Comparing the diagnostic classification accuracy of iTRAQ, peak-area, spectral-counting and emPAI methods for relative quantification in expression proteomics

Adam A. Dowle1\*, Julie Wilson2 and Jerry R. Thomas1

1. Bioscience Technology Facility, Department of Biology, University of York, York, UK

2. Departments of Mathematics and Chemistry, University of York, York, UK

ABSTRACT: Diagnostic classification accuracy is critical in expression proteomics to ensure that as many true differences as possible are identified with acceptable false positive rates. We present a comparison of the diagnostic accuracy of iTRAQ with three label-free methods – peak-area, spectral-counting, and emPAI – for relative quantification using a spiked proteome standard. We provide the first validation of emPAI for inter-sample relative quantification, and find clear differences among the four quantification approaches that could be considered when designing an experiment. Spectral counting was observed to perform surprisingly well in all regards. Peak area performed best for smaller fold-differences and was shown to be capable of discerning a 1.1-fold difference with acceptable specificity and sensitivity. The performance of iTRAQ was dramatically worse than the label-free methods with low abundance proteins. Using the iTRAQ data set for validation, we also demonstrate a novel iTRAQ analysis regime that avoids the use of ratios in significance testing and outperforms a common commercial alternative. KEYWORDS: *iTRAQ, peak area, spectral counting, emPAI, label free, relative quantification, expression proteomics.*

INTRODUCTION

In 1999, when LC-MS solutions for proteomic quantification were still in their infancy1, Blackstock and Weir defined the term “expression proteomics”.2 However, untargeted absolute quantification of protein expression still remains beyond the reach of most laboratories today.3,4 While approaches for untargeted absolute quantification continue to be investigated,5-7 relative quantification has become more amenable and commonplace in both research and service laboratories.8 Thankfully, relative quantification is well placed to answer the most fundamental of proteomics questions that is, which proteins are differently expressed?9

LC-MS-based relative quantification can be split into two categories: isotopic labelling and label free. In the former, different stable isotopes are introduced among samples either *in vivo* or *vitro*.10,11. Isotopically labelled samples are combined prior to LC-MS acquisition, with the imparted mass differences providing the link from mass spectra to sample provenance. Although isotopically different, labelled peptides are otherwise chemically identical, ensuring equal response factors during MS analysis. Measurement of relative MS intensity between isotopically labelled peptides is therefore an accepted measure of relative peptide amount.12 *In vivo* labelling strategies such as SILAC13 and 15N metabolic labelling14 potentially offer the lowest levels of technical variance15 as samples are combined at the earliest opportunity but have reduced applicability compared to *in vitro* approaches.16

A variety of *in vitro* labelling techniques are available including methods employing well characterized, low cost chemical reactions17-19 and more elaborate commercial approaches facilitating higher levels of multiplexing.1, 20-26 Two of the most frequently used *in vitro* labels are SCIEX’s isobaric tags for relative and absolute quantification (iTRAQ)21, 22 and Thermo’s tandem mass tags (TMT)23-26 each of which add isobaric tags to free amines at peptide N-termini and lysine side chains. Precursor ion spectra of differently labelled samples appear identical but unique *m/z* reporter ions are released upon fragmentation enabling relative quantification.11 It has been demonstrated that instrument parameters should be tuned in order to achieve maximum performance with isobarically labelled samples.27 Specifically, increasing fragmentation energies and tuning toward lower mass can aid observation of reporter ions and in turn quantitative accuracy but at the cost of sequence ions crucial to peptide identification.27 iTRAQ is available as either 4- or 8-plex21, 22 and TMT is offered as 6- or 10-plex versions.24, 26 Reporter ions are unit mass separated with the exception of TMT 10-plex, which requires mass resolution greater than 60,000 FWHM at low *m/z*, effectively limiting usage to Fourier transform mass analyzers.26

Label free quantification does not combine samples before LC-MS acquisition, instead unmodified peptide mixtures are acquired successively and relative responses extracted from the resulting datasets.28 Response can be measured directly, from precursor and/or product ion chromatogram peak areas29 or indirectly using spectral counting.30 Label free quantification is performed typically with multiple LC-MS acquisitions to generate sufficiently powered statistics for precise measurement of differential expression.31 Peak area-based quantification necessitates sophisticated software to align LC-MS chromatograms between runs and overlay peptide identifications.32 Acquisition parameters must also be tuned to deliver sufficient frequency and consistency of measurement across the peak widths of eluting peptides. By comparison, it could be argued that spectral counting approaches are the easiest to implement into established LC-MS identification-based work-flows. Spectral counting has been shown to be compatible with routine data dependent acquisition modes using dynamic exclusion33,34 and counts can be obtained from identification results without the need for additional software.35 Studies have demonstrated that normalization of spectral counting data can further improve measurement accuracy for both intra- and inter-sample relative protein quantification.36-38 Yet, spectral counting approaches are more abstracted from the raw data and often reported as semi-quantitative in the absence of more accepted approaches.39-41 It has been noted that emPAI quantification, a modified form of spectral counting designed for inter-protein, intra-sample, quantification42 is used as a form of semi-quantitative,43-45 relative inter-sample quantification with little evidence to demonstrate its performance.

With a variety of untargeted relative quantification techniques available it is logical that their performance be compared. Of particular importance is assessment of diagnostic accuracy – the veracity of a technique when classifying an outcome as positive or negative, both in respect of maximizing true positives and negatives, while minimizing false positives and negatives. For the purposes of a typical expression proteomics experiment the diagnostic classification is whether a protein is differentially expressed (positive outcome) or unchanged (negative outcome).

Previous comparisons between isobaric labelling and label free methodologies have been reported using biological systems, focusing on a variety of metrics including: number of quantifications,46-51 coefficient of variation,48, 49 Pearson’s coefficient,47, 49, 50 similarity to transcriptomic data46 and alignment with predicted or previously reported biological changes.51 While the studies provide valuable insight, expected fold differences were either unknown, estimated or zero, meaning diagnostic classification and measurement accuracy can only be extrapolated from precision.

An alternative is to mimic a biological system by adding exogenous proteins into a constant, complex background, such that the expected fold difference between samples is known for all proteins. Wang has presented a comparison of iTRAQ and peak area-based label free by measuring four exogenous proteins added at ratios from 1-10 fold difference into a quantitative analysis of two *Chlamydomonas reinhardtii* strains.52 Measurement accuracy is scrutinized but with exogenous proteins added among technical replicates of different biological samples no attempt is made to measure classification accuracy. Sandberg used a combination the Universal Proteomics Standard mix (UPS2) and an additional nine proteins in a human cell line. The comparison between iTRAQ 8-plex, TMT 6-plex and peak area-based label free quantification covered a wide range of fold differences and absolute abundances.53 The paper provides excellent insight into measurement accuracy, linearity and absolute limit of identification/quantification but no examination of classification accuracy or fold-difference limit. Sjödin compared iTRAQ, dimethyl labelling and peak area-based relative quantification by spiking 11 bovine proteins into an *E. coli* lysate background.54 The study assesses measurement accuracy and false quantification rate independently with no common assessment of diagnostic classification performance. Positives were accepted as true where the 95% confidence interval for the measured fold difference contains the expected fold-difference. True negatives were classified where the 95% C.I. spanned an arbitrary threshold of 1.2- or 1.5-fold. In combination, such criteria give the possibility for a value to be simultaneously a positive and a negative if the 95% C.I. encompasses both the expected fold difference and the threshold for no change. The criteria also have the potential to skew acceptance of true positives in favor of lower precision measurements with greater C.I.s. There is much value in Sjödin’s study, in particular that the false positive rate improves with increasing number of peptide measurements but the work leaves scope for further expansion.

In a study by Ramus *et al.*, exogenous UPS2 proteins were spiked into a common yeast background in a comparison of label free approaches.55 Predictive performance was determined from receiver operator characteristic curves, but quantification of exogenous proteins was restricted to 2-fold difference and above.

No study has fully assessed the comparative diagnostic performance of the most commonly used relative quantification techniques, especially in terms of diagnostic classification accuracy at low magnitude fold-differences. Furthermore, there is little comparison between isobaric labelling and spectral counting approaches and no evidence to validate the use of emPAI for inter-sample relative quantification. By expanding upon and combining Sjödin’s experimental design54 with the performance analysis of Ramus,55 we present a comparison of 8-plex iTRAQ with peak-area, spectral-counting and emPAI methodologies for relative quantification. We assess diagnostic accuracy using a spiked standard with fold differences ranging from 1.1-fold to ∞ in a complex background, with differences measured over two orders of magnitude of absolute abundance. The aim of the study is to compare diagnostic classification performance, measurement accuracy and the limit of fold-difference so that researchers are better placed to choose between the techniques.

Previous studies have shown that increased statistical manipulation can improve quantitative accuracy, such as Lou's use of Bayesian analysis to improve iTRAQ measurement accuracy.56 However, there is no universally accepted data analysis pipeline or statistical regime for quantitative proteomics data.57-60 The datasets we produced in this study can provide a valuable resource for testing software and data analysis schema. We use the data to test a novel iTRAQ data analysis strategy that outperforms a common commercial alternative.

MATERIALS AND METHODS

Digestion

Commercially sourced proteins (Sigma) were weighed from solid and diluted to 1 g/L in aqueous 0.5 M triethylammonium bicarbonate (TEAB) before taking 25 g for digestion. A total of 18 exogenous proteins were digested as detailed in table 1. *E. coli* lysate (Bio-Rad) was diluted to 10 g/L in aqueous 0.5 M TEAB before taking 2 mg for digestion. Protein solutions were digested with sequencing grade trypsin (Promega) at 1:50 enzyme to substrate ratio (*m:m*), post reduction with 50 mM tris(2-carboxyethyl)phosphine and alkylation with 200 mM methyl methanethiosulfonate (MMTS). Digests were incubated overnight at 37oC.

**Table 1. Exogenous proteins digested**

|  |  |  |
| --- | --- | --- |
| **Reference** | **Accession** | **Protein** |
| A | P00711 | Alpha-lactalbumin |
| B | P00698 | Lysozyme C |
| C | P02470 | Alpha-crystallin A chain |
| D | P02754 | Beta-lactoglobulin |
| E | P61823 | Ribonuclease pancreatic |
| F | P68082 | Myoglobin |
| G | P02662 | Alpha-S1-casein |
| H | P02666 | Beta-casein |
| I | P00125 | Cytochrome c1 |
| J | P12763 | Alpha-2-HS-glycoprotein |
| K | P01012 | Ovalbumin |
| L | P00924 | Enolase |
| M | P11412 | Glucose-6-phosphate dehydrogenase |
| N | P00432 | Catalase |
| O | P10594 | Invertase |
| P | Q29443 | Serotransferrin |
| Q | P02769 | Serum albumin |
| R | P00489 | Glycogen phosphorylase |

Construction of quantification standards

Four quantification standards (*Fine 1*, *Fine 2*, *Gross 1* and *Gross 2*) were created by adding exogenous protein digests to aliquots of a common *E. coli* digest as detailed in table 2. The design provided known differences over a range of ratios and absolute abundances. Comparing *Fine 1* and *Fine 2* provided exogenous protein ratios at: 1:1.1, 1:1.2, 1:1.3, 1:1.4, 1:1.5 and 1:1.75. Comparing *Gross 1* and *Gross 2* gave exogenous protein ratios of 1:1.5, 1:2, 1:5, 1:10, 1:100 and 1:∞. Each ratio was constructed at 100, 10 and 1 fmol/100 ng *E.coli* digest, using a different exogenous protein. A total of 33 unique fold-difference and abundance combinations were generated while maintaining equimolar total protein amount and a common unchanging background among all standards.

**Table 2. Construction of quantification standards**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Fine 1*** | | | | | | |
| **Ratio** | **1.75↓** | **1.5↑** | **1.4↑** | **1.3↓** | **1.2↑** | **↓1.1** |
| fmol  /100 ng *E. coli* | 72.7 (B) | 120 (H) | 117 (O) | 87.0 (C) | 109 (J) | 95.2 (N) |
| 7.27 (R) | 12.0 (A) | 11.7 (G) | 8.70 (M) | 10.9 (E) | 9.52 (F) |
| 0.73 (K) | 1.20 (Q) | 1.17 (D) | 0.87 (L) | 1.09 (P) | 0.95 (I) |
| ***Fine 2*** | | | | | | |
| **Ratio** | **1.75↑** | **1.5↓** | **1.4↓** | **1.3↑** | **1.2↓** | **1.1↑** |
| fmol  /100 ng *E. coli* | 127 (B) | 80.0 (H) | 83.4 (O) | 113 (C) | 91.0 (J) | 105 (N) |
| 12.7 (R) | 8.00 (A) | 8.34 (G) | 11.3 (M) | 9.10 (E) | 10.5 (F) |
| 1.27 (K) | 0.8 (Q) | 0.83 (D) | 1.13 (L) | 0.91 (P) | 1.05 (I) |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Gross 1*** | | | | | | |
| **Ratio** | **∞↓** | **1.5↑** | **100↑** | **10↑** | **5↓** | **2↓** |
| fmol  /100 ng *E. coli* | 0 (B) | 120 (H) | 198.02 (O) | 182 (C) | 33.3 (J) | 66.7 (N) |
| 0 (R) | 12.0 (A) | 19.8 (G) | 18.2 (M) | 3.33 (E) | 6.67 (F) |
| 0 (K) | 1.20 (Q) | 1.98 (D) | 1.82 (L) | 0.33 (P) | 0.67 (I) |
| ***Gross 2*** | | | | | | |
| **Ratio** | **∞↑** | **1.5↓** | **100↓** | **10↓** | **5↑** | **2↑** |
| fmol  /100 ng *E. coli* | 200 (B) | 80.0 (H) | 1.98 (O) | 18.2 (C) | 167 (J) | 133 (N) |
| 20.0 (R) | 8.00 (A) | 0.20 (G) | 1.82 (M) | 16.67 (E) | 13.33 (F) |
| 2.00 (K) | 0.80 (Q) | 0.02 (D) | 0.18 (L) | 1.67 (P) | 1.33 (I) |

Note. Four quantification standards (*Fine 1*, *Fine 2*, *Gross 1* and *Gross 2*) were created by adding exogenous protein digests to aliquots of a common *E. coli* digest in the amounts shown above. Rows marked ratio, list the fold difference in protein amount when comparing *Fine 1* with *Fine 2* and *Gross 1* with *Gross 2*. Arrows indicate the direction the fold-difference: ↑ = up; **↓** = down. Letters in parentheses provide a reference for the protein digest added as listed in table 1.

iTRAQ labelling

Four 45 g aliquots were taken from each quantification standard for 8-plex iTRAQ labelling, which was performed following the manufacturer’s protocol (SCIEX). Labelled samples were combined as pairwise comparisons of *Fine 1* *vs* *Fine 2* and *Gross 1* *vs* *Gross 2*, each comprising quadruple labelling replicates.

LC-MS/MS

Two replicate injections were performed for combined iTRAQ samples (500 ng loading per injection) and label-free samples were run with 10 replicate injections (100 ng loading per injection). All samples were diluted to 100 ng/L before loading onto a nanoAcquity UPLC system (Waters) equipped with a nanoAcquity Symmetry C18, 5 µm trap (180 µm x 20 mm Waters) and a nanoAcquity HSS T3 1.8 µm C18 capillary column (75 m x 250 mm, Waters). The trap wash solvent was 0.1% (v/v) aqueous formic acid and the trapping flow rate was 10 µL/min. The trap was washed for 5 min before switching flow to the capillary column. The separation used a gradient elution of two solvents (solvent A: 0.1% (v/v) formic acid; solvent B: acetonitrile containing 0.1% (v/v) formic acid). The flow rate for the capillary column was 300 nL/min. Column temperature was 60°C with the following gradient profiles:

Label-free – Linear 2-30% B over 125 minutes then linear 30-50% B over 5 minutes.

iTRAQ – Linear 2-30% B over 240 minutes then linear 30-50% B over 5 minutes.

All runs then proceeded to wash with 95% solvent B for 5 minutes. The column was returned to initial conditions and re-equilibrated for 25 minutes before subsequent injections.

The nanoLC system was interfaced to a maXis HD LC-MS/MS System (Bruker Daltonics) with a CaptiveSpray ionization source (Bruker Daltonics). Positive ESI-MS & MS/MS spectra were acquired using AutoMSMS mode. Instrument control, data acquisition and processing were performed using Compass 1.7 software (microTOF control, Hystar and DataAnalysis, Bruker Daltonics). Instrument settings were: ion spray voltage: 1,450 V, dry gas: 3 L/min, dry gas temperature 150 °C, ion acquisition range: *m/z* 150-2,000. AutoMS/MS settings were as stated in supporting information.

The collision energy and isolation width settings were automatically calculated using the AutoMSMS fragmentation table, absolute threshold 200 counts, preferred charge states: 2 – 4, singly charged ions excluded. A single MS/MS spectrum was acquired for each precursor and former target ions were excluded for 0.8 min unless the precursor intensity increased fourfold.

Database searching

Tandem mass spectral data were searched against the *E. coli* subset of the UniProt database appended with the added exogenous protein sequences (4,300 sequences; 1,357,294 residues). Searches were performed using a locally-running copy of the Mascot program (Matrix Science Ltd., version 2.5), through the Bruker ProteinScape interface (version 2.1). Search criteria specified: Enzyme, Trypsin; Fixed modifications, Methylthio (C); Variable modifications, Deamidated (NQ), Gln->pyro-Glu (N-term Q), Oxidation (M), Phospho (ST); Peptide tolerance, 10 ppm; MS/MS tolerance, 0.05 Da; Instrument, ESI-QUAD-TOF. For iTRAQ labelled samples search criteria also specified: Fixed modifications, iTRAQ8plex (N-term, K) and Variable modifications, iTRAQ8plex (Y).

Results were passed through Mascot Percolator to achieve a false discovery rate of 1% and further filtered to accept only peptides with an expect score of 0.05 or better.

Data analysis

iTRAQ

Data analysis was performed using a novel in-house procedure. Peptide identifications and reporter ion intensities were exported from Mascot search results and individual ion intensities normalized against total ion intensity for each channel. Percent intensity of each reporter ion was calculated for each spectrum before applying the student’s t-test (two-tailed, heteroscedastic) to the lists of reporter ion percentages at the protein level. Hochberg and Benjamini calculation of false discovery rate was applied to correct for multiple testing errors. Protein level fold-difference was calculated from the relative means of normalized reporter ion intensities.

Quantification was also performed using Mascot with median protein ratios calculated after normalization against the sum of intensities. Automatic outlier removal was applied. For both analyses a minimum of three unique peptides were required for quantification.

Peak area

Bruker .d files were loaded into Progenesis QI (Waters) for label free quantification by comparison of MS1 precursor area. LC-MS chromatograms were calibrated against two internal calibrants (*m/z* 299.2945 and *m/z* 1221.9906) before aligning and peak picking using the default settings. Default normalization to the total identified protein content was used. Peptide identifications were imported from Mascot and only unique peptides were used for quantification. Determination of significance was taken from QI’s ANOVA derived q-scores with a minimum of two unique peptides required for quantification.

emPAI

Mascot derived emPAI values were converted to molar percentages by normalizing against the sum of all emPAI values for the acquisition. Fold difference was calculated from the sum of molar percentages in all replicates, with the student’s t-test (two-tailed, heteroscedastic) applied to the same values. Hochberg and Benjamini calculation of false discovery rate was applied to correct for multiple testing errors. A minimum of three unique peptide sequences in at least three replicates were required for quantification.

Spectral counting

Mascot derived spectral count per protein were normalized against total number of spectral identifications for the acquisition. Fold difference was calculated from the sum of normalized spectral counts in all replicates, with the student’s t-test (two-tailed, heteroscedastic) applied to the same values. Hochberg and Benjamini calculation of false discovery rate was applied to correct for multiple testing errors. Spectral identification was required in at least three replicates for quantification to be considered.

Zero and infinity values

All methods avoided the use of ratios in statistical testing meaning zeros could be used as appropriate values in t-test calculations. A fold change of ∞ was generated where a protein is quantified in one sample but not the other. When assessing measurement accuracy, values measured as >10-fold for expected ∞-fold difference and values measured as ∞-fold difference for expected values of 100-fold difference are assigned an accuracy of 50%.

RESULTS

Quantification standards containing proteins at known ratios were analyzed by iTRAQ, peak area, emPAI and spectral counting relative quantification approaches. Two standards were used, one to assess fine differences (1.1- to 1.75-fold) and the other to measure gross differences (1.5- to ∞-fold), each prepared by spiking 18 exogenous proteins into a complex background of *E. coli* lysate at known amounts. Label-free acquisitions were run with 10 technical replicates and iTRAQ analysis used all eight available 8-plex labels for pairwise-comparison (4 *vs* 4), with two replicate injections. Fold differences and p-values were calculated as described in the methods and tables of results are presented as supporting information. Data sets are available to download from MassIVE (MSV000079803) and ProteomeXchange (PXD004290).

Diagnostic classification

Diagnostic classification accuracy was tested based on the following categorizations: a true positive (TP) is an exogenous protein correctly classified as different; a false positive (FP) is an *E. coli* protein classified as different; a true negative (TN) is an *E. coli* protein classified as unchanged and a false negative (FN) is an exogenous protein classified as unchanged.

Results were first compared by applying a common 1% false positive rate (FPR) threshold, as summarized in tables 3 and 4.

**Table 3. Summary of results from *Gross* standard at 1% FPR**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Method** | **TP** | **TN** | **FP** | **FN** |
| iTRAQ | 16 | 486 | 5 | 2 |
| Peak area | 14 | 396 | 4 | 4 |
| emPAI | 14 | 592 | 7 | 4 |
| spectral counts | 14 | 740 | 7 | 4 |

Note. Comparison of quantification approaches applied to analysis of *Gross* quantification standard containing 18 exogenous proteins added to a common background of *E. coli* lysate at ratios between 1.5- and ∞-fold. Results filtered to 1% FPR.

**Table 4. Summary of results from *Fine* standard at 1% FPR**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Method** | **TP** | **TN** | **FP** | **FN** |
| iTRAQ | 9 | 515 | 5 | 9 |
| Peak area | 15 | 463 | 5 | 3 |
| emPAI | 7 | 660 | 6 | 11 |
| Spectral counts | 12 | 789 | 8 | 6 |

Note. Comparison of quantification approaches applied to analysis of *Fine* quantification standard containing 18 exogenous proteins added to a common background of *E. coli* lysate at ratios between 1.1- and 1.75-fold. Results filtered to 1% FPR.

A key feature of any quantification approach should be to maximize both sensitivity and specificity, defined as:

1. Sensitivity = TP/(TP+FN)
2. Specificity = TN/(TN+FP)

Sensitivity is a measure of how likely a positive outcome is a true difference, while specificity is the probability that a value classed as a negative outcome really is unchanged. Another metric of diagnostic classification is positive predictive value (PPV). PPV measures the worth of a positive classification by expressing the number of true positives as a proportion of the total positive calls, as defined below:

1. PPV = TP/(TP+FP)

The sensitivity, specificity and PPV values achieved by the four quantification approaches at 1% FPR are summarized in tables 5 and 6.

For the gross standard, which measured fold-differences between 1.5- and ∞-fold iTRAQ was assessed to achieve the highest sensitivity (89%), with all three label-free approaches performing slightly worse than iTRAQ but equally well with each other (78%). Comparing PPV, iTRAQ and peak area both gave higher values (76% and 78%, respectively) than emPAI and spectral counting (both 67%).

The differences between the approaches were more dramatic with the fine standard, which compared proteins spiked at ratios between 1.1- and 1.75-fold. Peak area achieved the best sensitivity and PPV (83% and 75%, respectively). Spectral counting showed the second best sensitivity (67%), followed by iTRAQ, which only managed to identify half of the exogenous proteins as different. The emPAI approach performed least well giving a sensitivity of just 39% with a PPV of 54%. It should be noted that although the sensitivity of spectral counting was greater than iTRAQ the PPV was slightly lower – 64% for iTRAQ and 60% for spectral counting. These differences result from the fact that although spectral counting provides an additional three TPs over iTRAQ it is at the cost of three FPs.

**Table 5. Sensitivity, specificity and PPV achieved with *Gross* standard at 1% FPR**

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Sensitivity** | **Specificity** | **PPV (%)** |
| iTRAQ | 89 | 99 | 76 |
| Peak area | 78 | 99 | 78 |
| emPAI | 78 | 99 | 67 |
| Spectral counts | 78 | 99 | 67 |

Note. Comparison of sensitivity specificity and PPV achieved with different quantification approaches applied to analysis of *Gross* quantification standard. The standard comprised 18 exogenous proteins added to a common background of *E. coli* lysate at ratios between 1.5- and ∞-fold. Results filtered to 1% FPR.

**Table 6. Sensitivity, specificity and PPV achieved with Fine standard at 1% FPR**

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Sensitivity** | **Specificity** | **PPV (%)** |
| iTRAQ | 50 | 99 | 64 |
| Peak area | 83 | 99 | 75 |
| emPAI | 39 | 99 | 54 |
| Spectral counts | 67 | 99 | 60 |

Note. Comparison of sensitivity specificity and PPV achieved with different quantification approaches applied to analysis of *Fine* quantification standard. The standard comprised 18 exogenous proteins added to a common background of *E. coli* lysate at ratios between 1.1- and 1.75-fold. Results filtered to 1% FPR.

ROC analysis

The data presented above compare results at a single fixed threshold of 1% FPR. Although the 1% FPR gives an indication of how the techniques would for an unknown further analysis the underlying potential of each technique can be provided by comparing a range of thresholds. Receiver operator characteristic (ROC) curves show the relationship between sensitivity and 1-specificity with the area under the curve (AUC) denoting the overall diagnostic accuracy of the test across all thresholds.61

Sensitivity and specificity were calculated for all quantification techniques across a range of p-value thresholds. The data were plotted to create ROC curves as shown in figures 1A and B. Spectral counting applied to the sample with gross differences in protein fold-differences produced a highly accurate test with an AUC of 0.97. The other three methods also performed well in the number of correct assessments of gross differences with an AUC ≥ 0.85. For the sample with fine differences, emPAI produced an AUC of just 0.79, iTRAQ and spectral counting performed similarly with scores of 0.83 and 0.84, respectively, while peak area showed the highest classification accuracy with an AUC of 0.86.

**Figure 1A and 1B.** ROC curves demonstrating the diagnostic classification performance of different quantification techniques applied to the measurement of 18 exogenous proteins added to *E. coli* lysate with fold differences between 1.5- and ∞-fold (1A – Gross) and 1.1- and 1.75-fold (1B – Fine). The area under the curve (AUC) is given in the figure legend. AUC denotes the overall diagnostic accuracy of the test.

Figures 2A and 2B plot the PPVs calculated for the top 50 corrected p-value ranked differences observed with each quantification method. For gross fold-differences (>1.5-fold) iTRAQ yielded the highest PPVs although all methods performed well and were closely grouped. When fine changes were assessed peak area gave the highest PPVs. Spectral counting and iTRAQ performed very similarly and emPAI ranked lowest. These observations show correlation with the ROC curve data suggesting that peak area offers the highest classification accuracy for small fold-differences.

**Figure 2A and 2B.** PPV curves demonstrating the positive predictive value (PPV) of different quantification techniques applied to the measurement of 18 exogenous proteins added to *E. coli* lysate with fold differences between 1.5- and ∞-fold (2A – Gross) and 1.1- and 1.75-fold (2B – Fine). PPV is plotted for the top 50 proteins ranked by corrected p-value.

Effect of fold difference and absolute abundance

The quantification standards had fold differences from 1.1- to ∞-fold, with each ratio measured for three different proteins added at three different absolute abundances: 1, 10 and 100 fmol/100 ng *E. coli* lysate. To investigate whether these variables affect the diagnostic classification performance a heat map was produced showing the false positive rate (FPR = 1-specificity) threshold required for each exogenous protein to be called as a TP (figure 3). To summarize the inference of the heat map: the darker red the greater the cost to specificity to achieve correct classification.



**Figure 3.** Heat map displaying the minimum FPR (1-specificity) threshold required for different quantification approaches to correctly class exogenous protein measurements as changing. Exogenous proteins were added to an *E. coli* lysate at known ratios from 1.1-fold to ∞-fold difference. Each measurement was made with different proteins at three absolute abundances: 1, 10 and 100 fmol/100 ng *E. coli* lysate. Grey boxes indicate single peptide measurements excluded from quantification.

There are clear visual differences between techniques when comparing the specificity cost required to achieve the correct classification of specific exogenous proteins as changing. iTRAQ performance scales with the abundance at which the measurement is made. Fold-differences as low as 1.2-fold were correctly classified as changing with acceptable specificities when measured at high abundance (100 fmol) relative to the background. When the same fold-differences were measured at the lowest absolute abundance (1 fmol) the required specificity costs were much higher, limiting iTRAQ’s diagnostic use to 5-fold difference or greater at this level.

Spectral counting and emPAI show a contrasting trend to iTRAQ, in which classification performance scales not with the abundance at which the measurement is made but with the fold-difference requiring classification. Spectral counting performs marginally better than emPAI, with the former able to correctly classify differences in exogenous proteins at >1.2-fold. emPAI is a reliable method for the classification of differences >1.3-fold. There is no apparent trend with the abundance at which the measurements are made, implying that the absolute limit of quantification was not reached with these data.

Peak area performs well across all fold-differences and relative abundances and remarkably is able to correctly classify a 1.1-fold change at <5% FPR. Four of the exogenously spiked proteins were not identified with sufficient peptides (>2) for quantification and so could not be classified. These results may suggest that the limit of identification is reached at the same point as the limit of quantification.

Measurement accuracy

Although correct diagnostic classification is vital to expression proteomic analyses, measurement accuracy may also be important. Measurement accuracy was determined as defined in equation 3.

Figure 4 displays the measurement accuracy for each exogenous protein measured with four different quantification approaches.



**Figure 4.** Heat map displaying measurement accuracy with different quantification approaches applied to the relative quantification of exogenous proteins added to an *E. coli* lysate in known ratios from 1.1-fold to ∞-fold difference. Each measurement was made with different proteins at three absolute abundances: 1, 10 and 100 fmol/100 ng *E. coli* lysate. Values measured as >10-fold for expected ∞-fold difference and values measured as ∞-fold difference for expected values of 100-fold difference are assigned an accuracy of 50%. Symbols indicate whether the measured fold-difference was less than (-) or greater (+) than expected fold-difference. The symbol = is used for ∞-fold differences measured as ∞.

Peak area quantification shows the highest measurement accuracy among the techniques, although accuracy is reduced when measurements are made at lower abundance relative to the background or at extreme fold differences. The measurements show a mixture of under and over estimation indicating no systematic bias.

Spectral counting and emPAI methods appear to show little correlation with measurement accuracy, fold-change or absolute abundance. Spectral counting achieves over 50% measurement accuracy for 85% of the exogenous proteins, with greater accuracy than peak area when the greatest magnitude differences are measured at the lowest abundance. Both spectral counting and emPAI were shown to broadly underestimate the expected fold-difference.

iTRAQ accuracy is lower than all other methods, only generating accuracies greater than 50% for small magnitude fold differences measured at high absolute abundance. Inspection of fold differences measured by iTRAQ shows the data to be heavily attenuated with all measurements an underestimation of the true ratio.

iTRAQ data analysis

The novel in-house data analysis scheme, summarized in figure 5, was used with all iTRAQ data presented. The procedure normalizes against total reporter ion intensity before rescaling reporter ions within each spectrum to give percentage intensities, allowing direct comparison of all reporter ions assigned to a protein. The Student’s t-test (two-tailed, heteroscedastic) is applied at the protein level, decoupling reporter ions from individual peptides and avoiding the use of ratios in statistical testing. The method differs from standard methodologies, which calculate peptide level ratios before a protein level significance test against unity. Established workflows necessitate transformation of ratios, typically to log2 and often struggle with missing values, unlike our approach, which, is also more amenable to the use of replicates within groups. For example, when four replicates of A (A1-4) are compared to four replicates of B (B1-4), conventional methodologies would assign ratios in pairs, such as A1/B1, A2/B2, A3/B3 and A4/B4. Significance would then be tested against unity for the ratios calculated. In reality, there may be no more of a link between A1 and B1 than A1 and any other Bn sample. An alternative would be to define the peptide ratios ∑A1-4/∑B1-4, but this reduces the number of data points. Instead, our methodology compares all values from A1-4 to all values from B1-4 in a non-paired manner. The decoupling of reporter ions from individual peptide ratios has the potential to smooth peptide level variation.



**Figure 5.** Schematic of in-house iTRAQ data analysis scheme. Step 1, peptide identifications and reporter ion intensities exported from Mascot search result. Step 2, individual ion intensities normalized against total intensity for each channel. Step 3, percent intensity of each reporter ion calculated for each spectrum. Step 4, student’s t-test (two-tailed, heteroscedastic) applied at the reporter ion percentages at the protein level. Multiple test correction applied by converting p-values to Hochberg and Benjamini false discovery rates.

The diagnostic classification performance of the novel in-house iTRAQ data analysis scheme was compared to an established commercial alternative from Mascot (Matrix Science) by constructing ROC curves as shown in figures 6A and 6B.

Good classification accuracy (AUC values of 0.8-0.9) is achieved for both approaches when looking at gross differences (1.5- to ∞-fold difference), but there is a clear differentiation between the results for classification of small-fold differences (1.1- to 1.75-fold) with good results for the in-house method (AUC = 0.83) and poor results for Mascot (AUC = 0.60).

**Figure 6A and 6B.** ROC curves demonstrating the diagnostic classification accuracy of in-house and Mascot processing of the iTRAQ data from 18 exogenous proteins added to *E. coli* lysate with fold differences between 1.5- and ∞-fold (1A – *Gross*) and 1.1- and 1.75-fold (1B – *Fine*).

The trend of reduced classification performance at smaller magnitude fold difference and lower absolute abundance is consistent between the iTRAQ data analysis regimes, as shown in figure 7.



**Figure 7.** Heat map displaying the minimum FPR threshold required for different iTRAQ data analysis approaches to class exogenous protein measurements as true positives. Exogenous proteins were added to an *E. coli* lysate at known ratios from 1.1-fold to ∞-fold difference. Each measurement was made with different proteins at three absolute abundances: 1, 10 and 100 fmol/100 ng *E. coli* lysate. Blue and grey boxes indicate proteins excluded from quantification because of lack of normally distributed measurements, or insufficient peptide identifications, respectively.

The reduced classification accuracy obtained when using Mascot can be rationalized partly by the six measurements excluded from significance testing for lacking normality. Even with these six quantifications discounted, figure 7 demonstrates that the in-house approach shows increased diagnostic classification performance across multiple measurements. Particularly striking are the 100- and ∞-fold differences measured at 1 fmol loadings, which are significant with <0.5% FPR using the in-house method but require a FPR >5% for acceptance when processed through Mascot. Quantification of the 1 fmol, ∞-fold difference, measurement is derived from eight ovalbumin (sp|P01012|) peptide identifications. Using the in-house method gives a total of 32 quantification measurements per group (no. peptides x no. labels). Performing a t-test on the two groups gives a p-value of 3.43x10-09, which drops to 1.16 x10-07 upon Hochberg and Benjamini multiple test correction. The p-value generated using Mascot for the same protein is 0.03, with no correction applied.

All results are consistent with the in-house data analysis regime out-performing the commercial alternative across our datasets.

DISCUSSION

The production of standards for proteomic quantification has been presented expanding on the design employed by Sjödin.54 The standards mimic an expression proteomics analysis in which protein differences would be elucidated among samples within a common unchanging background; however expected differences are known, providing a system that methodologies can be reliably benchmarked against.

In addition to testing three common relative quantification approaches, iTRAQ, label-free peak area and spectra counting, emPAI has been validated against these more established methodologies for the first time. Although emPAI was never intended for inter-sample relative quantification42 is has been used as such in published studies,43-45 making its validation crucial. Metabolic labelling approaches, such as SILAC13 were not compared as these were felt to be less universally applicable.16 Acquisitions were performed to maximize the performance of the quantification techniques as detailed in the methods, rather than maintaining a common, compromised set of parameters. Standards were created from a common digestion pool meaning variance imparted by or preceding proteolytic digestion could be excluded. It is noted that such steps would be conducted independently in typical proteomic quantification, increasing the total technical variance. Previous studies have measured variance imparted by common processing steps up to the point of MS analysis.15 In practice, the specific techniques used and precision obtained are likely to be both sample and user dependent, furthermore biological variance is expected to exceed all levels of technical variance.62 The study presented here is intended to compare relative quantification techniques in isolation of upstream variation to test the maximum limits of diagnostic classification performance.

The datasets developed have been made publicly available through MassIVE (MSV000079803) and ProteomeXchange (PXD004290) and will provide a valuable training set for the development of software and statistical methods beyond the scope of one paper. Diagnostic classification is at the crux of any expression proteomics analysis, with the aim being to correctly classify differently expressed and non-changing proteins. Classification accuracy was investigated for all methodologies using ROC curves,61 and PPV.

Spectral counting

Spectral counting is often referred to as only semi-quantitative,39-41 yet these results show that when used with sufficient statistical rigor it provides a highly capable diagnostic classification approach. Spectral counting can be used with confidence as an alternative to the more traditionally recognized approaches of peak area and iTRAQ. At a fixed 1% FPR spectral counting falls below iTRAQ for detecting gross changes but provided more sensitivity for fine changes. Spectral counting outperforms all tested methodologies when classifying gross fold differences (>1.5-fold differences) as assessed using ROC analysis and is only exceeded by peak area when looking at the smallest fold-differences. Positive predictive values are on a par with iTRAQ for both gross and fine fold differences. Although spectral counting is limited to the classification of fold-differences above 1.2-fold with the data sets used in this study, such a limitation will be compatible with many biological analyses. Strikingly, over the range tested, classification accuracy appears independent of the abundance at which the difference is measured. The observation is in marked contrast to iTRAQ, where classification accuracy drops significantly when the same ratios are measured at lower absolute abundance relative to the background. The high performance of spectral counting for classifying changes in low abundance proteins can be rationalized against the specific nature of the measurement event - i.e. a peptide spectrum match (PSM). PSMs have multiple levels of stringency, the precursor mass must fall within the specified search tolerance and multiple product ions need to be observed. It is also common to perform additional global validation of PSMs such as FDR estimation or percolation. Such rigor makes non-specific contributions to quantification measurements very rare, meaning that for proteins observed above the threshold of three unique peptides quantification is amenable regardless of the absolute abundance. In contrast, the quantified measurement with peak area and iTRAQ relies on a single mass - the precursor area for peak area and the reporter ion intensity for iTRAQ. Although these metrics are contained within more stringent assessments (correlation of retention time, identification of PSMs, etc.) the individual measurements are more prone to contribution from non-specific ions. For example, while a precursor area can be confidently matched to a PSM, it is very difficult to exclude any underlying contribution from additional unidentified analytes at equal *m/z* and retention time. Likewise, although an iTRAQ product ion spectrum can be confidently matched to a PSM, additional peptides may be co-selected for fragmentation and contribute to reporter ion intensity, as has been reported previously.63 Such non-specific contributions are likely to be greater for lower abundance proteins where the relative intensity of the intended measurement is reduced. For small-fold differences spectral counting was shown to be less proficient, an observation likely due to the discrete nature of sampling. Classification of change requires a minimum difference of one PSM. Both iTRAQ and peak-area methods can be considered to have continuous measurement values only limited by the precision of the detector. These findings could be considered when deciding which technique is most appropriate for a given application. If identification of moderate fold-changes in low abundance proteins is valued over identification of small-fold changes in highly abundant proteins than spectral counting may be advised in place iTRAQ. Furthermore, spectral counting can be viewed as the simplest approach to integrate into standard proteomics workflows and does not require expensive reagents or specialist software for data analysis. Measurement accuracy for spectral counting is also surprisingly good across the range of abundances and fold-differences investigated and although shown to tend towards underestimation far exceeds iTRAQ in this regard.

emPAI

Although never intended as such, the use of emPAI for relative inter-sample quantification,42 has been creeping into the literature,43-45 but to our knowledge this study provides the first published validation against more accepted methodologies for inter-sample relative quantification. Our results suggest that emPAI is a valid methodology that can reliably classify differences above 1.3-fold difference. Spectral counting, which is equally simple to implement, outperforms emPAI in all aspects of this study suggesting that spectral counting would be a better approach for most data sets. The reduced performance of emPAI relative to spectral counting is likely due to the increased quantization of the quantification measurement from PSMs to unique peptide sequences. It should be noted that emPAI does have value and should be considered when information on relative inter-protein, intra-sample quantification is also required.42

Peak area

Peak area performed extremely well in all comparisons. It is set apart from the other techniques when classifying the smallest fold-differences and was shown to be capable of identifying differences as low as 10% at acceptable specificity. Peak area is also seen to provide the most accurate measurement of fold-difference with no systematic bias towards underestimation. The only negative result for this method was the loss of four protein identifications that were retained with the other methodologies. Such a result may suggest the quantification software is losing achievable identifications but also that the quantification threshold extends to the identification threshold, meaning that reliable quantification can be achieved on all identified proteins in the dataset, excluding single peptide matches.

iTRAQ

The classification performance of iTRAQ can be viewed as good for both ranges of fold-change analyzed. However, deeper interrogation of the iTRAQ data shows a significant drop in classification performance when differences are present at lower abundance relative to the background. Such an observation calls into question iTRAQ’s ability to identify changes in low abundance proteins, which are often considered the most interesting biologically. Chimeric components are known to be a ubiquitous feature of product ion spectra in complex isobaric labelled samples.57 It is expected that such background components would have a more dramatic effect when the selected precursor intensity is reduced: explaining the observed correlation between absolute abundance and classification performance. Measurement accuracy is lowest with iTRAQ quantification, where differences are attenuated throughout the measurements, a feature that has been well documented in previous analyses.63-65 While ion-trap instruments offer the capability of MS3 fragmentation to significantly improve iTRAQ measurement accuracy66, 67 no equivalent is available with conventional qTOF instruments. In spite of the negatives, iTRAQ is the most amenable technique to orthogonal separation as samples are combined prior to fractionation68 and the reduction of costly LC-MS time ensures its value for many analysis schemes.

Novel iTRAQ data analysis strategy

A novel processing strategy has been presented for iTRAQ data analysis and has been shown to outperform Mascot’s established commercial alternative. The strategy normalizes reporter ion intensities, decoupling the need for ratio pairs and allowing all peptide quantification measurements to be considered simultaneously at the protein level. The observed improvement in classification accuracy is consentient with work presented by Goeminne who demonstrated that peptide-based strategies for statistical testing outperform classical summarization methods.69

The processing strategy demonstrated here is extremely easy to implement and can be processed through Excel using the Mascot derived .csv exported results. It is envisioned that implementation into existing software solutions could be achieved with minimal difficulty. It was noted that Mascot quantification excluded multiple true positives for lacking normality. The presented method makes no test for normality and it is acknowledged that Gaussian distributions are an underlying assumption of the t-test used. However, it has been demonstrated empirically that the approach is able to reliably classify true differences. Furthermore, studies have suggested that t-tests are surprisingly tolerant to non-normality, especially where two-tailed comparisons are made.70-72 Alternatively, once reporter ions are normalized and grouped the method could be adjusted to incorporate non-parametric significance tests.

Limitations

The following considerations should be applied to the results presented in this manuscript. The analysis has been performed on a single LC-MS system meaning further work beyond the scope of this paper will be required to show that the trends translate across the field. Similarly, the quantification standards are derived from a common pool of 18 proteins and although trends are observed based on fold-difference and protein abundance, individual protein level biases cannot be fully excluded. We hope that the work presented here will lead to further investigation of these observations. Furthermore it is noted that Mascot is not the only available software for iTRAQ data analysis. A wide array of solutions exists from commercial software to in-house approaches and it is impractical to compare against all possible solutions. Our data sets are available to download from MassIVE (MSV000079803) and ProteomeXchange (PXD004290) so that groups can easily test their data analysis method of choice against the method we present.

CONCLUSION

A comparison of iTRAQ, peak area, spectral counting and emPAI methodologies for relative quantification was presented using a known, complex system. ROC curves were used to assess the diagnostic classification accuracy of each technique. Spectral counting was seen to perform surprisingly well in all regards and outperformed iTRAQ when assessing differences in proteins of low abundance relative to the background. Peak area is observed to perform best for low fold-differences (down to 1.1-fold) and exceeds the other methods in terms of measurement accuracy. We have presented the first systematic comparison of emPAI with more established methodologies for inter-sample relative quantification and although its use would not be favored above the alternatives it is shown to be a valid approach. A novel method of iTRAQ data analysis has been presented that avoids the use of ratios in statistical testing and is shown to outperform a common commercial alternative. Datasets have been made publicly available to provide valuable resource for development of software and testing statistical of regimes beyond the scope of a single paper.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website.

**Table S-1** Cover page (.docx)

**Table S-2** Auto-MS/MS parameters (.xlsx)

**Table S-3** Sensitivity and specificity results summary (.xlsx)

**Table S-4** iTRAQ Gross in-house data analysis workflow (.xlsx)

**Table S-5** iTRAQ Fine in-house data analysis workflow (.xlsx)

**Table S-6** iTRAQ unprocessed search results (.xlsx)

**Table S-7** Label free search results (.xlsx)

AUTHOR INFORMATION

Corresponding Author

\* Phone (44)-(0)-1904-328739; email adam.dowle@york.ac.uk

Author Contributions

All authors have given approval to the final version of the manuscript.

ACKNOWLEDGMENT

All data acquisitions were performed in York Centre of Excellence in Mass Spectrometry. The York Centre of Excellence in Mass Spectrometry was created thanks to a major capital investment through Science City York, supported by Yorkshire Forward with funds from the Northern Way Initiative.

REFERENCES

(1) Gygi S.P., Rist B., Gerber S.A., Turecek F., Gelb M.H., Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags *Nature Biotechnology* **1999**,17,994–999

(2) Blackstock W. P., Weir M. P. Proteomics: quantitative and physical mapping of cellular proteins *Trends in Biotechnology* **1999**,17(3), 121–127.

(3)Brownridge P., Holman S.W., Gaskell S.J., Grant C.M., Harman V.M., Hubbard S.J., Lanthaler K., Lawless C., OʼCualain R., Sims P., Watkins R., Beynon R. J. Global absolute quantification of a proteome: Challenges in the deployment of a QconCAT strategy *Proteomics* **2011**,11,2957–2970

(4) Picotti P. & Aebersold R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat Methods* **2012**, 9, 555–66

(5) Lawless C., Holman, S. W., Brownridge, P., Lanthaler, K., Harman, V. M., Watkins, R., Sims, P. F. G., Grant, C. M., Eyers, C. E, Beynon, R. J. and Hubbard, S. J. Absolute Quantification of over 1800 Yeast Proteins via Selected Reaction Monitoring. *Mol Cell Proteomics* **2016**, 15(4), 1309-22

(6) Rosenberger G, Ludwig C, Röst HL, Aebersold R, Malmström L. aLFQ: an R-package for estimating absolute protein quantities from label-free LC-MS/MS proteomics data. *Bioinformatics* **2014** 30(17), 2511-3

(7) Wiśniewski JR, Hein MY, Cox J, Mann M. A "proteomic ruler" for protein copy number and concentration estimation without spike-in standards. *Mol Cell Proteomics* **2014** 13(12), 3497-506

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

(9) Unwin R. D., Evans C. A., Whetton A. D. Relative quantification in proteomics: new approaches for biochemistry *Trends Biochem Sci*. **2006**, 31, 473–484

(10) Gevaert K., Impens F., Ghesquiere B., Van Damme P., Lambrechts A., Vandekerckhove J. Stable isotopic labeling in proteomics *Proteomics* **2008**, 8, 4873–4885

(11) Rauniyar N., Yates III J. R. Isobaric labeling-based relative quantification in shotgun proteomics *J Proteome Res*, **2014**, 13, 5293–5309

(12) Beynon R. J., Pratt J. M. Metabolic labeling of proteins for proteomics *Mol Cell Proteomics* **2005**, 4, 857–872

(13) Ong S. E., Blagoev B., Kratchmarova I., Kristensen D.B., Steen H., Pandey A., Mann M Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics *Mol Cell Proteomics* **2005**, 1(5)**,** 376–38

(14) Oda Y., Huang K., Cross F.R., Cowburn D., Chait B.T. Accurate quantitation of protein expression and site-specific phosphorylation *Proc Natl Acad Sci USA* **1999**, 96, 6591–6596

(15) Russell M.R., Lilley K.S. Pipeline to assess the greatest source of technical variance in quantitative proteomics using metabolic labelling *J Proteomics* **2012**, 77, 441–454

(16) Bantscheff M., Schirle M., Sweetman G., Rick J., Kuster B. Quantitative mass spectrometry in proteomics: a critical review *Analytical and Bioanalytical Chemistry* **2007**, 389, 1017–31

(17) Hsu J.L., Huang S.Y., Chow N.H., Chen S.H. Stable-isotope dimethyl labeling for quantitative proteomics *Anal Chem* **2003**, 75, 6843–6852

(18) Yao X., Freas A., Ramirez J., Demirev P. A., Fenselau C. Proteolytic 18O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Anal Chem.* **2001**, 73, 2836–2842

(19) Ji J., Chakraborty A., Geng M., Zhang X., Amini A., Bina M., Regnier F. Strategy for qualitative and quantitative analysis in proteomics based on signature peptides *Journal of Chromatography B* **2000**, 745(1), 197–210

(20) Lottspeich F., Kellermann J. ICPL labeling strategies for proteome research *Methods Mol Biol* **2011**, 753, 55–64

(21) Ross P.L., Huang Y.N., Marchese J.N., Williamson B., Parker K., Hattan S., Khainovski N., Pillai S., Dey S., Daniels S., Purkayastha S., Juhasz P., Martin S., Bartlet-Jones M., He F., Jacobson A., Pappin D. J. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents *Mol Cell Proteomics* **2004**, 3, 1154–1169

(22) Choe L., D’Ascenzo M., Relkin N. R., Pappin D., Ross P., Williamson B., Guertin S., Pribil P., Lee K. H. 8-Plex quantitation of changes in cerebrospinal fluid protein expression in subjects undergoing intravenous immunoglobulin treatment for Alzheimer’s disease *Proteomics* **2007**, 7, 3651–3660

(23) Thompson A., Schäfer J., Kuhn K., Kienle S., Schwarz J., Schmidt G., Neumann T., Johnstone R., Mohammed A. K., Hamon C. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS *Analytical Chemistry* **2003**, 75(8), 1895–1904

(24) Dayon L. Hainard A., Licker V., Turck N., Kuhn K., Hochstrasser D. F., Burkhard P. R., Sanchez J. C. Relative Quantification of Proteins in Human Cerebrospinal Fluids by MS/MS using 6-Plex Isobaric Tags *Anal. Chem.* **2008**,80, 2921–2931

(25) Werner T., Becher I., Sweetman G., Doce C., Savitski M. M., Bantscheff M. High-resolution enabled TMT 8-plexing *Anal. Chem.* **2012**, 84, 7188–7194

(26) McAlister G. C., Huttlin E. L., Haas W., Ting L., Jedrychowski M. P., Rogers J. C., Kuhn K., Pike I,. Grothe R. A., Blethrow J. D., Gygi S.P. Increasing the multiplexing capacity of TMT using reporter ion isotopologues with isobaric masses. *Anal Chem.* **2012**, 84, 7469–7478

(27) Ow S. Y., Noirel J., Salim M., Evans C., Watson R., Wright P. C. Balancing robust quantification and identification for iTRAQ: application of UHR-ToF MS *Proteomics* **2010**, 10(11), 2205–2213

(28) Neilson K. A., Ali N. A., Muralidharan S., Mirzaei M., Mariani M., Assadourian G., Lee A., van Sluyter S. C., Haynes P. A. Less label, more free: approaches in label-free quantitative mass spectrometry. Proteomics. **2011**, 11, 535–553

(29) Higgs R. E., Knierman M. D., Gelfanova V., Butler J. P., Hale J. E. Comprehensive label-free method for the relative quantification of proteins from biological samples. *J Proteome Res.* **2005**, 4, 1442–50

(30) Liu H., Sadygov R. G., Yates J. R. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem.* **2004**, 76(14), 4193–4201

(31) Clough T., Thaminy S., Ragg S., Aebersold R., Vitek O. Statistical protein quantification and significance analysis in label-free LC-MS experiments with complex designs *BMC Bioinformatics* **2012**, 13, S6

(32) Cappadona S., Baker P. R., Cutillas P. R., Heck A. J. R., van Breukelen B. Current challenges in software solutions for mass spectrometry-based quantitative proteomics. *Amino Acids* **2012**, 43(3), 1–2

(33) Hoehenwarter W., Wienkoop S. Spectral counting robust on high mass accuracy mass spectrometers *Rapid Commun Mass Spectrom* **2010**, 24(24), 3609-14

(34) Cooper B., Feng J., Garrett W. M. Relative, label-free protein quantitation: spectral counting error statistics from nine replicate MudPIT samples. *J. Am. Soc.* *Mass Spectrom*. **2010**, 21, 1534–1546

(35) Lundgren D. H., Hwang S.-I., Wu L., Han D. K. Role of spectral counting in quantitative proteomics. Expert Review of Proteomics. **2010**, 7(1), 39–53

(36) Zybailov B, Mosley AL, Sardiu ME, Coleman MK, Florens L, Washburn MP. Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *J Proteome Res*. **2006** 5(9), 2339-47

(37) Lu P, Vogel C, Wang R, Yao X, Marcotte EM. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol.* **2007** 25(1):117-24.

(38) Kannaste O, Suomi T, Salmi J, Uusipaikka E, Nevalainen O, Corthals GL .Cross-correlation of spectral count ranking to validate quantitative proteome measurements. *J Proteome Res.* **2014** 13(4), 1957-68.

(39) Ulrich C., Quilici D. R., Schlauch K. A., Buxton I. L. Proteomic network analysis of human uterine smooth muscle in pregnancy, labor, and preterm labor *Integr Mol Med.* **2015**, 2(4), 261-269

(40) Hayashi J, Kihara M, Kato H, Nishimura T. A proteomic profile of synoviocyte lesions microdissected from formalin-fixed paraffin-embedded synovial tissues of rheumatoid arthritis *Clin Proteomics* **2015**, 12(1)**,** 20

(41) Orton D. J., Doucette A. A., Maksym G. N., MacLellan D. L. Proteomic analysis of rat proximal tubule cells following stretch-induced apoptosis in an in vitro model of kidney obstruction *J Proteomics* **2014**, 100, 125-35

(42) Ishihama Y., Oda Y., Tabata T., Sato T., Nagasu T., Rappsilber J., Mann M. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein *MCP* **2005**, 4(9), 1265-72

(43) Carvalhais V., Cerca N., Vilanova M., Vitorino R. Proteomic profile of dormancy within Staphylococcus epidermidis biofilms using iTRAQ and label-free strategies *Appl Microbiol Biotechnol.* **2015**, 99(6), 2751-62

(44) Ternan N. G., Jain S., Graham R. L., McMullan G. Semiquantitative analysis of clinical heat stress in Clostridium difficile strain 630 using a GeLC/MS workflow with emPAI quantitation. *PLoS One* **2014**, 24;9(2)

(45) Urbani A., Sirolli V., Lupisella S., Levi-Mortera S., Pavone B., Pieroni L., Amoroso L., Di Vito R., Bucci S., Bernardini S., Sacchetta P., Bonomini M. Proteomic investigations on the effect of different membrane materials on blood protein adsorption during haemodialysis. *Blood Transfus.* **2012**, 10(2), 101-12

(46) Latosinska A., Vougas K., Makridakis M., Klein J., Mullen W., Abbas M., Stravodimos K., Katafigiotis I., Merseburger A. S., Zoidakis J., Mischak H., Vlahou A., Jankowski V. Comparative Analysis of Label-Free and 8-Plex iTRAQ Approach for Quantitative Tissue Proteomic Analysis *PLoS One* **2015**, 2;10(9)

(47) V. Trinh H. V., Grossmann J., Gehrig P., Roschitzki B., Schlapbach R., Greber U. F., Hemmi S. iTRAQ-Based and Label-Free Proteomics Approaches for Studies of Human Adenovirus Infections *International Journal of Proteomics*, **2013**, Article ID 581862

(48) Patel V. J., Thalassinos K., Slade S. E., Connolly J. B., Crombie A., Murrell J. C., Scrivens J. H. A comparison of labeling and label-free mass spectrometry-based proteomics approaches *J Proteome Res.* **2013**, 8(7), 3752-9

(49) Li Z., Adams R. M., Chourey K., Hurst G. B., Hettich R. L., Pan C. Systematic comparison of label-free, metabolic labeling, and isobaric chemical labeling for quantitative proteomics on LTQ Orbitrap Velos. *J Proteome Res.* **2012**, 11(3), 1582-90

(50) Megger D. A., Pott L. L., Ahrens M., Padden J., Bracht T., Kuhlmann K., Eisenacher M., Meyer H. E., Sitek B. Comparison of label-free and label-based strategies for proteome analysis of hepatoma cell lines *Biochim Biophys Acta.* **2014**, 1844(5), 967-76

(51) Schmidt C., Grønborg M., Deckert J., Bessonov S., Conrad T., Lührmann R., Urlaub H. Mass spectrometry-based relative quantification of proteins in precatalytic and catalytically active spliceosomes by metabolic labeling (SILAC), chemical labeling (iTRAQ), and label-free spectral count *RNA* **2014**, 20(3), 406-20

(52) Wang H., Alvarez S., Hicks L.M.. Comprehensive comparison of iTRAQ and label-free LC-based quantitative proteomics approaches using two Chlamydomonas reinhardtii strains of interest for biofuels engineering *J Proteome Res.* **2012**, 11(1), 487-501

(53) Sandberg A, Branca RM, Lehtiö J, Forshed J. Quantitative accuracy in mass spectrometry based proteomics of complex samples: the impact of labeling and precursor interference. *J Proteomics.* **2014**, 96, 133-44

(54) Sjödin M. O., Wetterhall M., Kultima K., Artemenko K. Comparative study of label and label-free techniques using shotgun proteomics for relative protein quantification. *J Chromatogr B Analyt Technol Biomed Life Sci.* **2013**, 928, 83-92

(55) Ramus C., Hovasse A., Marcellin M., Hesse A. M., Mouton-Barbosa E., Bouyssié D., Vaca S., Carapito C., Chaoui K., Bruley C., Garin J., Cianférani S., Ferro M., Van Dorssaeler A., Burlet-Schiltz O., Schaeffer C., Couté Y, Gonzalez de Peredo A. Benchmarking quantitative label-free LC-MS data processing workflows using a complex spiked proteomic standard dataset. *J Proteomics.* **2015**, 132, 51-62

(56) Luo R, Colangelo CM, Sessa WC, Zhao H. Bayesian Analysis of iTRAQ Data with Nonrandom Missingness: Identification of Differentially Expressed Proteins. *Stat Biosci.* **2009** 1(2), 228-245

(57) Cappadona S., Baker P. R., Cutillas P. R., Heck A. J., van Breukelen B. (2012) Current challenges in software solutions for mass spectrometry-based quantitative proteomics *Amino Acids* **2012**, 43(3), 1087-108

(58) Kumar C., Mann M. Bioinformatics analysis of mass spectrometry-based proteomics data sets. *FEBS Lett.* **2009**, 583(11), 1703-12

(59) Clough T., Thaminy S., Ragg S., Aebersold R., Vitek O. Statistical protein quantification and significance analysis in label-free LC-MS experiments with complex designs *BMC Bioinformatics* **2012**, 13

(60) Spooncer E., Brouard N., Nilsson S. K., Williams B., Liu M. C., Unwin R. D., Blinco D., Jaworska E., Simmons P. J., Whetton A. D. Developmental fate determination and marker discovery in hematopoietic stem cell biology using proteomic fingerprinting *MCP* **2008**, 7(3), 573-8

(61) Fawcett T. An introduction to ROC analysis *Pattern Recognition Letters* **2006**, 27, 861–874

(62) Gan C. S., Chong P. K., Pham T. K., Wright P. C. Technical, experimental, and biological variations in isobaric tags for relative and absolute quantitation (iTRAQ) *J Proteome Res* **2007**, 6(2), 821-7

(63) Christoforou A., Lilley K. S. Taming the isobaric tagging elephant in the room in quantitative proteomics *Nat Methods* **2011**, 8(11), 911-3

(64) Ow S. Y., Salim M., Noirel J., Evans C., Rehman I., Wright P. C. iTRAQ underestimation in simple and complex mixtures: "the good, the bad and the ugly" *J Proteome Res* **2009**, 8(11), 5347-55

(65) Shirran S.L., Botting C.H. A comparison of the accuracy of iTRAQ quantification by nLC-ESI MSMS and nLC-MALDI MSMS methods *J Proteomics* **2010**, 73(7), 1391-403

(66) Ting L., Rad, R., Gygi S. P., Haas, W. MS3 eliminates ratio distortion in isobaric labeling-based multiplexed quantitative proteomics *Nature Methods* **2011**, 8(11), 937–940

(67) McAlister G. C., Nusinow D. P., Jedrychowski M. P., Wühr M., Huttlin E. L., Erickson B. K., Rad R., Haas W., Gygi S. P. MultiNotch MS3 Enables Accurate, Sensitive, and Multiplexed Detection of Differential Expression across Cancer Cell Line Proteomes *Analytical Chemistry* **2014**, 86(14), 7150–7158

(68) Abdallah C, Sergeant K, Guillier C, Dumas-Gaudot E, Leclercq CC, Renaut J. Optimization of iTRAQ labelling coupled to OFFGEL fractionation as a proteomic workflow to the analysis of microsomal proteins of Medicago truncatula roots. *Proteome Sci.* **2012**, 10(1), 37

(69) Goeminne LJ, Argentini A, Martens L, Clement L. Summarization vs Peptide-Based Models in Label-Free Quantitative Proteomics: Performance, Pitfalls, and Data Analysis Guidelines. *J Proteome Res.* **2015** 14(6), 2457-65

22672774; PubMed Central PMCID: PMC3442994.

(70) Ratcliffe J. F. The effect on the t distribution of non-normality in the sampled population. *Appl. Stat.* **1968**, 17, 42–48

(71) Sawilowsky S. S., Blair R. C. A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychol. Bull.* **1992**, 111, 352–60

(72) Lumley T., Diehr P., Emerson S., Chen L. The importance of the normality assumption in large public health data sets. *Annu Rev Public Health* **2002**, 23, 151-69

Authors are required to submit a graphic entry for the Table of Contents (TOC) that, in conjunction with the manuscript title, should give the reader a representative idea of one of the following: A key structure, reaction, equation, concept, or theorem, etc., that is discussed in the manuscript. Consult the journal’s Instructions for Authors for TOC graphic specifications.

For TOC only