A hierarchical statistical modeling approach to analyze proteomic isobaric tag for relative and absolute quantitation data

Cong Zhou1#, Michael J Walker2#, Andrew J K Williamson2, Andrew Pierce2, Carlo Berzuini3, Caroline Dive1* and Anthony D Whetton2*

1Clinical and Experimental Pharmacology Group, CRUK Manchester Institute, University of Manchester, Wilmslow Road, Manchester M20 4BX, United Kingdom.
2Stem Cell and Leukaemia Proteomics Laboratory, Institute of Cancer Sciences, Manchester Academic Health Science Centre, Wolfson Molecular Imaging Centre, University of Manchester, M20 3LJ, United Kingdom.
3Centre for Biostatistics, Institute of Population Health, University of Manchester, Oxford Road, Manchester, M13 9PL, United Kingdom.

Running head: modeling iTRAQ variation to identify protein changes

ABSTRACT

Motivation: Isobaric tag for relative and absolute quantitation (iTRAQ) is a widely used method in quantitative proteomics. A robust data analysis strategy is required to determine protein quantification reliability, i.e., changes due to biological regulation rather than technical variation, so that proteins that are differentially expressed can be identified.

Methods: samples were created by mixing 5, 10, 15 and 20 ug E.coli cell lysate with 100 ug of cell lysate from mouse, corresponding to expected relative fold changes of 1 for mouse proteins and from 0.25 to 4 for E.coli proteins. Relative quantification was carried out using 8 channel isobaric tagging with iTRAQ reagent and proteins were identified using a TripleTOF 5600 mass spectrometer. Technical variation inherent in this iTRAQ dataset was systematically investigated.

Results: A hierarchical statistical model was developed to utilize quantitative information at peptide level and protein level simultaneously to estimate variation present in each individual peptide and protein. A novel data analysis strategy for iTRAQ, denoted in short as WHATraq, was subsequently proposed with its performance evaluated by the proportion of E.coli proteins that are successfully identified as differentially expressed. Compared with two benchmark data analysis strategies WHATraq was able to identify at least 62.8% more true positive proteins that are differentially expressed. Further validated using a biological iTRAQ dataset including multiple biological replicates from varied murine cell lines, WHATraq performed consistently and identified 375% more proteins as being differentially expressed among different cell lines than the other data analysis strategies.

Contact: cdive@picr.man.ac.uk & tony.whetton@manchester.ac.uk

Supplementary Information: Supplementary data are available at Bioinformatics online.

#Authors contribute equally.
*To whom correspondence should be addressed.

1 INTRODUCTION

Over the past decade, mass spectrometry (MS) has become the method of choice for large scale proteomics research to identify and quantify proteins present in different biological contexts (Domon and Aebersold, 2006; Gingras, et al., 2007). Several techniques including isotopic labeling and label free methods have been developed in association with MS to monitor quantitative changes in protein expression and post-translational modification states (Gygi, et al., 1999; Ong, et al., 2002; Ross, et al., 2004; Thompson, et al., 2003). The isobaric tag for relative and absolute quantitation (iTRAQ) approach is particularly of interest among isotopic labeling methods due to its ability to analyze in vivo samples and to compare up to 8 samples simultaneously (Ross, et al., 2004).

In general terms, the goal of an iTRAQ experiment is to identify differentially expressed peptides/proteins as a function of different biological conditions. There is a requirement for sophisticated methodologies to define significance for the observed changes in relative quantification, so that changes as a result of biological regulation can be distinguished from those as a result of technical issues. Value to the investigators comes from focusing on identifying differentially expressed proteins as we describe below, yet identifying differentially expressed peptides is a much simpler process. A fundamental assumption for all the discussion is that peptides have been identified and iTRAQ reporter ion intensities are available for these peptides.

Analysis of iTRAQ data starts from the processing of raw MS signal to match spectra to peptides and extracting the areas under iTRAQ reporter ion peaks, followed by empirical or statistical methods to identify differentially expressed proteins. Early efforts reported for iTRAQ data analysis included applying ad hoc thresholds to the observed quantitative changes(Unwin, et al., 2006; Williamson, et al., 2008), and using analysis of variance (ANOVA) to identify significant changes after fitting the observed changes with a linear model (Hill, et al., 2008; Oberg, et al., 2008).
It was proposed more recently that significance of the observed changes can be defined based on the standard deviation of technical variation, which can be estimated using technical replicates (Zhou, et al., 2012). The most recent advance in analyzing iTRAQ data came from the observation that peptides with lower iTRAQ reporter ion intensities tend to suffer from larger technical variation (Breitwieser, et al., 2011; Hultin-Rosenberg, et al., 2013; Hundertmark, et al., 2009; Karp, et al., 2010; Mahoney, et al., 2011; Zhang, et al., 2010). Many approaches have been proposed to address the observed variance heterogeneity, and as a result to achieve more accurate protein quantities (Gan, et al., 2007; Hu, et al., 2006; Hultin-Rosenberg, et al., 2013; Lin, et al., 2006; Onsongo, et al., 2010) and an improved capability to identify differentially expressed proteins (Breitwieser, et al., 2011; Karp, et al., 2010).

A novel hierarchical statistical modelling approach was developed in this work to systemically evaluate technical variation inherent in iTRAQ data, in order that variation inherent in relative quantification can be estimated for each individual peptide and protein. Statistical models currently available considered technical variation either as a single random variable (Breitwieser, et al., 2011; Hill, et al., 2008; Oberg, et al., 2008) or a discrete variable with empirical values (Hultin-Rosenberg, et al., 2013; Onsongo, et al., 2010), based on quantitative information from peptides only. In this work, however, variation in relative quantification was modelled as coming from three different types of sources, namely instrument noise, bias on peptides and bias on proteins. The developed hierarchical model enabled estimation of these variations by simultaneously taking quantitative information from both peptides and proteins into account. A novel data analysis strategy was then proposed based on the hierarchical statistical model, and was denoted as Workflow for Hierarchical Analysis of iTRAQ datasets (in short as WHATraq).

Spiked-in proteins have often been used to verify data analysis strategies (Breitwieser, et al., 2011; Hundertmark, et al., 2009; Karp, et al., 2010; Mahoney, et al., 2011). However, the limited number of proteins involved makes it extremely difficult to represent iTRAQ datasets generated using biological samples.Datasets using cell line samples with pre-defined quantities, i.e., known iTRAQ ratios, have also been generated (Hultin-Rosenberg, et al., 2013; Zhang, et al., 2010). While these datasets were sufficient to explore variation in iTRAQ, they were not ideal to evaluate the capability of a data analysis strategy to identify a discrete subset of proteins that are differentially expressed. In this work, a dataset was designed to simulate a real-world biological system under regulation, in which a proportion of proteins are differentially expressed and the remaining proteins keep unchanged. Such a dataset was generated by mixing identical amount of mouse cell lysate and varied pre-defined amount of E.coli cell lysate then labeling with different isobaric iTRAQ tags. Each sample was analyzed in duplicates to allow investigation of technical variation within the dataset. The dataset allowed easy evaluation of data analysis strategies in that only E.coli proteins were differentially expressed and the expected quantitative changes were known.

WHATraq was firstly validated using the E.coli/murine test dataset, with its performance evaluated in terms of true positive (TP) and false positive (FP) identifications. It was then further validated by analyzing a biological iTRAQ dataset including biological replicates from varied murine cell lines. The dataset was designed to further our understanding of the role of mutant isocitrate dehydrogenase gene (IDH) in the pathogenesis of acute myeloid leukemia (AML). We demonstrated in this analysis that although WHATraq was developed based on modelling technical variation, it can be readily applied on the analysis of iTRAQ datasets including biological replicates. The performance of WHATraq was compared with two benchmark data analysis strategies currently available (Isobar and STDP, see Methods). It was observed that WHATraq performed markedly better in discriminating differentially expressed proteins from technical/biological variation, and was capable of identifying larger number of differentially expressed proteins with improved accuracy.

2 METHODS

Experimental protocol of the first test dataset (E.coli/murine)

Both murine and E.coli cells were lysed in 0.5 M TEAB, 0.1% (w/v) SDS before trypsin proteolytic digestion and isobaric tag (iTRAQ) labeling was carried out as previously published (Pierce, et al., 2008) following the experimental scheme for label use (Table 1). The dataset comprised four pairs of technical replicates and the expected fold changes were: 0.25, 0.33, 0.5, 0.67, 0.75, 1.33, 1.5, 2, 3 and 4. Prior to MS peptide fractions were fractionated off line using a reversed phase chromatography column (Agilent Zorbax extend-C18 3.5 µm particle size, 4.6 mm x 15 cm. Reversed phase column) at high pH using an Agilent LC system. The gradient was run at 700 µl/min using initially 99.5% (v/v) high pH buffer A (0.1% (v/v) Ammonium hydroxide, 0.5% (v/v) high pH buffer B (0.1% (v/v) Ammonium hydroxide, 99.9% (v/v) acetonitrile) up to 40 % (v/v) buffer B over 40 minutes. Twenty 15 second fractions were collected between 10 and 15 minutes. Each fraction was then dried under vacuum.

Peptide was reconstituted in 30 µl of loading buffer (0.1% (v/v) Acetonitrile, 0.1% (v/v) formic acid, 20 mM citrate) with a third loaded onto a nanoAQUITY UPLC system (Waters). A reverse phase nano LC column (Waters BEH130 C18, 1.7 µm, 75 µmx25cm) with a 90 minute gradient running from 0.3% (v/v) acetonitrile, 0.1% (v/v) formic acid to 40% (v/v) acetonitrile, 0.1% (v/v) formic acid separated peptides for analysis. MS was carried out by a TripleTOF 5600 (Absciex) set up to analyse the top 20 ions by MS/MS per an MS scan. The MS scanned between 350 and 1250 with an accumulation time of 250 ms. Ions were only selected for MS/MS if they were over 150 counts per a second and had a charge state of between 2 and 5, ions previously selected were excluded for 30 seconds. The MS/MS was carried out in high sensitivity mode with 100 ms accumulation time and a rolling collision energy based upon mass and charge with a spread of 20. The MS/MS scanned between 100 and 1600.

Table 1. Experimental and labelling strategy for the E.coli/murine dataset

<table>
<thead>
<tr>
<th>E.colA Protein (µg)</th>
<th>Mouse Protein (µg)</th>
<th>iTRAQ label</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>100</td>
<td>113, 114</td>
</tr>
<tr>
<td>20</td>
<td>100</td>
<td>115, 116</td>
</tr>
<tr>
<td>15</td>
<td>100</td>
<td>117, 118</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>119, 121</td>
</tr>
</tbody>
</table>

The iTRAQ labelling strategy of the E.coli/murine dataset was listed in the table. The dataset was designed to comprise 4 pairs of duplicate iTRAQ channels. Each iTRAQ channel contained an identical amount of mouse cell lysate and varied amount of E.coli lysate.

Experimental protocol of the second test dataset (IDH)
Murine Ba/F3 cells were transfected with either MSCV-GFP retroviral vector, or MSCV-GFP containing wild type IDH (wild IDH) or mutant (R132H) IDH as previously described (Pierce, et al., 1998). The resultant cell lines were maintained in Fischers medium (Invitrogen) with 10% (v/v) horse serum (Gibco) supplemented with 5% mIL-3 (conditioned media from X63-Ag-653 cells).

Nuclear proteins were enriched as previously described (Pierce, et al., 2012) then subjected to tryptic enzymatic digestion, iTRAQ labelling and processing as performed above. Quantitative data on the effect of IDH mutant activity on protein expression was obtained by 8 channel iTRAQ isobaric tagging of biological replicates of Ba/F3, MSCV, wtIDH and mutant IDH. The 8 channel iTRAQ tagging experiment were repeated twice to obtain data on four biological replicates.

**Protein identification**

All MS/MS data were submitted to ProteinPilot software version 4.2 (AbSciex) for database searching and iTRAQ reporter ion quantification. Searches were performed twice against the Ensembl *mus musculus* core 63 database and Ensembl *E.coli* K12-111-database. Cys alkylation with methanethiosulfate (MMTS) and trypsin as the digestion enzyme were specified in the search. Biological modifications and amino acid substitutions were also permitted. The false discovery rate (FDR) of protein identification was controlled using a target-decoy searching strategy (Elias and Gygi, 2007) where forward and reverse sequences from a database were in equal competition to be the highest ranking identification for each spectrum. The q-value approach (Storey and Tibshirani, 2003) was then applied to define a cut-off for peptide confidence so that the control criteria of FDR can be met. The maximum allowed FDR for protein identification was set to 1%.

**Statistical modelling of variation present in iTRAQ datasets**

In general terms, the purpose of modelling an iTRAQ dataset was to estimate the technical variation inherent in the dataset so that the quality of the quantitative information provided by this dataset can be evaluated. As a result, genuine changes in protein expression levels could be discriminated from technical variation. In this section, a novel statistical modelling approach was briefly described, with more detail of the modelling in Supplementary material 1.

Denote the observed iTRAQ reporter ion intensity of a peptide as $I_o$, the genuine (technical variation free) but unknown iTRAQ reporter ion intensity as $I$. It was observed that $I$ is heteroscedastic (Supplementary Figure 1) and this observation was in agreement with literature. Therefore the following multiplicative model can be assumed:

$$I_o = I \cdot \epsilon'$$

(1)

where $\epsilon'$ represents technical variation.

In this study, technical variation was modelled to come from three different types of sources, namely, instrument noise, bias on peptides and bias on proteins:

$$t = \epsilon + b_{pep} + b_{pro}$$

(2)

where $\epsilon$ represents instrument noise, which refers to technical variation caused by imprecision of instruments. The symbol $b_{pep}$ represents bias on peptides, and is used to model variation caused by the characteristics of each individual peptide. Precursor co-elution typically falls into this category. Similarly, $b_{pro}$ represents bias on proteins and is used to model variation caused by the characteristics of each individual protein. It is worth noting that $b_{pep}$ and $b_{pro}$ are channel-dependent, meaning that they have different values in different iTRAQ channels for a same peptide. For more detail on the validity of this model please refer to supplementary material.

Naturally, $\epsilon$ can be modelled as a random variable following a Gaussian distribution:

$$\epsilon = f(I) \cdot \epsilon'$$

$$\epsilon' \sim N(0,1)$$

(3)

where $\epsilon'$ is a standard normal distribution. In Eq. 3, the variance of instrument noise is modelled as a function of $I$ to address the observed heteroscedasticity. The form of function $f(I)$ is intended to be determined by the observed data. Although $b_{pep}$ and $b_{pro}$ are not random variables, their impact on a dataset with sufficient amount of peptides/proteins can be modelled as following Gaussian distributions:

$$b_{pep} \sim N(0, \sigma_{pep}^2)$$

$$b_{pro} \sim N(0, \sigma_{pro}^2)$$

(4)

This assumption was made based on the fact that $b_{pep}$ and $b_{pro}$ cannot be estimated accurately for a single peptide or protein, therefore they were approximated based on their average impact on the whole proteome.

Denote a protein being successfully quantified in a given iTRAQ dataset as $P_i$ ($i = 1 \ldots n$), where $n$ is the total number of proteins, and the peptides belonging to this protein as $p_{ij}$ ($i = 1 \ldots n, j = 1 \ldots k_i$), where $k_i$ is the total number of peptides being successfully quantified for protein $P_i$. The relative quantification of $p_{ij}$, measured as log ratio of its iTRAQ reporter ion intensities under two different biological states, can be expressed as:

$$\log \left( \frac{I_{o_{ij}}}{I_{o_{ij}'}} \right) = \log \left( \frac{I_{o_{ij}}}{I_{o_{ij}'}} \right) + t_a - t_{b2}$$

With regard to relative quantification between technical replicates, in particular, the above equation can be expressed as:

$$r_{ij} = I_{o_{ij}} - I_{o_{ij}'} = \epsilon_{ij} - \epsilon_{ij}' + b_{pep_{ij}} - b_{pep_{ij}'} + b_{pro_{ij}} - b_{pro_{ij}'}$$

(5)

where $r_{ij}$ is relative quantification between technical replicates. The above equations indicate that technical variation inherent in an observed relative quantification value, as the information of our prime interest, can be estimated with the knowledge of $f(I)$, $\sigma_{pep}^2$ and $\sigma_{pro}^2$. These functions and variables could be determined using technical replicates present in an iTRAQ dataset.

A step-by-step approach was described in this section on how to derive the formula of $f(I)$ as well as the values of $\sigma_{pep}^2$ and $\sigma_{pro}^2$ from a given iTRAQ dataset. Full detail on how the approach was derived can be found in Supplementary material 1.

Step 1: modelling technical variation on peptides

Divide all peptides into multiple small groups within which peptides share close values of observed iTRAQ reporter ion intensities. For each group, calculate the average of all $I_o$ (denoted as $\bar{I}$) and the variance of all $r_{ij}$ (denoted as $\text{var}(r)$). The function between $\text{var}(r)$ and iTRAQ signal intensity was assumed to have an empirical formula as shown in Eq.4.

$$\text{var}(r) = 2a(\bar{I} + b)^2 + 2d$$

(4)

Parameter $a$, $b$ and $c$ in Eq.4 can be estimated using a maximum likelihood estimation (MLE) approach.

Step 2: reducing peptide redundancies

Merge the redundant peptides within a protein into one entry. Here redundant peptides refer to peptides sharing a same amino acid sequence and post-translational modification. For example, assume protein $P_i$ has $k_i$ unique peptides out of a total of $k_i$ peptides, and $m$ of them have redundant peptides. The starting and ending index of the redundant peptides are $s_u$ and $t_u$, $u=1...m$. They can be merged into one peptide entry denoted as $p_u$, using a weighted average approach:

$$r_{iu} = \sum_{j=s_u}^{t_u} \frac{w_j r_{ij}}{\sum_{j=s_u}^{t_u} w_j}$$

where $w_j$ is the weight applied on peptide $p_{ij}$. One option for $w_j$ is:

$$w_j = \frac{1}{a(x + b)^c}$$

(6)
Weights of the merged peptides should be updated accordingly. This step results in \( k' \) unique peptides for \( P_s \), each associate with its own relative quantification \( r_{ij}' \), and weight \( w_{ij}' \).

Step 3: deriving protein quantification using only unique peptides

Assemble the redundancy-free peptides from last step to derive relative quantification of proteins using a weighted average approach:

\[
R_i = \frac{\sum_{j=1}^{k} w_{ij} r_{ij}'}{\sum_{j=1}^{k} w_{ij}'}
\]

where \( R_i \) is the relative quantification between technical replicates of \( P_i \).

Step 4: Modelling technical variation on proteins

For an easy annotation, define function:

\[
z(i, j, s, t, x) = \frac{w_{ij}}{\sum_{j=1}^{k} w_{ij}}(x)
\]

Then the following model can be derived:

\[
\begin{align*}
\text{var}(R_i) &= \text{var}(R_{eij}) + \text{var}(R_{p,ij}) + \text{var}(R_{p,ij}) \\
R_{eij} &\sim N\left(0, 2\sigma_{pe}^2\right) \\
R_{p,ij} &\sim N\left(0, 2\sigma_{pe}^2 \left(\mu_i, 1, k_i, \sigma_{pro}^2\right)\right) \\
R_{ij} &\sim N\left(0, 2\sigma^2 \left(\mu_i, 1, k_i, \sigma_{pro}^2\right)\right)
\end{align*}
\]

In this equation, \( f \) introduced parameters \( a, b, c \) and \( d, a, b \) and \( c \) has been estimated in step 1, and the remaining parameter \( d \) together with \( \sigma_{pe}^2 \) and \( \sigma_{pro}^2 \) can be estimated using a MLE approach with the protein relative quantification calculated in step 3.

Step 5: Estimate technical variation and assign significance for relative changes of a given protein

Using the model derived above, technical variation on the relative quantification of a given protein can be estimated:

\[
\begin{align*}
\text{var}(R_i) &= \text{var}(R_{eij}) + \text{var}(R_{p,ij}) + \text{var}(R_{p,ij}) \\
\text{var}(R_{eij}) &= \sum_{s} \sum_{t} z(i, j, s, t, x) \sum_{s} \sum_{t} z(i, j, s, t, x) \\
\text{var}(R_{p,ij}) &= 2\sigma_{pro}^2 \\
\text{var}(R_{p,ij}) &= 2\sigma_{pro}^2 \left(\mu_i, 1, k_i, \sigma_{pro}^2\right) \\
\end{align*}
\]

Step 6: Identify proteins that are differentially expressed

Statistical significance of observed protein changes can be assigned based on the technical variation estimated in step 5, in forms of \( p \)-values. They are subject to FDR controls to identify proteins that are differentially expressed.

**WHATraq: a novel data analysis strategy for iTRAQ**

Successful estimation of technical variation allowed a statistical significance \( (p\text{-value}) \) to be assigned to the quantitative change of each observed protein. A cut-off value \( (q\text{-value}) \) can be derived so that the expected FDR can be controlled to be below 5%. Proteins with \( p \)-values smaller than the cut-off were considered as being differentially expressed. A novel data analysis strategy, denoted as Workflow for Hierarchical Analysis of iTRAQ datasets (WHATraq), was developed based on the proposed hierarchical statistical model and its complete workflow was illustrated in Figure 1. The strategy required an iTRAQ dataset to include at least two iTRAQ channels to be technical/biological replicates, and should have identification and iTRAQ reporter ion intensity data available for each peptide. It was capable of identifying proteins as being differentially expressed. A detailed description of each step within the workflow can be found in Supplementary material 2. WHATraq was implemented in house using Matlab R2009a (The Mathworks, Inc.). The source code and the E.coli/murine dataset (database search results), has been made publicly available at http://www.scalpl.org/-czhou/program code WHATraq.zip

**Performance evaluation of WHATraq:**

The first test dataset (E.coli/murine) included defined fold changes ranging from 0.25 to 4, i.e., from 4 fold down regulation to 4 fold up regulation. Bigger changes were not included because they could easily be identified by any data analysis strategy (Ting, et al., 2011). Performance evaluation was carried out by using WHATraq to identify proteins that were differentially expressed. The E.coli proteins identified as being differentially expressed were considered as true positives (TP) and the identified mouse proteins were considered as false positives (FP). E.coli proteins with expected relative changes of 1.5, 2, 0.5 and 4 folds were used in this study.

A novel data analysis strategy denoted as WHATraq has been developed for iTRAQ based on the proposed hierarchical statistical model. It provided a complete workflow starting from identification and iTRAQ reporter ion intensity data accumulation for each peptide, followed by a parallel modelling of technical variation on both peptide and protein levels. Technical variation on a specific peptide/protein can be estimated using the derived models and can subsequently be used to identify proteins that were differentially expressed.

We next applied WHATraq to a second iTRAQ test dataset (IDH) which comprised multiple biological replicates. Important assumptions for the analysis of this dataset were described in the next section. For this dataset, we reported only the number of proteins that were identified as being differentially expressed because it was not feasible to discriminate TP with FP. The dataset was designed to explore proteomic differences between wtIDH and mutant IDH, which was of potential value in understanding the role of mutant IDH in leukaemogenesis (Dang, et al., 2009; Figueroa, et al., 2010). We also assessed the effects of expressing an "empty" retroviral vector as compared to a non-transfected control to understand any artefacts that may arise from retroviral mediated gene transduction.

The performance of WHATraq was compared with two other benchmark strategies. The first one modelled technical variation of all the observed proteins as following a normal distribution, and the statistical significance of an observed change was assigned based on the standard deviation of protein technical variation, as described in literature (Simpson, et al., 2013; Zhou, et al., 2012). This strategy was denoted as STDP (standard deviation of proteins). The second one, named *Isobar* (Breitwieser, et al., 2011),
modelled technical variation of a specific peptide as an exponential function of its iTRAQ reporter ion intensity. It is an application modelling peptide heteroscedasticity and has been shown in its original publication to be a state-of-the-art strategy that was able to identify more differentially expressed proteins than data analysis strategies developed earlier. STDP was implemented in-house to have the same workflow as WHATraq except for the statistical model applied. Isobar was downloaded from its developer (version 1.6.2) and installed in R (version 3.0.1).

Applying WHATraq on biological replicates

Although WHATraq was originally developed to model technical variation from technical replicates, it can be readily applied on biological replicates resulting in an estimation of technical variation + biological variation, with assumptions and alteration described in this section:

1. It was assumed that the biological variation of all proteins follows one unique distribution, e.g. a normal distribution. This assumption enabled biological variation to be modelled as a part of bias on peptide/proteins (see Results section).

2. In the original workflow of WHATraq (Figure 1), peptides with extremely large variation will be removed from contribute to protein quantification. This step was abandoned when biological variation was modeled because biological variation was more likely to induce extremely large relative changes.

3. According to the original definition, bias on peptides / proteins, are not random variables and therefore will not change as a result of averaging multiple replicates. If biological variations are modeled as a part of the biases, this definition becomes invalid and a proportion of biases will indeed be reduced by averaging multiple replicates. Unfortunately it is impossible to estimate the proportionality because biological replicates don’t allow discrimination of technical variation and biological variation. If it is assumed that biological variation is dominant, biases should be treated as random and their standard deviations are reduced according to the number of replicates. Otherwise, it is safe to assume that biological and technical variation are in a 1:1 ratio and therefore half of the biases will be unchanged and the other half will be reduced according to the number of replicates.

Normalizing iTRAQ channels for FDR control

We demonstrated that technical variations can be channel-specific (Results section). Therefore using technical variation estimated from one pair of iTRAQ channels to identify differentially expressed proteins in other iTRAQ channels, a typical scene in analyzing iTRAQ datasets, may lead to an unexpectedly high number of FP identifications. Here we developed a normalization algorithm to help control FDR. We can denote $c1$ and $c2$ as technical replicates that were used to estimate technical variation, $P_{00.1.2}$ and $P_{00.3.4}$ as the 40th and 60th quantile of the log ratio between $c1$ and $c2$. Similarly we can denote $c3$ and $c4$ as the iTRAQ channels to calculate protein relative quantification, and $Q_{00.3.4}$ and $Q_{00.1.2}$ as the 40th and 60th quantile of the log ratio between $c3$ and $c4$. The normalization algorithm is developed based on the commonly accepted assumption that the majority of proteins will not be differentially expressed, i.e., $P_{00.1.2} = P_{00.3.4} = Q_{00.3.4} = Q_{00.1.2}$.

Define normalization term $N = \frac{Q_{00.3.4} - Q_{00.1.2}}{P_{00.1.2} - P_{00.3.4}}$.

Standard deviation of protein relative quantifications from $c3$ and $c4$ needed to times the normalization term $N$ for a more accurate estimation on technical variation and subsequently for a better FDR control.

3 RESULTS

3.1 General characteristics of the first test dataset

In the E.coli/murine test dataset peptide spectral matches (PSMs) identified to come from mouse proteins were 337042 in number, and the other 82301 PSMs were identified to come from E.coli proteins. PSMs with peptide sequences shared by both species were removed, resulting in a total of 336745 valid PSMs. With the FDR control using the q-value approach, the PSMs translated into 2938 mouse proteins and 750 E.coli proteins, of which 2837 mouse proteins and 741 E.coli proteins had valid quantitative information. These proteins were listed in Supplementary table 1.
The dataset was designed to comprise 4 pairs of technical replicates. Protein quantification was calculated based on the average of iTRAQ reporter ion intensities from technical replicates. The observed changes for peptides and proteins were plotted in Figure 2, with expected relative changes of 1.5, 2, 0.5 and 4 fold respectively. A considerable amount of technical variation was observed in peptides but it was significantly reduced in proteins. The observed changes on E.coli peptides and proteins were notably smaller than expected, representing approximately 40% of the expected changes (log space). The observation indicated that iTRAQ as a quantification technique tends to underestimate biological changes, which was in agreement with literature (Karp, et al., 2010; Mahoney, et al., 2011). The observed underestimation was not improved by using more stringent peptide selection criteria or aggregation into proteins.

3.2 Statistical modelling of technical variation

The relationship between technical variation and iTRAQ reporter ion intensities, peptide masses and peptide retention times were explored using the E.coli/murine dataset. iTRAQ reporter ion intensities was the only factor found to have dependency on technical variation (Supplementary Figure 1-a,b,c). Modelling the dependency using Eq.4 (Methods) enabled derivation of 95% confidence interval (CI) of the observed technical variation, as illustrated with red solid curves in Figure 3-a. The close matches of the CIs with the empirical 95th quantile of the observed technical variation (blue dashed curves) confirmed that the proposed statistical model provided an accurate estimation of the technical variation on peptide level. Technical variation on proteins was modelled to come from three origins: instrument noise, bias on peptides and bias on proteins. Their contribution to the total variation was demonstrated in Figure 3-b. Instrument noise, as indicated by the red lines, was the highest source of technical variation for proteins with low numbers of quantified peptides, but its impact quickly reduced for proteins with increased number of quantified peptides. Bias on proteins (blue lines), however, was stable for all proteins and therefore was dominant for proteins with high numbers of quantified peptides. Bias on peptides, as demonstrated with green lines, showed moderate impact for proteins with more than 5 peptides. In general terms, the model assigned smaller expected technical variation on proteins with higher number of peptides, making them easier to be identified as being differentially expressed. This can be interpreted as changes in protein levels with more data accrued are more likely to be genuine.

Variation models were built up for each individual pair of technical replicates. As illustrated in Figure 4, all iTRAQ channels demonstrated similar level of instrument noise, indicating that the MS instrument performed consistently in the experiment. Bias on peptides and bias on proteins, on the contrary, were clearly channel-specific. iTRAQ channel 117/118 exhibited considerably smaller biases than the other channels (35.0% and 58.8% smaller than the average bias level). These observations indicated that using technical variation estimated from one pair of iTRAQ channels to identify differentially expressed proteins in other iTRAQ channels may lead to unexpected high number of FP identifications. Normalization algorithms such as the one proposed in the Methods section has to be applied in order to identify a reliable list of proteins that are differentially expressed.
Variation models were built up using each individual technical replicates. It was found that the instrument noise* have similar values in all iTRAQ channels, whereas the bias on peptides and bias on proteins are considerably smaller in iTRAQ channel 117 and 118 compared to the other channels.

* Instrument noise was calculated at iTRAQ reporter ion intensity equals to 1000.

### 3.3 Performance evaluation of WHATraq

The proposed data analysis strategy denoted as WHATraq was used to identify differentially expressed proteins in the E.coli/murine dataset, together with two benchmark data analysis strategies denoted as STDP and Isobar. The expected fold changes of the differentially expressed proteins were 1.5, 2, 0.5 and 4. The performance of all the data analysis strategies was evaluated in terms of true positive rate and false positive rate as listed in Table 2. Receiver operating characteristic (ROC) curves were plotted to provide an overview of their performance (Figure 5).

#### Table 2. Comparison of TPR and FPR

<table>
<thead>
<tr>
<th></th>
<th>1.5 fold up regulation</th>
<th>2 fold up regulation</th>
<th>2 fold down regulation</th>
<th>4 fold up regulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHATraq</td>
<td>14 (1.8%)</td>
<td>304 (40.9%)</td>
<td>217 (29.2%)</td>
<td>464 (62.8%)</td>
</tr>
<tr>
<td>STDP</td>
<td>1 (0.1%)</td>
<td>47 (6.3%)</td>
<td>18 (2.4%)</td>
<td>282 (38.0%)</td>
</tr>
<tr>
<td>Isobar</td>
<td>1 (0.1%)</td>
<td>139 (18.7%)</td>
<td>3 (0.4%)</td>
<td>285 (38.4%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1.5 fold up regulation</th>
<th>2 fold up regulation</th>
<th>2 fold down regulation</th>
<th>4 fold up regulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHATraq</td>
<td>2 (0.1%)</td>
<td>23 (0.8%)</td>
<td>9 (0.3%)</td>
<td>47 (1.7%)</td>
</tr>
<tr>
<td>STDP</td>
<td>9 (0.3%)</td>
<td>19 (0.7%)</td>
<td>17 (0.6%)</td>
<td>33 (1.2%)</td>
</tr>
<tr>
<td>Isobar</td>
<td>4 (0.2%)</td>
<td>13 (0.5%)</td>
<td>7 (0.3%)</td>
<td>19 (0.7%)</td>
</tr>
</tbody>
</table>

The proposed strategy WHATraq and STDP were both implemented in house whereas Isobar was downloaded from its developer. The results listed in the table were achieved by setting the expected FDR to 0.05.

WHATraq demonstrated a marked advantage over STDP and Isobar (Figure 5). In general term, all data analysis strategies performed better on proteins with bigger expected changes, in the context of TP proteins they were able to identify. For proteins with 1.5 fold expected changes, STDP and Isobar failed to deliver meaningful results with only 1 TP protein being identified and FDR over 400% (Table 2). This is because protein changes were too small to be separated from the dominant technical variation, as demonstrated by Figure 2-b (top left panel). WHATraq struggled but was still able to control FDR properly with 14 TP proteins and 2 FP proteins identified. A clear advantage from WHATraq was observed in proteins with larger changes. For proteins subjected to 2 fold down regulation, both STDP and Isobar failed to identify any proteins whereas WHATraq successfully identified 217 TP proteins and more importantly controlled the FDR to 4.0%. Similarly, for proteins with 2 fold and 4 fold up regulation WHATraq identified no less than 118.7% and 62.8% more TP proteins than Isobar and STDP at the same FDR level.

In general terms, proteins with larger number of quantified peptides, proteins with constituent peptides that associated with higher iTRAQ reporter ion intensities, and proteins with less redundant peptides were modelled to have lower technical variation, and changes in these proteins could be identified more readily. For example, an E.coli protein with an observed 1.29 fold change (EBESCP00000000712) was identified as differentially expressed, thanks to its 58 peptides being successfully quantified. The observed fold change was considerably underestimated (expected fold change 2), demonstrating that the proposed data analysis strategy was very sensitive in detecting biological changes. On the other hand, many proteins with big observed changes such as a mouse protein (ENSMUSP00000045111) bearing 1.64 fold changes, were successfully identified as associated with technical variation, indicating that the proposed data analysis strategy had an outstanding specificity on identifying differentially expressed proteins.

A full list of the differentially expressed proteins being identified could be found in Supplementary table 2.

### Fig. 5. ROC curve analysis of identifying differentially expressed proteins

The ROC curves of WHATraq were illustrated using red solid lines, and the red squares on the line represented the results listed in Table 2. Similarly, the ROC curves of STDP were illustrated using blue dashed lines, with the data analysis results represented using the blue squares. The ROC curves of Isobar were illustrated using green dotted lines and the data analysis results were demonstrated using green squares.

### 3.4 Data analysis of the IDH dataset

This dataset was designed based on the observation that point mutations in the IDH have been detected in 15%-20% of newly diagnosed AML patients examined. The mutations confer novel enzymatic activity on IDH facilitating the reduction of α-ketoglutarate...
To d-2-hydroxyglutarate (2HG) thus affecting the epigenetic state of the cells but the biologic outcomes can vary with respect to cell growth (Stepanenko, et al., 2013). We have expressed mutant and wild type IDH in the Ba/F3 cell line. Measurement of 2HG in cells expressing wild type and mutant IDH confirmed novel enzymatic activity of the mutant IDH (1440 ng/ml +/- 129 compared to 6933 ng/ml +/- 1022, n=3). Furthermore, cells expressing mutant IDH had a lower colony forming efficiency than those expressing wild type IDH (47 +/- 4 versus 60 +/- 4 colony forming cells per 100 cells plated respectively, n=3, p=0.01).

To further our understanding of the potential effect of mutant IDH on the proteome we undertook a proteomic assessment of the consequences of wild type and mutant IDH expression respectively in a cell line model. A total of 1742751 PSMs were identified from the IDH dataset. Processed using the method described in the Methods section, the PSMs translated into 5572 murine proteins with valid quantitative information (Supplementary Table 3). WHATraq was applied to model variation between biological replicates, enabling estimation of a combination of biological and technical variation for the relative quantification of each protein. The resulting model clearly demonstrated that the total variation was dominated by bias on peptides and bias on proteins, a pattern different with that in the E.coli/murine dataset (Figure 6). Further analysis showed that the instrument noise of the two test datasets were very close in value whereas the bias on peptide and bias on proteins calculated from the IDH dataset nearly double the amount of those calculated from the E.coli/murine dataset. It was reasonable to conclude that the additional biases were transformed from biological variation. The number of proteins that were differentially expressed between different cell lines is listed in Table 3. Similar to the results of the E.coli/murine dataset, WHATraq identified considerably more proteins (375%) as being differentially expressed than both STDP and Isobar.

Table 3. Number of proteins identified as being differentially expressed

<table>
<thead>
<tr>
<th>Method</th>
<th>wtIDH / mutant IDH</th>
<th>Ba/F3 / MSCV</th>
<th>Total number</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHATraq</td>
<td>13</td>
<td>30</td>
<td>38</td>
</tr>
<tr>
<td>STDP</td>
<td>6</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>Isobar</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

The number of proteins that were identified as being differentially expressed between different murine cell lines were listed in the table. We are unable to discriminate TP with FP in this test dataset.

4 DISCUSSION

In this work, a novel data analysis strategy denoted as WHATraq was proposed for the analysis of iTRAQ proteomics datasets, allowing differentially expressed proteins to be identified. The strategy was primarily based on a novel hierarchical statistical model that was capable of estimating technical/biological variation inherent in relative quantification of any peptide/protein. Notably, the model benefited from making use of quantitative information at both peptide and protein levels, and estimated not only random noise as other published models do, but also bias present in peptides and proteins. Here bias refers to the non-random effect from factors such as noise from co-eluting peptides and nearly isobaric peptides. It was also demonstrated that biological variation can also be modeled as biases. WHATraq was shown to have superior ability in identifying proteins that were differentially expressed due to biological regulation, from changes due to technical noise compared to other benchmark data analysis strategies such as STDP and Isobar.

Biological replicates can be modeled using WHATraq resulting in an estimation of a combination of technical variation and biological variation. However, such analysis was carried out based on an assumption that the biological variation of all proteins within the sample belonged to one unique normal distribution. Though the assumption was valid for the majority of the proteins, indeed many proteins could exhibit extremely large biological variation. The invalidation of the assumption resulted in an over-estimation of bias on peptides and bias on proteins, which in turn reduced the number of differentially expressed proteins being identified. On the other hand, proteins with large biological variation can also be identified by mistake as being differentially expressed because their biological variation was under-estimated by the model. We would like to argue that including an additional pair of technical replicates to estimate technical variation may be the best approach to accurately identify proteins being differently expressed, because it allows biological variation to be estimated for each individual protein. In large scale studies that include multiple subjects being analyzed in multiple iTRAQ experiments, it was especially useful to include a pair of common reference samples (such as pooled samples) as technical replicates. This will allow comparison of

Fig. 6. Modelling a combination of technical and biological variation

The figure showed the statistical model of a combination of technical and biological variation at protein level. Bias on peptides and bias on proteins dominated the total variation. Comparison with the model built up on technical variation alone demonstrated that the values of instrument noise in the two models are very close, whereas bias on peptide/proteins nearly doubled with the additional biological variation.

* Instrument noise was calculated at iTRAQ reporter ion intensity equals to 1000.
protein expression level across different experiments, e.g. for the purpose of estimating inter-subject variation.

WHATraq is more likely to identify proteins with higher number of peptides as being differentially expressed. However, sometimes researchers are particularly interested in proteins with low expression level, which typically exhibit low number of peptides in an iTRAQ dataset. In this case, an additional analysis can be carried out to search for differentially expressed proteins with only a limited number of peptides, using the same procedure as standard analysis present in this study. Such an analysis will increase the global FDR to a limited extent, but provides a unique chance for proteins with low expression level to be identified as being differentially expressed. For example, the analysis on proteins with three peptides or less allowed identification of an additional five TP proteins (no FP) from the E.coli/murine dataset.

Some proteins were identified to be changing probably as a consequence of mutant IDH expression compared to the control cells expressing wild type IDH. We also showed proteins potentially expressed as a consequence of retroviral mediated gene transfer with an empty vector. Thus it was demonstrated that the methodology is applicable directly to real experiments.

It has been proposed in earlier publications that an intensity cut-off should be applied to eliminate peptides with low iTRAQ reporter ion intensities, because these peptides usually associate with high technical variation and provide inaccurate quantitative information (Hu, et al., 2006; Oberg, et al., 2008). In this study, we tested various cut-off values ranging from 6 to 230 which were derived using the technical variation model at peptide level, corresponding to the standard deviation of technical variation to be 1.5, 1.2, 1.0, 0.7 and 0.5 respectively. We discovered that applying intensity cut-offs provided no benefit to WHATraq probably because it has already taken the uncertainty from peptides with low iTRAQ reporter ion intensities into account (Supplementary Figure 2). Instead, applying intensity cut-offs can indeed decrease the FPR of data analysis strategies that do not have technical variation models at peptide level, such as STDP.

It has been shown that iTRAQ quantification often underestimates fold change due to several factors (Bantscheff, et al., 2008; Karp, et al., 2010; Ow, et al., 2009; Ow, et al., 2011). Underestimation is believed to primarily be a result of reporter ion mixing from multiple co-eluting peptides or nearly isoobaric peptides. This background effect has been reduced but not removed by increased fractionation of samples reducing fraction complexity; this however is at the expense of increased analysis time (Ow, et al., 2011). Another approach to eliminate the quantitative inaccuracy of isoobaric tagging techniques uses MS3 to add an additional fractionation step to target peptide signals (Ting, et al., 2011). This is only available on mass spectrometry instrumentation capable of higher order fragmentation events and so cannot be applied in many situations. Therefore a model which can differentiate between genuine changes and technical variation, such as the one proposed here, is important for the correct identification of low level changes in the proteome.

5 CONCLUSION

In this work, we proposed a novel data analysis strategy called WHATraq for quantitative proteomics experiments using iTRAQ. Based on a hierarchical statistical model the strategy is capable of estimating technical/biological variation for each individual peptide and protein, and subsequently identifying proteins that were differentially expressed. We demonstrated that the strategy was able to identify significantly larger number of proteins as being differentially expressed compared to two other benchmark data analysis strategies STDP and Isobar.

ACKNOWLEDGEMENTS

CZ was funded by Clinical Research Initiative Grant Code C357/A12197, MJW was funded by Experimental Cancer Medicine Centre Network, AJKW and AP were funded by Leukaemia Lymphoma Research UK, CB was partially supported by the FP7-305280 MIMOms European Collaborative Project, as part of the HEALTH-2012 INNOVATION scheme. CD was funded by Cancer Research UK core funding C5759/A12328 and Cancer Research UK Center Award A12197. ADW is funded by Leukaemia Lymphoma Research UK.

REFERENCES


Ow, S.Y., et al. (2009) iTRAQ underestimation in simple and complex mixtures: "the good, the bad and the ugly", J Proteome Res, 8, 5347-5355.


Ting, L., et al. (2011) MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics, Nat Methods, 8, 937-940.


