–1242.

[7] Desiere, F., Deutsch, E. W., Nesvizhskii, A. I., Mallick, P. et al., Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.* 2005, *6*, R9.

[8] Martens, L., Hermjakob, H., Jones, P., Adamski, M. et al., PRIDE: the proteomics identifications database. *Proteomics* 2005, *5*, 3537–3545.

[9] Slotta, D. J., Barrett, T., Edgar, R., NCBI Peptidome: a new public repository for mass spectrometry peptide identifications. *Nat. Biotechnol.* 2009, *27*, 600–601.

[10] Klie, S., Martens, L., Vizcaíno, J. A., Côté, R. et al., Analyzing large-scale proteomics projects with latent semantic indexing. *J. Proteome Res.* 2008, *7*, 182–191.

[11] Mueller, M., Vizcaíno, J. A., Jones, P., Côté, R. et al., Analysis of the experimental detection of central nervous system-related genes in human brain and cerebrospinal fluid datasets. *Proteomics* 2008, *8*, 1138–1148.

[12] Mallick, P., Schirle, M., Chen, S. S., Flory, M. R. et al., Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.* 2007, *25*, 125–131.

[13] Sherwood, C. A., Eastham, A., Lee, L. W., Peterson, A. et al., MaRiMba: a software application for spectral library-based MRM transition list assembly. *J. Proteome Res.* 2009, *8*, 4396–4405.

[14] Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K. et al., Building consensus spectral libraries for peptide identification in proteomics. *Nat. Methods* 2008, *5*, 873–875.

[15] Tabb, D. L., Vega-Montoto, L., Rudnick, P. A., Variyath, A. M. et al., Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J. Proteome Res.* 2010, *9*, 761–776.

[16] Rudnick, P. A., Clauser, K. R., Kilpatrick, L. E., Tchekhovskoi, D. V. et al., Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses. *Mol. Cell. Proteomics* 2010, *9*, 225–241.

[17] Qian, W. J., Kaleta, D. T., Petritis, B. O., Jiang, H. et al., Enhanced detection of low abundance human plasma proteins using a tandem IgY12-SuperMix immunoaffinity separation strategy. *Mol. Cell. Proteomics* 2008, *7*, 1963–1973.

[18] Liu, T., Qian, W. J., Gritsenko, M. A., Xiao, W. et al., Inflammation and the host response to injury large scale collaborative research programm. High dynamic range characterization of the trauma patient plasma proteome. *Mol. Cell. Proteomics* 2006, *5*, 1899–1913.

[19] Bell, A. W., Deutsch, E. W., Au, C. E., Kearney, R. E. et al., HUPO Test Sample Working Group. A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. *Nat. Methods* 2009, *6*, 423–430.

[20] Cox, J., Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* 2008, *26*, 1367–1372.

[21] Chelius, D., Bondarenko, P. V., Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. *J. Proteome Res.* 2002, *1*, 317–323.

[22] Bondarenko, P. V., Chelius, D., Shaler, T. A., Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. *Anal. Chem.* 2002, *74*, 4741–4749.

[23] Schiess, R., Wollscheid, B., Aebersold, R., Targeted proteomic strategy for clinical biomarker discovery. *Mol. Oncol.* 2009, *3*, 33–44.

[24] Vizcaíno, J. A., Côté, R., Reisinger, F., Foster, J. M. et al., A guide to the Proteomics Identifications Database proteomics data repository. *Proteomics* 2009, *9*, 4276–4283.

[25] Yang, H. J., Hong, J., Lee, S., Shin, S. et al., Pressure-assisted tryptic digestion using a syringe. *Rapid Commun. Mass Spectrom.* 2010, *24*, 901–908.

[26] Keller, B. O., Sui, J., Young, A. B., Whittal, R. M., Interferences and contaminants encountered in modern mass spectrometry. *Anal. Chim. Acta* 2008, *627*, 71–81.

[27] Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B. et al., Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* 2004, *3*, 1154–1169.

[28] Thompson, A., Schäfer, J., Kuhn, K., Kienle, S. et al., Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* 2003, *75*, 1895–1904.

[29] Houel, S., Abernathy, R., Renganathan, K., Meyer-Arendt, K. et al., Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *J. Proteome Res.* 2010, *9*, 4152–4160.

[30] Yi, J., Kim, C., Gelfand, C. A., Inhibition of intrinsic proteolytic activities moderates preanalytical variability and instability of human plasma. *J. Proteome Res.* 2007, *6*, 1768–1781.

