Journal of Integrated Omics

JOURNAL OF INTEGRATED OMICS

A METHODOLOGICAL JOURNAL

HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE

Assessing the Loss of Information through Application of the ‘Two-hit Rule’ in iTRAQ Datasets

Josephine Kilner[2], Liangjie Zhu[1], Saw Yen Ow[2], Caroline Evans[2,a], Bernard M. Corfe*[1,a].

[1]Department of Oncology, The University of Sheffield,The Medical School, Beech Hill Road, Sheffield, S10 2JF, UK; [2]Department of Chemical and Biological Engineering, ChELSI Institute, The University of Sheffield, Sheffield , Mappin Street, Sheffield, S1 3JD, UK; [a]These authors contributed equally to this work.

Received: 1 November 2010 **Accepted:** 17 November 2010 **Available Online**: 23 November 2010

ABSTRACT

High-throughput studies of complex protein mixtures using proteomic workflows typically employ tandem mass spectrometric analysis of peptides obtained by tryptic digestion. Protein identification is achieved by comparing the experimentally obtained peptide MS/MS spectra to theoretical spectra. Protein identifications based on peptide fragment sequences are often judged valid using the so called ‘two-peptide’ rule whereby any protein identified by sequencing of fragment ions must be justified by the identification of two sequence unique peptides from the same protein. This excludes proteins identified on the basis of a single peptide ‘hit’ (often termed a one-hit wonder, or OHW). Applying the ‘two hit’ stringency may result in the loss of potentially valuable meta-data: information yielded or consolidated by valid OHW proteins may be overlooked. This study tests the hypothesis that certain groups of OHW proteins (and thus related biological events or pathways) are more likely to be identified by single peptide due to various physical or biochemical characteristics (molecular weight and isoelectric point). We have undertaken analysis on data from three independent quantitative iTRAQ based proteomic studies of a human colon cell line and human colon tissue to correlate the differences between OHW and “valid” protein sets for molecular weight, isoelectric point and for associated biological pathways. The results show that there is a possible trend of inverse correlation between the pI value of a protein and the number of peptide hits for identification. Molecular weights range from 30-60 kDa. Pathway analysis using EBI-EMBL Reactome SkyPainter found that by excluding OHWs, several biological pathways were consistently not mapped, suggesting that exclusion of OHW potentially limits the understanding the biological processes potentially identified within the whole dataset. Future work should address strategies for evaluation of validity and reproducibility of these conclusions in other tissues.

Keywords: iTRAQ Mass-spectroscopy; One-hit Wonders; Protein Identification; GeneBio Phenyx search engine; Reactome-SkyPainter Pathway analysis.

Abbreviations

FDR false discovery rate; **iTRAQ LC MS/MS** Liquid-chromatography with tandem mass-spectroscopy using isobaric tag for relative and absolute quantitation; **MCP** Molecular & Cellular Proteomics; **MW** Molecular weight; **OHW** One-peptide wonders; **pI** Isoelectric point; **PSM** Peptide-spectrum matching; **SCFA** Short-chain fatty acidsReferences.

1. Introduction

The successful identification and relative quantification of the entire complement of proteins expressed by a whole cell or organism under certain conditions is a key goal of high throughput proteomics. The most common method of protein analysis is via the bottom-up approach, whereby enzymatically derived peptides (typically tryptic peptides) are analysed by mass spectrometry to determine their intact masses (MS) and the complement of ions from their dissociated fragments during gas phase fragmentation (MS/MS). Under these cases, detectable peptides (often called proteotryptic peptide) sequences from MS/MS spectra are matched against curated protein databases to give either positive or

***Corresponding author:** Dr. Bernard Corfe, Department of Oncology, The University of Sheffield,The Medical School, Beech Hill Road, SHEFFIELD, S10 2JF, UK; Email Adress: b.m.corfe@shef.ac.uk.

negative peptide identifications based on statistical and experimental supervised pattern matching criteria [1]. High throughput methods inevitably generate large datasets with spectra of varying quality. Peptide-identification tools and their underlying algorithms employ multivariate minimum score-threshold to segregate true/false results, with a 5% false discovery rate deemed as the upper limit by the Paris Guidelines (published in Molecular and Cellular Proteomics). A seemingly arbitrary 'two-peptide rule', was also originally recommended by 'MCP guidelines' [2], and is often applied under the collective assumption that more peptide identifications lead to a higher confidence protein identification. Whilst the two-hit rule remains one of the most applied unsupervised filters for high throughput proteomics techniques, there have been very few theoretical studies to support and describe the 'two-hit rule'. Interestingly, even the 'Molecular & Cellular Proteomics (MCP) guidelines' state that: "The two-peptide rule was discussed in the context of studies where there was little or no analysis done at all." [3,4]. The consequent flow-through from this filter implies that any protein identification that carries only single peptide evidence is viewed with a high level of uncertainty and is discarded from subsequent analysis [3,5].

In fact, protein identification that is supported by a single peptide with high confidence can theoretically be more valid than by two or more peptides with lower confidence. For example an analysis of human and Shewanella oneidensis datasets [4] showed that OHWs with a high peptide-spectrum matching (PSM) score were better for protein identification purposes than 'two-peptide proteins' with low PSM scores, especially when taking other proteomic information into consideration such as protein length. With the development of new technologies, the false positive rate (FPR) of protein identification is becoming lower. Global proteomic studies typically show FPRs below 5% for peptide identifications and more accurate methods for distinguishing false identifications are continuingly being found. These limitations apply, at least part, to the analysis of data derived from multiplex iTRAQ tandem mass-spectrometry. Briefly, iTRAQ is a gel-free approach that allows the identification and quantification of proteins across a diverse range of molecular weights (MW), pI values (isoelectric point), cellular locations and functional categories. Population of protein fragments (peptides) that are produced through bulk enzymatic digestion (typically tryptic) of the proteome are covalently labeled with a isotopically modified piperazine tag at both the N-terminus and amine side-chain amino acids [6]. As tags have been isotopically tampered to give isobaric masses, both non-intrusive relative and absolute quantifications for multiple comparisons can be made during the gas-phase dissociation (MS/MS) stage when mass different reporter tags are released for detection [7]. Subsequent database searches to match MS spectra with known peptide and protein sequences are then mostly analogous to other techniques, relying on search-engines and algorithms to filter identifications. As the analysis of a complex proteome is limited by the dynamic range of the sample and the practicing instrumentation, often every dataset analysis produces a significant number of OHWs. Apart from the limitation of low abundance and size of protein [8], certain groups of proteins may be consistently excluded by virtue of their physical characteristics and their dependency to crude extraction protocols. Thus, in cases where OHWs are the only detectable peptides of an enzymatic digest or are key regulators of a process, arbitrary removal of these potentially true positive identifications becomes particularly detrimental [4,9,10]; leading to a loss of potentially valuable information on their related pathways and mechanisms.

The outcome of an iTRAQ analysis is a list of differentially expressed proteins. A useful approach to aid interpretation of the resulting meta-information is metabolic pathway analysis. A pathway is defined as a set of interlinked, sequential biochemical steps that drives a cellular biological process and there are a number of packages (both proprietary and open-source) for bioinformatic analysis. Reactome SkyPainter [11] bases its analysis on a hypergeometric, or Fisher's exact test, a statistical significance test used when sample sizes are small which gives the probability of observing at least N genes from an event if the event is not overrepresented in the submitted list of genes. For further information see for example [12,13, 14,15].

This study forms part of our ongoing research into how short-chain fatty acids (SCFAs) act as chemo-preventives in colorectal cancer [16]. Three large datasets generated from multi-plex iTRAQ MS/MS experiments on colon cancer cells treated with SCFAs or from tissue from an in vivo study [16]. SCFAs are histone-deacetylase inhibitors (HDACi) which promote a range of transcriptional and post-translational changes in cells [16]. By using high-throughput proteomic techniques, including iTRAQ MS/MS, coupled with pathways analysis, a more global view of the range of actions of SCFA might be developed. The study also addresses the hypothesis that: "By discarding all 'OHW', valid protein identifications are missed with the potential loss of entire metabolic pathways. This has implications in the overall understanding of biological processes."

In order to assess the validity of this hypothesis statistical and pathway analyses were carried on the three datasets generated from iTRAQ quantification. The relationship between pI, Mw and peptide coverage with respect to peptide hits was described to characterize and potentially correlate relationship between 'valid' and OHW proteins. SkyPainter pathways analysis [17] was used to assess representation of pathways in the datasets, comparing 'valid' OHW depleted datasets with complete unfiltered (two-peptide rule) identifications.

2. Material and methods

2.1 Datasets and software

The three datasets used in our study were generated by 8-plex iTRAQ tandem mass spectrometry. Two datasets were based on HCT116 cell culture experiments, the third dataset was derived from analysis of human colorectal biopsies [16]. The spectra were analysed using GeneBio's 'Phenyx' search engine. For peptide analysis and protein identification (http://www.genebio.com/products/phenyx) searches were undertaken against various publicly available datasets: NCBInr, UniProt and IPI. The results of these searches are available in Supplementary data 1a-c for each dataset. These contain details of all the proteins identified, including those by a single unique peptide, and related information on the number of validated unique peptide sequences on the protein, MS score, percent coverage, MW, pI value, etc. Table 1 gives a summary of the number of protein identifications and percentages of OHWs for each dataset. GraphPad Prism (GraphPad Prism 5 Demo) was used to analyze the correlation between MW and pI and the number of valid peptides for each protein identification and the means of each protein group.

Table 1. *The protein numbers for the three datasets of our study.* The columns give the total number of protein identifications; the number identified by 1 or 2 valid peptides; the number identified by >2 valid peptides and the percentage of '1&2-hit' proteins.

Dataset	n, total number of proteins identified by Phenyx	Number of proteins identified by 1&2 peptides	Number of proteins identified by >2 peptides	Percent of proteins identified by 1&2 peptides
Dataset 1	262	163	99	62.2%
Dataset 2	599	419	180	69.9%
Dataset 3	209	188	21	90.0%

Microsoft Excel was used to construct surface plots for visualizing the relationship between all 3 parameters; MW, pI value, and valid peptide hits.

Pathway analysis was carried out using EBI-EMBL Reactome 'SkyPainter' (http://www.reactome.org/cgi-bin/skypainter2). This peer-reviewed and manually curated knowledgebase [14,17] includes biological pathway steps inferred to exist based on experimental data and provides an infrastructure for computation across the entire metabolic reaction network for multiple species, principally Homo sapiens. Pathways in Reactome are described as a number of molecular events that transform input physical entities into output entities in catalyzed or regulated pathways by other entities. By imputing a range of protein identifiers, 'SkyPainter' calculates which pathways are statistically over- or under-represented in a set of identifications, using a hypergeometric testing.

3. Results

3.1 Correlation between molecular weight and number of valid peptides.

We sought to establish whether or not there is a relationship, as previously suggested, between number of representative peptides and molecular weight of the protein [4]. Fig 1 shows the relationship between MW and the number of valid peptide peptides: Datasets 1 & 3 (Figs. 1a & 1c) show positive linear correlation, although this relationship is only significant (p-value <0.0001) for dataset 1. Dataset 2 (Fig. 1b) shows non-significant negative linear trend. These data suggest that proteins with higher Mw are identified by more peptides, however taking the three analyses together, caution should be exercised in this interpretation and analysis of further datasets is required to establish such a relationship. The scatter-plots do show that the spread of MW is wider for OHWs than for proteins identified by more peptide peptides, consistent with observations by other groups [8] that typically >60% of proteins are identified by only one or two peptides.

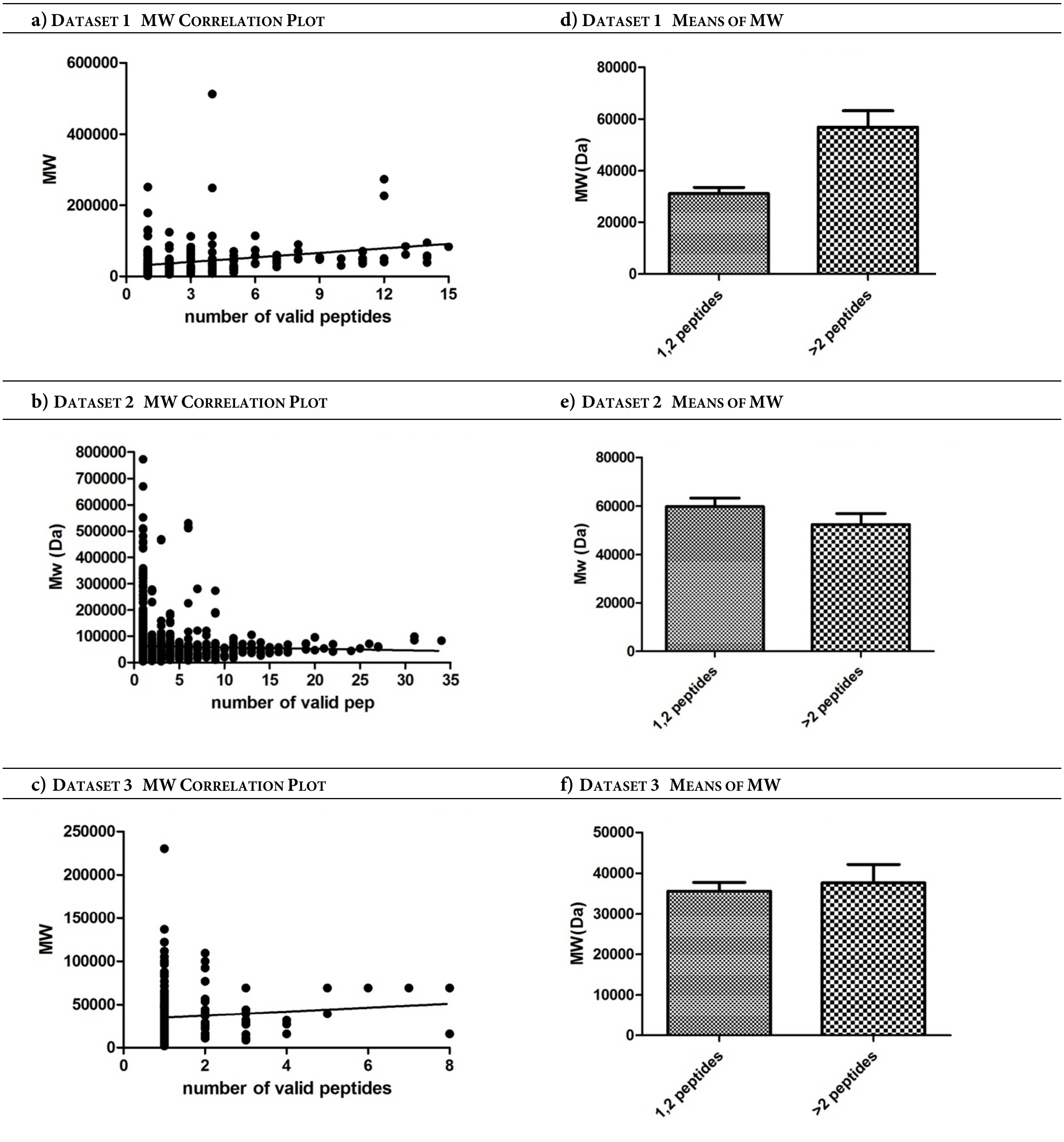
Figure 1 panels d-f show the mean MW of proteins for the two groups: Dataset 1 (Fig. 1d) shows the mean MW of OHW's is significantly smaller than for proteins identified by more peptides (p-values <0.0001). However, no significant difference was observed for datasets 2 & 3 (Figs. 1e & f, p-values =0.2227 and 0.7612 respectively).

3.2 Correlation between pI value and number of valid peptides.

Next we sought to investigate whether there is a relationship between pI and number of valid peptides. Figure 2a-c shows a consistent negative correlation between pI and the number of valid peptide peptides between the two groups for all 3 datasets. This correlation was significant (p-value <0.0001 & 0.0081 for datasets 1 & 2 respectively) but not for dataset 3 (Figs. 2a, b & c) which had a p-value of 0.0603. Although a consistent trend was observed, similar analysis of further datasets will be required to substantiate the significance of the correlation. The plots also show that proteins identified by fewer valid peptides show higher fluctuation of pI value, validating observations reported earlier [18].

3.3 Three-way interaction between Mw, pI and valid peptide number.

As there were trends to relationships between both Mw and peptides and pI and peptide number, the three-way interaction between these variables was explored using surface plots. These surface-plots allow the relationship between MW, pI value, and valid peptide hits to be visualized and qualitative interpretations to be drawn as well as provided data to support the interpretations. Figure 3a-c show surface plots of MW vs. pI by valid peptide hits. These were constructed by first grouping the results by valid peptide hits (OHW; 2 hits; 3 to 5 hits; etc.) and then calculating the mean MW for each group by pI band (pI 4 = ≥4.5 to <5.5; pI 5 = ≥5.5 to <6.5; etc.) Of note are the twin peaks in MW that occur at a similar position for each dataset; i.e. at an approximate pI of 6, with



a) Dataset 1 MW Correlation Plot

d) Dataset 1 Means of MW

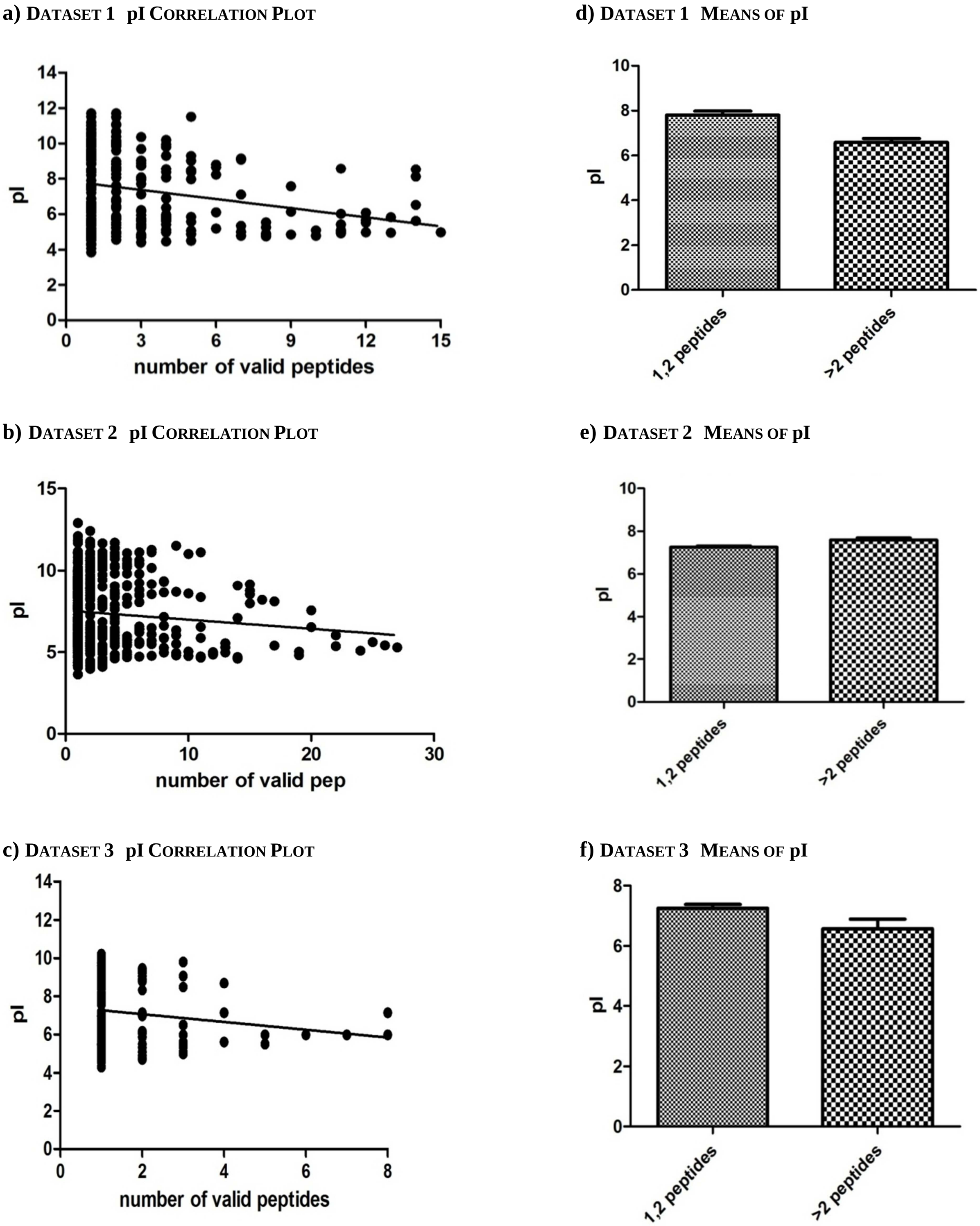
b) Dataset 2 MW Correlation Plot

e) Dataset 2 Means of MW

c) Dataset 3 MW Correlation Plot

f) Dataset 3 Means of MW

Figure 1. *The relationship between the M_W (Da) of a protein and the number of valid peptide hits in its identification***.** Figs. 1a-1c are correlation plots for the three datasets respectively. With the exception of Dataset 1, there are no trends observed and no significant correlation. The results are as follows: **(a)** Dataset 1, slope= 4044±690 (n=262), p-value<0.0001; **(b)** Dataset 2, slope= -557.7±479.4 (n=599), p-value=0.2449; **(c)** Dataset 3, slope = 2319±1822 (n=209), p-value=0.2046. Figs. 1d-1f compares the mean M_W between the two protein groups: (i) proteins identified by 1or 2 peptide-hits; and (ii) proteins identified by >2 peptides. Again no relationship is observed and only Dataset 1 shows a significant difference (by unpaired t-test). The results are as follows: **(d)** Dataset 1, 1&2-hits (n=163) mean M_W=31,140±2,429 Da; >2-hits (n=99) mean M_W=56,830±6,305 Da, p<0.0001; **(e)** Dataset 2, 1&2-hits (n=419) mean M_W=59,769 ±3,470 Da; >2-hits (n=180) mean M_W=52,361±4,525 Da, p=0.2227; (f) Dataset 3, 1&2-hits (n=188) mean M_W=35,620±2,167 Da; >2-hits (n=21) mean M_W=35,640±4,512 Da, p=0.7612.


a) Dataset 1 pI Correlation Plot

d) Dataset 1 Means of pI

b) Dataset 2 pI Correlation Plot

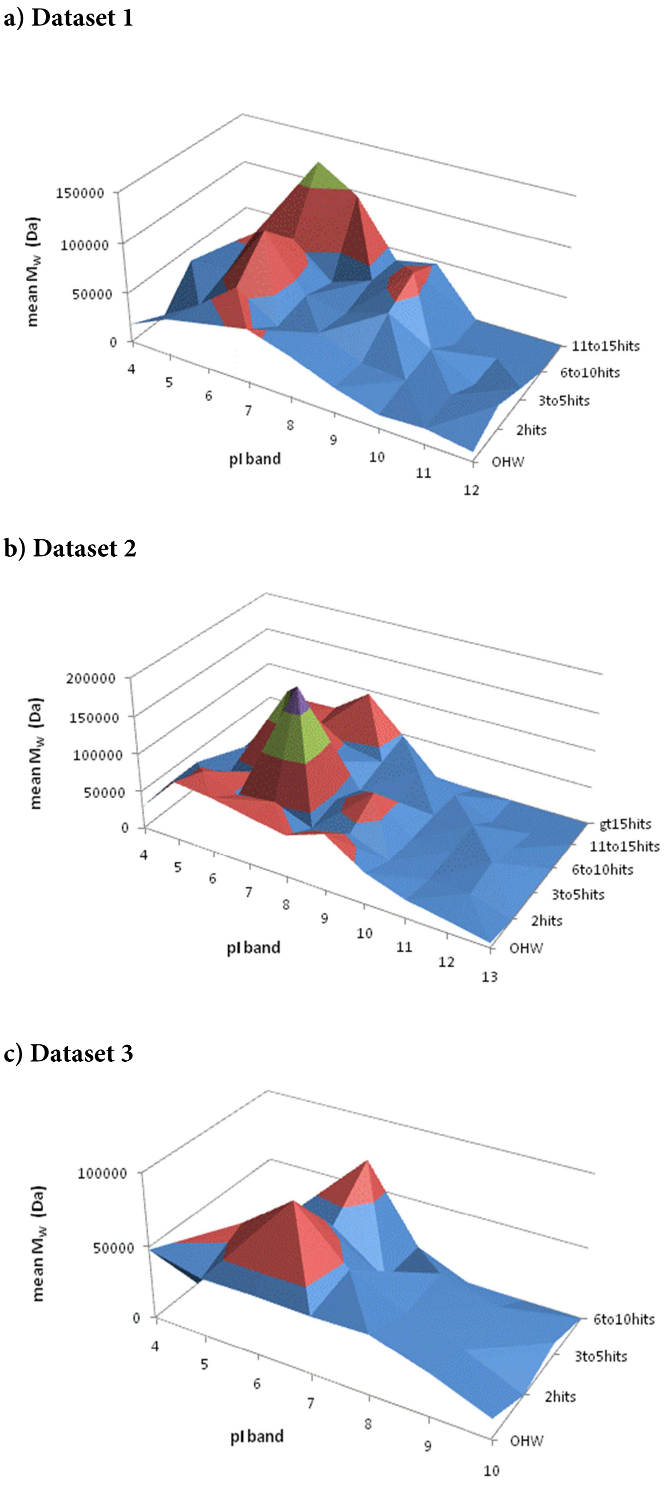
e) Dataset 2 Means of pI

c) Dataset 3 pI Correlation Plot

f) Dataset 3 Means of pI

Figure 2. *The relationship between the isoelectric point (pI) of a protein and the number of valid peptide hits in its identification***.** Figs. 2a-c are correlation plots for the three datasets respectively. Negative correlation is observed in all 3 datasets, although only Dataset 1 shows this to be significant. This suggests that peptides with high or low pI values are more likely to be represented by OHWs and so this group of proteins will potentially be excluded from any results. The results are as follows: (a) Dataset 1, slope= -0.1708±0.0411 (n=262), p-value<0.0001; (b) Dataset 2, slope= -0.5583±0.02101 (n=599), p-value=0.0081; (c) Dataset 3, slope = -0.2035±0.1078 (n=209), p-value=0.0603. Figs. 2d-f compare the mean pI values between the two protein groups: (i) proteins identified by 1or 2 peptide-hits; and (ii) proteins identified by >2 peptides. These results are inconsistent, only Dataset 1 shows a significant difference (by unpaired t-test) and there is no trend across the three datasets. The results are as follows: (d) Dataset 1, 1&2-hits (n=163) mean pI=7.803±0.1755 Da; >2-hits (n=99) mean pI=6.587±0.1758 Da, p<0.0001; (e) Dataset 2, 1&2-hits (n=419) mean pI=7.255±0.06110 Da; >2-hits (n=180) mean pI=7.584±0.1040 Da, p=0.0037; (f) Dataset 3, 1&2-hits (n=188) mean pI=7.248±0.1264 Da; >2-hits (n=21) mean pI=6.568±0.3215 Da, p=0.0854.

the first peak at 3 to 5 peptides and the second coinciding with the maximum peptide hits for each dataset (11-15 hits; >15 hits; 6-10 hits for datasets 1, 2 and 3 respectively).


a) Dataset 1

b) Dataset 2

c) Dataset 3

Figure 3. *Surface plots of* M_W *vs. pI by valid peptide hits for each of the 3 datasets.* The results have been grouped by valid peptide hits, then by pI band and the mean M_W for each group was calculated before the plots were constructed. The plots not only provide quantitative data but also provide a method of visualizing the relationships between these three parameters allowing qualitative interpretations and assessments to be made, for example the similarity between the position of the peaks in the plots for all 3 datasets and the observation that proteins in the OHW and 2-hit groups are only represented by low M_W proteins.

3.4 Pathway Analysis.

EBI-EMBL's Reactome SkyPainter application [13] was used to compare events and pathways between the "OHW group" and "valid group" of proteins for each of the three datasets, where the "OHW group" includes proteins identified by 1 or 2 peptide hits and the "valid group" only contains proteins identified by more than 2 valid peptide hits. Each identified protein in the MS experiments was assigned to a reaction event and biochemical or metabolic pathway in the Reactome SkyPainter knowledgebase, (which contains a computationally-accessible human pathway network that has been manually curated as described in "Materials and Methods"). For user-submitted lists of protein identifiers, as carried out for the three datasets of this study, pathway overrepresentation analysis was performed and the results returned in the form of pathway trees colour-coded by probability. As such the relationships between our proteins and other complexes, reactions and pathways can be visualized. Data mining allows the pathway-trees to be expanded and the contributing reactions and events to be viewed. Hyperlinks allow the supporting literature to be accessed.

Full details of the SkyPainter analyses including event and pathway results for each dataset are provided in Supplementary information 2a-c. Supplementary-3 gives a summary of the event data with rows highlighted in red for results involving 2 or more of the datasets. Table 2 gives a list of events that were exclusively identified by the 1-2 peptide group for at least two of the datasets. Table 2 also gives the un-adjusted probability of seeing N or more genes in these events by chance (where a low probability indicates that the genes are statistically overrepresented in the pathway [15]). In total 7 events were identified that would have been lost in two or more of the datasets if the 1-2 peptide groups are discarded, with apoptosis being the only event consistently lost by all three datasets, as detailed in Table 2. Complete pathway-trees, including child-branches are given in Supplementary information 4a-c. The root-pathways are shown in Figure 4 and summarized in Table 3 as follows: (Table 3a) Pathway Trees unique to the group of proteins identified by only 1 or 2 peptides; (Table 3b) Pathway Trees unique to the group of proteins identified by more than 2 peptides; (Table 3c) Pathway Trees common to proteins in both groups.

A familiarization and working knowledge of the Reactome SkyPainter tool is important when interpreting results. For example, identification of an individual event that would be lost if OHW are discarded does not necessarily mean the pathway will also be lost because a pathway can be composed from many related events. We identified apoptosis as being a consistently lost event for all three datasets, but only a lost pathway for datasets 2 and 3 (see Table 3). Further investigation of the pathway-tree for dataset 1 (see Supplementary information 4a) shows this to be by virtue of related proteins

Table 2. *Events that were unique to proteins identified by only 1 or 2 valid peptides in two or more of the datasets***.** These are events that would potentially be overlooked in any interpretation of the results if the 'two-hit rule is applied and all OHWs are excluded (full event lists can be found in the supplementary information).

Events unique to the group of proteins identified by only 1 or 2 peptide hits	Un-adjusted probability of seeing N or more genes in the event by chance			
	Dataset 1	Dataset 2	Dataset 3	Count
Apoptosis	0.00924	0.01356	0.02875	3
Cytosolic tRNA aminoacylation	0.04163	2.07E-05	-	2
Formation and Maturation of mRNA Transcript	0.00153	0.00060	-	2
mRNA Processing	0.00044	0.00015	-	2
Processing of Capped Intron-Containing Pre-mRNA	0.00016	0.00028	-	2
Processing of Capped Transcripts	0.00019	5.48E-05	-	2
Release of platelet cytosolic components	0.01105	-	0.00469	2

involved in the apoptotic execution phase (including Plectin; Importin subunit beta-1; and Lamin-A/C). Similarly, mRNA processing was identified as a lost event in OHWs in datasets 1 and 2 (see Table 2) but again was an identified pathway for dataset 1 (see Table 3) with further investigation revealing other related proteins and reactions including mRNA splicing (Supplementary information 4a).

4. Discussion

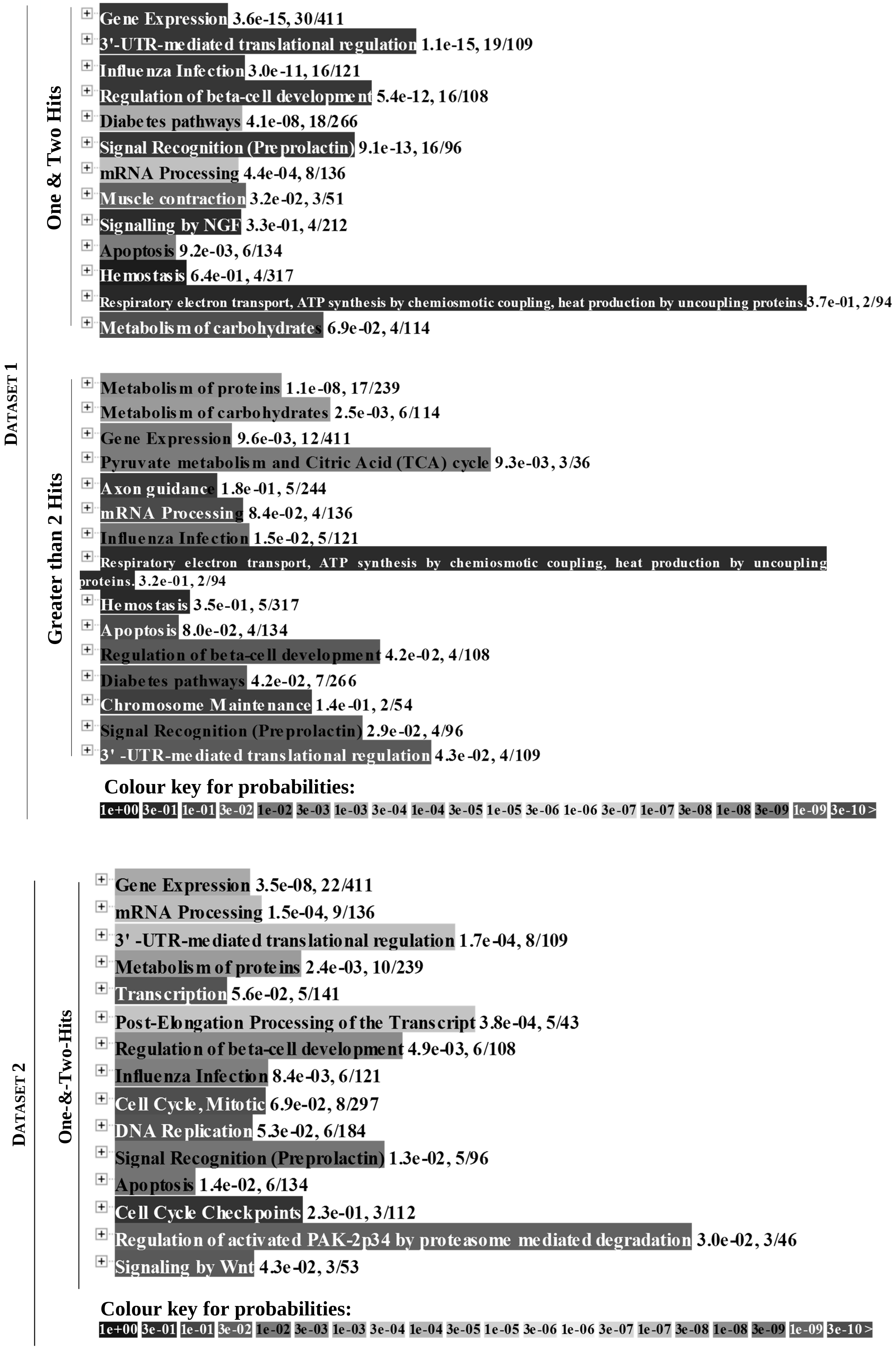
The ratio of MW to charge (m/z) is a key measurement metric in most MS protein identification and quantification. We had adopted the assumption, as proposed by others [7], that the number of trypsin cleavage sites increases with increased protein size, therefore the number of potentially detectable peptides should also increase accordingly. However our results (Fig. 1) showed an inconsistent correlation between MW and number of valid peptides and no significance and no consistency in the direction of difference between the mean MW of proteins identified by 1or 2 peptides or by more than 2 peptides. The lack of a strong and consistent correlation is surprising based on assumptions on purely stochastic grounds.

Using the second parameter, isoelectric point (pI), of the classes of proteins identified, we showed (Fig. 2) that although there was a trend, suggesting that proteins with high pI values produce fewer peptide fragments for identification, the results were only achieved significance for the first two datasets (Fig. 2a & b). The pI values of proteins identified in all three datasets are all between 4 and 14, with the exception of one protein having a pI<4. The results were unsurprising: a previous study [18] suggested that proteins with pI<4 have fewer arginine or lysine residues available for digestion by trypsin, thereby reducing the number of compatible peptides (for ion source protonation) available for positive ion mode (+ve MS) MS detection. Their study showed significant correlation in more than 95% of the proteins, with positive correlation for acidic proteins and inverse correlation for basic proteins. These data suggest the high fluctuations of pI values are therefore a consequence of protein length and amino acid composition, leading to broader and more sporadic pI shifts in shorter proteins.

Other studies [19] have also reported a higher proportion of negatively charged residues in peptides identified by MS with high confidence, with on average 16.8% of the residues in the high-scoring peptides being negatively charged, suggesting that the presence of acidic residues in a peptide may lead to more comprehensive and intense fragmentation of ions. EBI-EMBL Reactome SkyPainter was used for pathway and reaction event analysis (see Figure 4, Tables 2 & 3, and full details in the Supplementary information). A hypergeometric test is used to show events that are statistically overrepresented in the pathways. We had initially hypothesized that some pathways may be consistently lost if all OHWs are discarded. A pathway analysis of all three datasets has provided some evidence to support this and most importantly, a number of events were found to be uniquely identified by only one or two peptides in two or more of our datasets (see Table 2). While we have shown that one is able to validate and correlate extraneous layers of information given by traditionally discarded OHW proteins, we also recommend caution on the steps necessary to interpret these and other OHW data; for example, loss of event information does not necessarily mean loss of pathway information depending on which other proteins, events and reactions are involved, as discovered in our analysis of dataset 1 for the apoptosis and mRNA processing events, as discussed in detail earlier. We demonstrate that there are tangible benefits to compare and correlate OHW data, in turn minimizing overlooked pathways, and impairing the overall understanding of the biological process.

5. Concluding remarks

In this report, we have investigated the possibility of a relationship between MW and pI value of peptides and proteins


DATASET 1
One & Two Hits
Gene Expression 3.6e-15, 30/411
3'-UTR-mediated translational regulation 1.1e-15, 19/109
Influenza Infection 3.0e-11, 16/121
Regulation of beta-cell development 5.4e-12, 16/108
Diabetes pathways 4.1e-08, 18/266
Signal Recognition (Preprolactin) 9.1e-13, 16/96
mRNA Processing 4.4e-04, 8/136
Muscle contraction 3.2e-02, 3/51
Signalling by NGF 3.3e-01, 4/212
Apoptosis 9.2e-03, 6/134
Hemostasis 6.4e-01, 4/317
Respiratory electron transport, ATP synthesis by chemiosmotic coupling, heat production by uncoupling proteins. 3.7e-01, 2/94
Metabolism of carbohydrates 6.9e-02, 4/114
Greater than 2 Hits
Metabolism of proteins 1.1e-08, 17/239
Metabolism of carbohydrates 2.5e-03, 6/114
Gene Expression 9.6e-03, 12/411
Pyruvate metabolism and Citric Acid (TCA) cycle 9.3e-03, 3/36
Axon guidance 1.8e-01, 5/244
mRNA Processing 8.4e-02, 4/136
Influenza Infection 1.5e-02, 5/121
Respiratory electron transport, ATP synthesis by chemiosmotic coupling, heat production by uncoupling proteins. 3.2e-01, 2/94
Hemostasis 3.5e-01, 5/317
Apoptosis 8.0e-02, 4/134
Regulation of beta-cell development 4.2e-02, 4/108
Diabetes pathways 4.2e-02, 7/266
Chromosome Maintenance 1.4e-01, 2/54
Signal Recognition (Preprolactin) 2.9e-02, 4/96
3' -UTR-mediated translational regulation 4.3e-02, 4/109
Colour key for probabilities:
1e+00 3e-01 1e-01 3e-02 1e-02 3e-03 1e-03 3e-04 1e-04 3e-05 1e-05 3e-06 1e-06 3e-07 1e-07 3e-08 1e-08 3e-09 1e-09 3e-10 >
DATASET 2
One-&-Two-Hits
Gene Expression 3.5e-08, 22/411
mRNA Processing 1.5e-04, 9/136
3' -UTR-mediated translational regulation 1.7e-04, 8/109
Metabolism of proteins 2.4e-03, 10/239
Transcription 5.6e-02, 5/141
Post-Elongation Processing of the Transcript 3.8e-04, 5/43
Regulation of beta-cell development 4.9e-03, 6/108
Influenza Infection 8.4e-03, 6/121
Cell Cycle, Mitotic 6.9e-02, 8/297
DNA Replication 5.3e-02, 6/184
Signal Recognition (Preprolactin) 1.3e-02, 5/96
Apoptosis 1.4e-02, 6/134
Cell Cycle Checkpoints 2.3e-01, 3/112
Regulation of activated PAK-2p34 by proteasome mediated degradation 3.0e-02, 3/46
Signaling by Wnt 4.3e-02, 3/53
Colour key for probabilities:
1e+00 3e-01 1e-01 3e-02 1e-02 3e-03 1e-03 3e-04 1e-04 3e-05 1e-05 3e-06 1e-06 3e-07 1e-07 3e-08 1e-08 3e-09 1e-09 3e-10 >

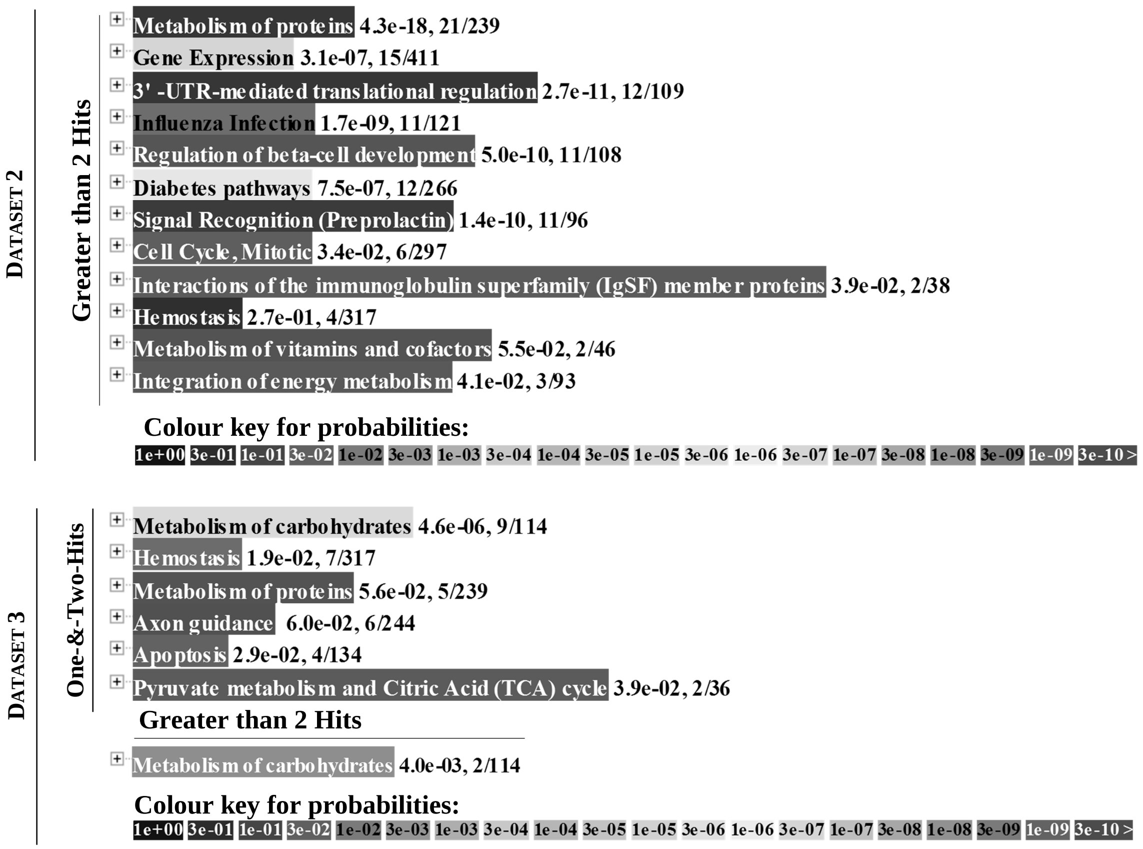


Figure 4. *Root-level pathway trees for the three datasets***.** This shows the main pathways represented by the two protein groups: [(i) proteins identified by 1or2 peptide-hits; and (ii) proteins identified by >2 peptides]. (The complete pathway trees are given in the Supplementary information). Of note is how these pathways compare to the events shown in Table 2 indicating that caution needs to be taken when interpreting pathway and event data at face-value, as events not represented appear in the pathway tree by virtue of related events, e.g. the apoptosis event and the apoptotic execution phase which is a branch of the apoptosis pathway for the group >2-peptide hits in Dataset 1.

Table 3. Summary of the top-level pathways for the three datasets**.**

a) Pathways unique to proteins identified by 1 or 2 valid peptides

DATASET 1	DATASET 2	DATASET 3
Signalling by NGF	Apoptosis	Apoptosis
Muscle contraction	Cell Cycle Checkpoint	Axon guidance
	DNA Replication	Hemostasis
	mRNA Processing	Metabolism of proteins
	Post-Elongation Processing of the Transcript	Pyruvate metabolism and Citric Acid (TCA) cycle
	Regulation of activated PAK-2p34 by proteasome mediated degradation	
	Signalling by Wnt	
	Transcription	

b) Pathway Trees unique to the group of proteins identified by more than 2 peptides

DATASET 1	DATASET 2	DATASET 3
Axon guidance	Diabetes Pathway	
Chromosome Maintenance	Hemostasis	
Metabolism of proteins	Integration of energy metabolism	
Pyruvate metabolism and Citric Acid (TCA)	Interactions of the immunoglobulin super-	

cycle	family (IgSF) member proteins Metabolism of vitamins and cofactors	

c) Pathway Trees common to proteins in both groups

DATASET 1	DATASET 2	DATASET 3
3'-UTR mediated translational regulation	3'-UTR mediated translational regulation	Metabolism of carbohydrates
Apoptosis	Cell Cycle, Mitotic	
Diabetes Pathway	Gene Expression	
Gene Expression	Influenza Infection	
Hemostasis	Metabolism of proteins	
Influenza Infection	Regulation of beta-cell development	
Metabolism of carbohydrates	Signal Recognition (Preprotactin)	
mRNA Processing		
Regulation of beta-cell development		
Respiratory electron transport, ATP synthesis by chemioscopic coupling and heat production by uncoupling proteins		
Signal Recognition (Preprotactin)		

identified in iTRAQ experiments with the number of valid peptides to determine if any protein groups are consistently lost to any analyses, when OHWs are discarded. Although no relationship between MW was established, there was a trend towards negative correlation between pI value and number of peptide identification.

Pathway analyses highlighted several events that were only attributed to proteins identified by only one (OHW) or two peptides in two or more of our datasets.

While we acknowledge that the confirmation of our observations requires further analysis using orthogonal validation, for example by western blot analysis, we advocate the importance of the 'lost' information for a global interpretation of the data and therefore suggest that a more open approach should be taken when analyzing MS data since all candidate proteins/pathways will require verification. With the continued development of new technologies, software algorithms and bioinformatics tools, we believe the validation of OHW should become much more feasible [20, 21, 22].

6. Supplementary material

Supplementary data and information is available at: http://www.jiomics.com/index.php/jio/rt/suppFiles/53/0

Supplementary 1a to c: Phenyx Protein Information for Datasets 1 to 3 respectively. Supplementary 2a to 2c: EBI-EMBL Reactome SkyPainter event and pathway results for Datasets 1 to 3 respectively. Supplementary 3: Summary of SkyPainter Events for the 3 Datasets. Supplementary 4a to c: Complete Pathway trees from SkyPainter for Datasets 1 to 3 respectively.

Acknowledgements

EPSRC funding, EPSRC grant GR/S84347/01, for the ChELSI initiative to support OSY and JK through a doctoral training grant to Chemical Engineering at the Life Science Interface at the University of Sheffield.

References

1. Aebersold, R., and Mann, M. (2003). Mass spectrometry-based proteomics. Nature 422, 198-207.
2. Carr, S., Aebersold, R., Baldwin, M., Burlingame, A., Clauser, K., and Nesvizhskii, A. (2004). The need for guidelines in publication of peptide and protein identification data - Working group on publication guidelines for peptide and protein identification data. Molecular & Cellular Proteomics 3, 531-533.
3. Eddes, J.S., Kapp, E.A., Frecklington, D.F., Connolly, L.M., Layton, M.J., Moritz, R.L., and Simpson, R.J. (2002). CHOMPER: A bioinformatic tool for rapid validation of tandem mass spectrometry search results associated with high-throughput proteomic strategies. Proteomics 2, 1097-1103.
4. Gupta, N., and Pevzner, P.A. (2009). False Discovery Rates of Protein Identifications: A Strike against the Two-Peptide Rule. Journal of proteome research 8, 4173-4181.
5. Bradshaw, R.A., Burlingame, A.L., Carr, S., and Aebersold, R. (2006). Reporting protein identification data - The next generation of guidelines. Molecular & Cellular Proteomics 5, 787-788.
6. Ernoult, E., Gamelin, E., and Guette, C. (2008). Improved proteome coverage by using iTRAQ labelling and peptide OFFGEL fractionation. Proteome Science 6.
7. Aggarwal, K., Choe, L.H., and Lee, K.H. (2006). Shotgun proteomics using the iTRAQ isobaric tags. Briefings in functional genomics & proteomics 5, 112-120.
8. Adkins, J.N., Varnum, S.M., Auberry, K.J., Moore, R.J., Angell, N.H., Smith, R.D., Springer, D.L., and Pounds, J.G. (2002). Toward a human blood serum proteome - Analysis by multidimensional separation coupled with mass spectrometry. Molecular & Cellular Proteomics 1, 947-955.
9. Pan, S.Q., Gu, S., Bradbury, E.M., and Chen, X. (2003). Single peptide-based protein identification in human proteome through MALDI-TOF MS coupled with amino acids coded mass tagging. Analytical Chemistry 75, 1316-1324.
10. Veenstra, T.D., Conrads, T.P., and Issaq, H.J. (2004). Com-

mentary: What to do with "one-hit wonders"? Electrophoresis 25, 1278-1279.

11. www.reactome.org (2009). Reactome: a human pathway database (Cambridge, UK, EMBL-EBI).

12. Deutsch, E.W., Lam, H., and Aebersold, R. (2008). Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. Physiological Genomics 33, 18-25.

13. Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G.R., Wu, G.R., Matthews, L., et al. (2005). Reactome: a knowledgebase of biological pathways. Nucleic Acids Research 33, D428-D432.

14. Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., et al. (2009). Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Research 37, D619-D622.

15. Vastrik, I., D'Eustachio, P., Schmidt, E., Gopinath, G., Croft, D., de Bono, B., Gillespie, M., Jassal, B., Lewis, S., Matthews, L., et al. (2009). Reactome: a knowledge base of biologic pathways and processes (vol 8, pg 39, 2007). Genome Biology 10.

16. Drake, P.J., Griffiths, G.J., Shaw, L., Benson, R.P., and Corfe, B.M. (2009). Application of high-content analysis to the study of post-translational modifications of the cytoskeleton. Journal of proteome research 8, 28-34.

17. Vastrik, I., D'Eustachio, P., Schmidt, E., Joshi-Tope, G., Gopinath, G., Croft, D., de Bono, B., Gillespie, M., Jassal, B., Lewis, S., et al. (2007). Reactome: a knowledge base of biologic pathways and processes. Genome Biology 8.

18. Kiraga, J., Mackiewicz, P., Mackiewicz, D., Kowalczuk, M., Biecek, P., Polak, N., Smolarczyk, K., Dudek, M.R., and Cebrat, S. (2007). The relationships between the isoelectric point and: length of proteins, taxonomy and ecology of organisms. BMC Genomics 8, 163.

19. Benjamini, Y., and Hochberg, Y. (1995). Controlling The False Discovery Rate - A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological 57, 289-300.

20. Goodlett, D.R., Bruce, J.E., Anderson, G.A., Rist, B., Pasa-Tolic, L., Fiehn, O., Smith, R.D., and Aebersold, R. (2000). Protein identification with a single accurate mass of a cysteine-containing peptide and constrained database searching. Analytical Chemistry 72, 1112-1118.

21. Nesvizhskii, A.I., Keller, A., Kolker, E., and Aebersold, R. (2003). A statistical model for identifying proteins by tandem mass spectrometry. Analytical Chemistry 75, 4646-4658.

22. Zhen, Y.J., Xu, N.F., Richardson, B., Becklin, R., Savage, J.R., Blake, K., and Peltier, J.M. (2004). Development of an LC-MALDI method for the analysis of protein complexes. Journal of the American Society for Mass Spectrometry 15, 803-822.