Journal of proteome research

Technical Note

A comparison of labelling and label-free mass spectrometry-based proteomics approaches.

Vibhuti Jitendra Patel, Konstantinos Thalassinos, Susan Slade, Joanne B Connolly, Andrew Crombie, J. Colin Murrell, and James Scrivens

J. Proteome Res., **Just Accepted Manuscript** • DOI: 10.1021/pr900080y • Publication Date (Web): 12 May 2009

Just Accepted

Journal of Proteome Research is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

A comparison of labelling and label-free mass spectrometry-based proteomics approaches

Vibhuti J. Patel[†§], Konstantinos Thalassinos[†§], Susan E. Slade[†], Joanne B. Connolly[‡], Andrew Crombie[†], J. Colin Murrell[†], James H. Scrivens[†]*

Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, UK,

Waters Corporation, Atlas Park, Simonsway, Manchester M22 5PP, UK

[§] Both authors contributed equally to this work.

*To whom correspondence should be addressed. Tel: +44(0)24-7657-4189. Fax: +44(0)24-7652-3568 E-mail: j.h.scrivens@warwick.ac.uk

[†] Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, UK

[‡] Waters Corporation, Atlas Park, Simonsway, Manchester M22 5PP, UK

The proteome of the recently discovered bacterium *Methylocella silvestris* has been characterised using three profiling and comparative proteomics approaches. The organism has been grown on two different substrates enabling variations in protein expression to be identified. The results obtained using the experimental approaches have been compared with respect to number of proteins identified, confidence in identification, sequence coverage and agreement of regulated proteins. The sample preparation, instrumental time and sample loading requirements of the differing experiments are compared and discussed. A preliminary screen of the protein regulation results for biological significance has also been performed.

Keywords: proteomics, quantification, iTRAQ, label-free, methanotroph, mass spectrometry

Introduction

Since the mid-1990s, mass spectrometry-based strategies have been the mainstream method for protein identification [1]. There remain, however, a number of issues to be tackled. Intrinsic characteristics of proteomes raise a number of experimental challenges. By nature, proteomes are large and complex. A single gene can often give rise to multiple, distinct proteins due to alternative splicing, sequence polymorphisms and post-translational modifications. Protein databases generated from the genome of an organism may, therefore, not be a true reflection of the potential protein complement [2]. There has been significant progress in the development of new approaches to tackle these issues, but technical challenges persist.

An ideal approach would enable the comprehensive characterisation of proteomes in a high-throughput manner. Currently, the techniques involved can be complex, costly and involving time-consuming data analysis. A low number of replicate experiments conducted – often due to a lack of sample availability – means that reproducibility is a concern. In addition, any given technique may only yield information on a fraction of the relevant peptides in any single analytical run [3].

An established proteomics approach is based on the separation of proteins via one- or two-dimensional polyacrylamide gel electrophoresis (PAGE). Proteins are digested within the gel, and the resulting peptides extracted for MS analysis. Drawbacks associated with PAGE include dynamic range, insufficient resolving power to fully separate all proteins within a sample, and restricted sample throughput [4].

Non gel-based techniques have been developed for the analysis of complex proteomic samples: so-called 'shotgun' experiments, where a whole proteome is digested without prior protein separation. Typically, the resulting peptides are separated by strong cation exchange chromatography (SCX) before reversed-phase LC-MS/MS analysis [5], an example of an approach known as multi-dimensional protein identification techniques (MudPIT). This method has been shown to provide increased proteome

coverage compared to gels, although it still suffers from problems with reproducibility and dynamic range. This approach has gained popularity within proteomic studies in preference to gels [2].

In addition to providing a profile of what proteins are present within a system at a given time, information on the expression levels of these proteins is increasingly required. Techniques in comparative and quantitative proteomics have, therefore, also developed significantly in recent years. Relative quantification can be performed on proteins separated by two-dimensional PAGE, using image analysis software, sometimes incorporating selective labelling approaches such as difference gel electrophoresis (DiGE) [6]. This approach is subject to the restrictions imposed by the gel methods.

A number of labelling approaches can also be incorporated into 'shotgun' type experiments. These include stable isotope labelling by amino acids in cell culture (SILAC) [7], isotope dilution [8], stable isotope labelled peptides [9], radiolabelled amino acid incorporation [10], chemically synthesised peptide standards [11], tandem mass tags (TMT) [12], isotope-coded affinity tags (ICAT) [13], and more recently, isobaric tags for relative and absolute quantification (iTRAQ) [14]. The iTRAQ system is now commercially available with eight isobaric tags [15], having only initially been available with four tags, and has been widely used in proteomic studies [16].

Most label-based quantification approaches have potential limitations: complex sample preparation, the requirement for increased sample concentration, and incomplete labelling. There has, therefore, recently been a focus in the area of non-labelled quantification in order to address some of these issues [17].

Non-labelled techniques which have been developed include peptide match score summation (PMSS) [18] and spectrum sampling (SpS) [19], both of which can be combined with statistical evaluation to detect differentially expressed proteins [20]. Another approach utilises a protein abundance index (PAI), [21] which can be converted to exponentially modified PAI (emPAI) for absolute protein quantification [22].

It has been observed that electrospray ionisation (ESI) provides signal responses that correlate linearly with increasing concentration [23], but there have been concerns regarding the nonlinearity of signal response [24]. Previous works have introduced quantitative, label-free LC-MS-based strategies for global profiling of complex protein mixtures [25] [26]. More recently, a simple LC-MS-based methodology was published which relies on changes in signal response from each accurate mass measurement and corresponding retention time (AMRT) to directly reflect concentrations in one sample relative to another [27], which has since been developed into a label-free system capable of relative and absolute quantification [28] [29]. All detectable, eluting peptides and their corresponding fragments are observed via rapid switching between high and low collision energy during the LC-MS/MS experiment, giving a comprehensive list of all ions that can subsequently be searched [30].

In this work three proteomics approaches have been used to identify and relatively quantify the proteins within a bacterium when grown under different substrates. Samples have been analysed both qualitatively and quantitatively by: (i) one-dimensional PAGE, (ii) MudPIT incorporating iTRAQ tags, and (iii) a data-independent, alternate scanning LC-MS method enabling label-free quantification. Comparisons have been made regarding experimental considerations such as ease of use, amount of biological sample required, time required to prepare samples for analysis and total instrument time. The data obtained have been evaluated with respect to number of protein identifications, confidence of the assignments, sequence coverage and agreement of regulated proteins. All approaches have been carried out using equivalent instrumentation, enabling the results to be more directly compared. The organism used in this study is the methanotrophic bacterium *Methylocella silvestris*, an environmentally important organism involved in global methane cycling. Unlike other methanotrophs, *M. silvestris* is able to grow on multi-carbon compounds as well as on methane [31]. In this work, cultures of *M. silvestris* were grown on acetate and methane.

Materials and methods

Bacterial growth and sample preparation

Methylocella silvestris was grown in fermenter cultures on diluted nitrate mineral salts (NMS) medium with methane or acetate (5 mM) as previously described [31]. Cells, grown to late exponential phase (OD_{540} ~1.0), were harvested by centrifugation (17,700 × g, 20 min, 4°C), washed in growth medium, resuspended in 0.1 M PIPES buffer (piperazine-N,N′-bis[2-ethanesulfonic acid], pH 7.0), and frozen in liquid nitrogen. Subsequently, frozen cells were thawed and resuspended in PIPES buffer containing 1 mM benzamidine and broken by four passes through a French pressure cell at 125 MPa (4°C) (American Instrument Co., Silver Spring, MD). Cell debris and membranes were removed by two centrifugation steps (13,000 × *g*, 30 min, 4°C, followed by 140,000 × *g*, 90 min, 4°C), and the supernatant, containing soluble cytoplasmic proteins, used for analysis. A protein assay was conducted on the soluble extract, using a Micro BCA Protein Assay Kit (Pierce Protein Research Products, Thermo Scientific, Cramlington, UK) according to the manufacturer's protocol.

Protein separation by gel electrophoresis

Proteins were resolved by 1D SDS-PAGE (14 µg per lane) and stained with Coomassie Blue. 30 to 40 slices were excised from each lane, and subjected to tryptic digestion. All processing of the gel plugs was performed by a MassPrep robotic protein handling system (Waters Corporation, Manchester, UK) using the manufacturer's protocol. In brief, the gel plugs were destained, the disulfide bonds were reduced by the addition of dithiothreitol and the free cysteine residues were alkylated with iodoacetamide. The gel plugs were washed prior to a dehydration step, followed by the addition of trypsin (Promega, Southampton, UK), and incubated for 4.5 h. The resultant tryptic peptides were extracted twice and transferred to a cooled 96-well microtitre plate; if necessary, they were stored at –20 °C.

iTRAQ labelling and strong cation exchange chromatography (SCX)

Labelled quantification was carried out using the iTRAQ 4-plex labelling kit (Applied Biosystems, Warrington, UK). Protein extracts from the two growth conditions were digested and labelled according to the manufacturer's standard protocol, and the samples pooled and lyophilised. A total of 400 μg protein from each growth condition was labelled, giving a total protein loading of 800 μg. As SCX was carried out offline, the potential for sample losses is higher. A larger initial protein loading was therefore used in order to minimise such losses and optimise the number of proteins identified by this approach. 200 μg of acetate-grown sample was labelled with the 114 reporter tag, and 200 μg with the 116 reporter tag. 200 μg of the methane-grown sample was labelled with the 115 tag and 200 μg with the 117 tag. As per the manufacturer's protocol, a maximum of 100 μg of protein was labelled per vial of iTRAQ label, i.e. two vials were used per label. The labelling of one growth condition with two different iTRAQ tags provides the means for an internal control to monitor labelling efficiency. The labelled tryptic peptides were partially resolved using a PolySULFOETHYL A SCX column, 2.1 mm × 20 cm, 5 μm particles, 300 Å pore size (PolyLC, Columbia, USA), using a stepwise gradient of KCl, adapted from Link *et al*. [32], from 2.5–50% salt solution over a period of 75 minutes. In total, 64 fractions were collected.

In-solution tryptic digestion

100 μg of soluble protein extract was resuspended in 1 mL of 0.1% Rapigest (Waters Corporation, Milford, MA) and concentrated using a 5 kDa cut-off spin column. The solution was then heated at 80°C for 15 minutes, reduced with DTT at 60°C for 15 minutes, alkylated in the dark with iodoacetamide at ambient temperature for 30 minutes, and digested with 1:50 (w/w) sequencing grade trypsin (Promega, Southampton, UK) for 16 hours. RapiGest was hydrolysed by the addition of 2 μL 15 M HCl, centrifuged, and each sample diluted 1:1 with a 50 fmol/μl glycogen phosphorylase B standard

tryptic digest to give a final protein concentration of 500 ng/µl per sample and 25 fmol/µl phosphorylase B.

LC-MS/MS acquisition for gel-separated samples

Peptides extracted from the digested gel were transferred to a nanoACQUITY system (Waters Corporation). A 6.4 µl aliquot of extract was mixed with 13.6 µl of 0.1% formic acid and loaded onto a 0.5 cm LC Packings C18 5 µm 100Å 300 µm i.d µ-precolumn cartridge. Flushing the column with 0.1% formic acid desalted the bound peptides before a linear gradient of solvent B (0.1% formic acid in acetonitrile) at a flow rate of approximately 200 nl/min eluted the peptides for further resolution on a 15 cm LC Packings C18 5 µm 5Å 75 µm i.d. PepMap analytical column. The eluted peptides were analysed on a Micromass Q-Tof Global Ultima (Waters Corporation) mass spectrometer fitted with a nano-LC sprayer with an applied capillary voltage of 3.5 kV. The spectral acquisition scan rate was 1.0 s with a 0.1 s interscan delay. The instrument was calibrated against a collisionally induced dissociation (CID) spectrum of the doubly charged precursor ion of [Glu^1]-fibrinopeptide B (GFP – Sigma Aldrich, St. Louis, USA), and fitted with a GFP lockspray line. The instrument was operated in data dependent acquisition (DDA) mode over the mass/charge (*m/z*) range of 50-2000. During the DDA analysis, CID experiments were performed on the three most intense, multiply charged peptides as they eluted from the column at any given time. Once these data have been collected, the next three most intense peptides are selected, and this process repeated.

LC-MS/MS acquisition for iTRAQ samples

Fractions collected from the SCX separation of iTRAQ-labelled peptides were snap-frozen on dry ice and lyophilised to dryness. The samples were resuspended in 20 µl 0.1% formic acid and transferred to a CapLC system (Waters Corporation). A 6.4 µl aliquot of extract was mixed with 13.6 µl of 0.1%

formic acid and loaded onto a 0.5 cm LC Packings C18 5 μm 100Å 300 μm i.d precolumn cartridge. Flushing the column with 0.1% formic acid desalted the bound peptides before a linear gradient of solvent B (0.1% formic acid in acetonitrile) at a flow rate of approximately 200 nl/min eluted the peptides for further resolution on a 15 cm LC Packings C18 5 μm 5Å 75 μm i.d. PepMap analytical column. The eluted peptides were analysed on a Micromass Q-Tof Global Ultima (Waters Corporation) mass spectrometer fitted with a nano-LC sprayer with an applied capillary voltage of 3.5 kV. The spectral acquisition scan rate was 1.0 s with a 0.1 s interscan delay. The instrument was calibrated against a CID spectrum of the doubly charged precursor ion of GFP, and fitted with a GFP lockspray line. The instrument was operated in data dependent acquisition (DDA) mode as described above.

LC-MS configurations for label-free analysis

Nanoscale LC separations of tryptic peptides for qualitative and quantitative multiplexed LC-MS analysis were performed with a nanoACQUITY system (Waters Corporation) using a Symmetry C_{18} trapping column (180 μm x 20 mm 5 μm) and a BEH C_{18} analytical column (75 μm x 250 mm 1.7 μm). The composition of solvent A was 0.1% formic acid in water, and solvent B, 0.1% formic acid in acetonitrile. Each sample (total protein 0.5 μg) was applied to the trapping column and flushed with 1% solvent B for 5 minutes at a flow rate of 15 μL/min. Sample elution was performed at a flow rate of 300 nL/min by increasing the organic solvent concentration from 3 to 40% B over 90 min. All analyses were conducted in triplicate. The precursor ion masses and associated fragment ion spectra of the tryptic peptides were mass measured with a Q-ToF Premier mass spectrometer (Waters Corporation) directly coupled to the chromatographic system.

The time-of-flight analyzer of the mass spectrometer was externally calibrated with NaI from *m/z* 50 to 1990, with the data post-acquisition lockmass-corrected using the monoisotopic mass of the doubly charged precursor of GFP, fragmented with a collision energy of 25V. The GFP was delivered at 500

fmol/μL to the mass spectrometer via a NanoLockSpray interface using the auxiliary pump of a nanoACQUITY system at a flow rate of 500 nL/min. The reference sprayer was sampled every 60 seconds.

Accurate mass data were collected in data independent mode of acquisition (LC-MS^E) by alternating the energy applied to the collision cell between a low energy and elevated energy state. The spectral acquisition scan rate was 0.6 s with a 0.1 s interscan delay. In the low energy MS mode, data were collected at constant collision energy of 4 eV. In elevated energy MS mode, the collision energy was ramped from 15 eV to 35 eV during each integration.

Data processing for DDA acquisitions

The uninterpreted MS/MS data from the gel-separated and iTRAQ-labelled samples were processed using ProteinLynx Global Server (PLGS) v2.3. The data were smoothed, background subtracted, centred and deisotoped. All data were lockspray calibrated against GFP using data collected from the reference line during acquisition.

Data processing for label-free acquisitions

The LC-MS^E data were processed using PLGS v2.3. The ion detection, data clustering and normalisation of the data independent, alternate scanning LC-MS^E data has been explained in detail elsewhere [33]. In brief, lockmass-corrected spectra are centroided, deisotoped, and charge-state-reduced to produce a single accurately mass measured monoisotopic mass for each peptide and the associated fragment ion. The initial correlation of a precursor and a potential fragment ion is achieved by means of time alignment.

Database searches

All data were searched using PLGS v2.3 against a *Methylocella silvestris* database (http://genome.ornl.gov/microbial/msil). Fixed modification of carbamidomethyl-C was specified, and variable modifications included were acetyl N-terminus, deamidation N, deamidation Q and oxidation M. For the iTRAQ data, variable modifications for the isobaric tags were specified. One missed cleavage site was allowed. Search parameters specified were a 50 ppm tolerance against the database-generated theoretical peptide ion masses and a minimum of one matched peptide.

For the LC-MS^E data, the time-based correlation applied in data processing was followed by a further correlation process during the database search that is based on the physicochemical properties of peptides when they undergo collision induced fragmentation [34]. The precursor and fragment ion tolerances were determined automatically. The protein identification criteria also included the detection of at least three fragment ions per peptide, at least one peptide determined per protein and the identification of the protein in at least two out of three technical replicates. By using protein identification replication as a filter, the false positive rate is minimised, as false positive protein identifications, i.e. chemical noise, have a random nature and as such do not tend to replicate across injections. This approach rules out systematic search events errors due to the repeated ambiguity of a particular spectrum and the subsequent sequence assignment by a search algorithm, as could be the case with peptide-centric searches.

Protein quantification using iTRAQ labelling

PLGS was also used for quantitative evaluation of MS/MS data generated from the analysis of the iTRAQ-labelled peptides. A relative quantification was conducted using a merged dataset comprising the results from the database search. Concentration ratios of iTRAQ-labelled proteins were calculated

based on signal intensities of reporter ions observed in peptide fragmentation spectra, with the relative areas of the peaks corresponding to proportions of the labelled peptides [14].

Protein quantification using label-free system

Relative quantitative analysis across conditions was performed by comparing normalised peak area/intensity of each identified peptide. Normalisation of the data was conducted by the use of an internal protein digest standard. In brief, peak areas/intensities are corrected using those of the internal protein digest. Intensity measurements are typically further adjusted on those components, that is de-isotoped and charge-state reduced accurate mass retention time pairs, that replicate throughout the complete experiment. Next, the redundant, proteotypic quantitative measurements provided by the multiple tryptic peptide identification from each protein were used to determine an average, relative protein fold-change. The algorithm performs binary comparisons for each of the conditions to generate an average normalised intensity ratio for all matched proteins. Proteins with a likelihood of quantification smaller than 0.05 were considered to be significantly regulated. The entire data set of differentially expressed proteins was further filtered by considering only the identified peptides that replicated two out of three technical instrument replicates. A likelihood of regulation higher than 95%, as reported by the quantification algorithm, was considered.

Results and discussion

Protein identifications

Three distinct experimental approaches have been employed in order to provide profiling and quantitative information regarding the proteome of *M. silvestris*. The numbers of proteins identified via each approach are summarised in **Table 1**. The total number of non-redundant proteins identified, when single peptide identifications are included, is comparable for all three techniques at 389, 384 and 425 proteins respectively.

	Gel-based	**iTRAQ**	**Label-free**
Total number of protein identifications	389	384	425
Single-peptide identifications	154	206	4
Proportion of identifications which were from a single peptide	40 %	54 %	0.9 %
Identifications with more than one peptide	235	178	421

Table 1: Total protein identifications for the three experimental approaches

Differences arise, however, when looking at the number of peptides per protein identification. There have been questions raised in the literature regarding the validity of identifications performed using a single peptide, so-called 'one-hit wonders', and whether they should be included in the list of proteins identified [35]. Of the gel separation identifications, 154 were from a single peptide as are 206 in the iTRAQ experiment. As a proportion of the total proteins identified, these values are 40% and 54% respectively. This is typical of many results in the literature [36]. In the label-free results, of the 425 identifications, only 4 are from a single peptide: proportionally less than 1%. As the label-free analysis is performed in triplicate, only an identification observed in at least two of the three replicates was taken to be valid; therefore, single-peptide identification in label-free data means that a single peptide was found in at least two of three data sets. Proteins identified by each experimental setup are listed in

Supplementary Data, Tables 1-5. **Figure 1** shows the overlap of protein identifications between the three approaches; **1a** uses all data, including single-peptide identifications, **1b** illustrates filtered data, with only identifications obtained with two or more peptides. All proteins identified are listed in Supplementary Tables 1 to 5, giving information on the molecular weight and pI of the identifications, and also the number of peptides identified. When including single peptide based identifications, there are a total of 699 proteins identified. Each of the techniques uniquely provides approximately 17% of those identifications. The remaining 49% of the identifications overlapped as shown. To overcome the uncertainty involved in the inclusion of single peptide-based identification, Figure 2b shows the data presented only including identifications made using a minimum of two peptides. This gives a total of 509 protein identifications, of which 9% were unique to the gel-based approach, 6% to iTRAQ, and 38% to label-free. This shows a significant increase in the proportion of unique identifications by the label-free method, a reduction in gel-based unique identifications and a considerable decrease in those uniquely identified by iTRAQ.

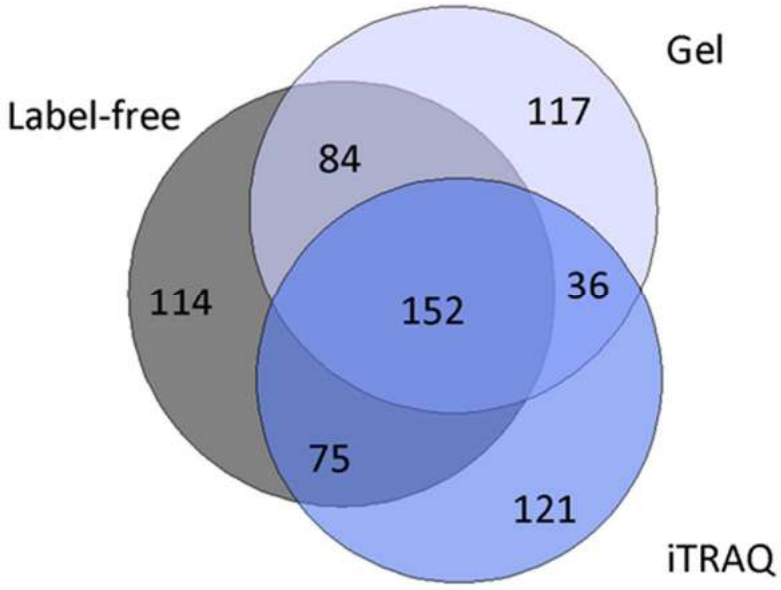


Figure 1a: Number of proteins identified by the various experimental approaches, including single-peptide identifications.

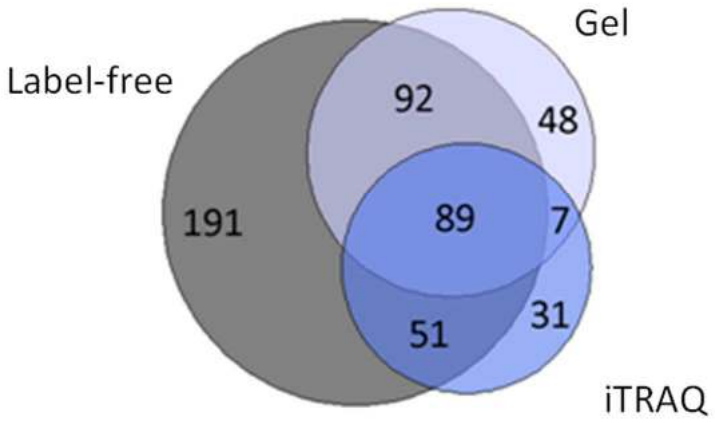


Figure 1b: Proteins identified by the various experimental approaches, identifications based on a minimum of two peptides.

A closer inspection of the number of proteins identified with and without the inclusion of single-peptide identifications reveals some interesting observations. As one would expect, the total number of proteins identified is lower when single-peptide identifications are excluded (509 when excluded, compared to 699 when included), including those identifications common to all three methods (89 when excluded, compared to 152 when included). In contrast, the number of proteins unique to the label-free method, and those common to the label-free and gel methods, has increased. This is because all but four of the proteins identified by the label-free method were done so with two or more peptides, whereas the gel and iTRAQ methods generated a large number of single-peptide identifications. The fact that 152 proteins were independently identified by all three methods provides strong evidence that although some of these (63 in total) were identified with a single-peptide by one or more technique, they should possibly not be discarded as false-positive identifications. This raises the questions as to what should be done with protein identifications based on a single-peptide. While the majority of these are likely to

correspond to false-positive identifications, there are a small number that are potentially valid and should also be included in the list of confidently-identified proteins, although this is not definitive. Further discussions of results will therefore exclude single-peptide identifications.

Relative quantification of identified proteins

Supplementary Figure 1 shows the 1D SDS-PAGE separation of the *M. silvestris* proteome obtained from different growth conditions. The difference in intensity of staining in a 1D gel is indicative of differential expression, and some representative changes are highlighted. A band on a 1D gel, however, can often contain multiple proteins due to the limitations of the resolving power of this technique. Although the analysis of gel-separated samples provided a comparable number of protein identifications, quantitative analysis using a 1D separation is difficult. Quantification via gel methods is more routinely performed using two-dimensional separations, which were not carried out here. Further results, focussing on differential expression, use only iTRAQ and label-free data.

Figure 2 represents the differential expression of proteins as characterised by iTRAQ labelling, plotted on a $\log_e$ scale; the values are included alongside protein identifications in **Supplementary Data, Table 3**. Tags 115 and 117, which correspond to methane-grown samples, and tag 116, which corresponds to an acetate-grown sample, were normalised to tag 114, which corresponds to an acetate-grown sample. The values for the 116 sample are clustered close to a line along the x-axis as would be expected since the 114 and 116 samples should be identical. The 115 and 117 samples should also be identical and we would therefore expect good agreement between their ratios, as is observed. This experiment provides a good indication of the reproducibility of the iTRAQ approach. As can be seen by the 115 and 117 trends, distinct up- and down-regulated proteins may be identified in *M. silvestris* when grown under methane as compared to when the organism is grown under acetate. The standard deviation of all the 116:114 ratios is 0.17, providing an indication of what can be considered true up- or down-regulation. If

these values are considered to be a normal distribution around a calculated mean of 0, then any proteins with 115:114 and 117:114 ratios within -0.5 and 0.5 cannot be said to be regulated, using the value of three standard deviations to provide filtering parameters. Only those identifications showing ratios outside these values have been accepted as up- or down-regulated.

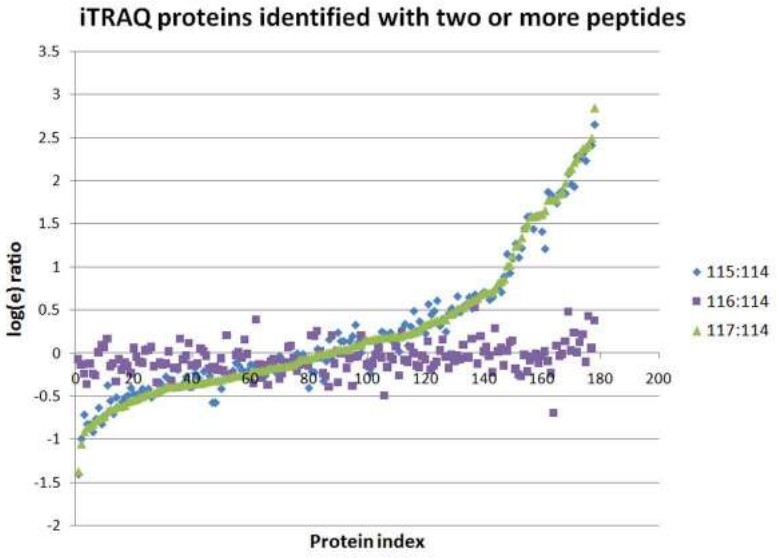


Figure 2: Differential expression as determined by iTRAQ labelling; all tags have been normalised to the 114 label.

In the iTRAQ method, samples from different growth conditions are pooled together. Quantification depends entirely upon the isobaric tags; if insufficient data is available from the isobaric tags, the protein identification will still be provided in the overall results table, but will not appear in the quantification data. In the label-free system, samples from differing growth conditions are kept separate, so a distinct set of protein identifications is generated for each sample. 231 proteins were identified unique to the acetate-grown sample, 70 were unique to the methane-grown sample and 124 were common to both conditions. Data relating to these 124 proteins were then processed to provide information on relative expression between the samples.

Figure 3 shows the relative protein expression for the regulated proteins (common to acetate and methane substrates) identified using the label-free approach; this is the output from the relative quantification software, which generates peptide signal intensity measurements, using all the peptides identified for any particular protein identification. These represent deisotoped, charge-state reduced and accurately mass measured ion lists, which are used for both qualitative identification and relative quantification [28]. $Log_{(e)}$ values used as the quantitative measurement can be found in **Supplementary Data, Table 6**, including indication of proteins assigned to only one of the two growth conditions. Error measurements are automatically generated as standard deviation values, which have been plotted. For an MS^E acquisition, the technical variation with respect to signal intensity has been shown to be 10-15% with highly consistent reproducibility [27, 38]. For the label-free quantitative data, the significance level of regulation was determined at 30% fold change, which is an average relative fold change between -0.3 and 0.3 on a natural log scale [38]. This is typically 2-3 times higher than the estimated error on the intensity measurement [27, 38, 39]. Those identifications with relative expression values between -0.3 and 0.3 cannot be taken as regulated; only those identifications outside these values can be said to be regulated.

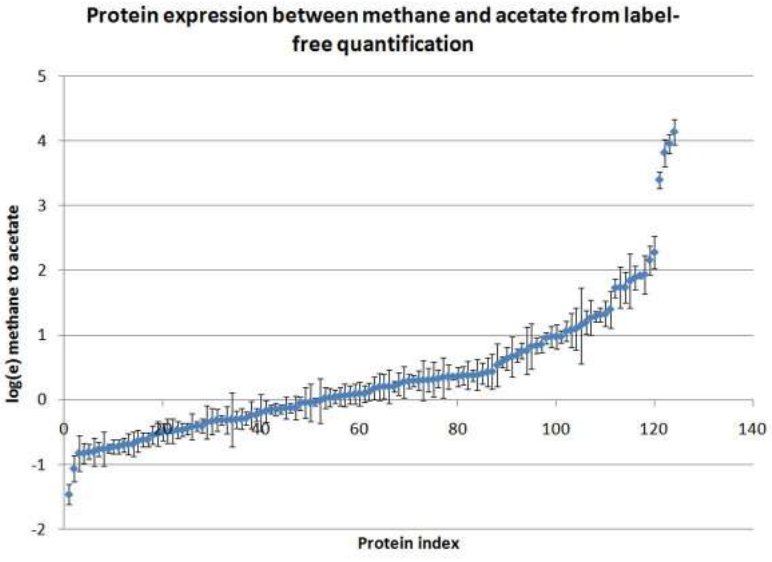


Figure 3: Automated protein-level quantification of regulated proteins using the label-free system; error bars correspond to the automatically generated standard deviation values.

Both iTRAQ and label-free allow profiling and relative quantitative data to be concurrently collected. The ability to do this, particularly in a high throughput manner, is desirable but often difficult. In total, 79 confident identifications (i.e. more than one peptide) are common between the two methods, which is a much larger overlap than previous studies comparing methods of quantification [14, 40]. A scatter plot comparing the regulation as assessed by the two methods is shown in **Supplementary Figure 2a**. There is reasonable correlation, with an R^2 value of 0.69, with one distinct outlier. If the overall trend of regulation is compared, all of the common identifications for which quantification data is available are in agreement, bar the outlier. 21 proteins are indicated to be up-regulated in the methane sample compared to acetate and 6 are indicated to be down-regulated; the remaining proteins show no distinct differential expression when filters for both datasets are applied. If the one distinct outlier is removed from the dataset then the correlation improves significantly (R^2 value 0.80), as shown in **Supplementary Figure 2b**. The outlier, corresponding to the enzyme citrate synthase, presented down-regulation in the methane-grown sample according to the label-free analysis, but up-regulation according to iTRAQ. Interrogation of the raw data showed good correlation between all three replicates of the label-free acquisition in both growth conditions. In the iTRAQ data, however, there was a disparity in the data from the isobaric tags. Five peptides were used for identification, with quantification data available for four of these. Three peptides showed down-regulation in the methane growth condition; the one peptide which indicated up-regulation was the shortest of the five (four residues), the others matching at least eight residues within the assigned MS/MS spectrum. If the short peptide is removed, there is down-regulation of citrate synthase within the filtering parameters, and in-line with the label-free data, suggesting that this was a mis-assignment by the software. MS/MS spectra of the matched sequences and isobaric tags are shown in **Supplementary Figure 3**. Although this is only one anomalous data point, it indicates potential problems if single-peptide identifications are used to provide quantitative data from an iTRAQ experiment.

The label-free approach differs from iTRAQ in that each growth condition is analysed independently, while in iTRAQ samples from different conditions are pooled together. Of the 425 non-redundant

identifications obtained by the label-free method, 231 were unique to the acetate-grown sample and 70 unique to the methane-grown. From these 301 proteins, 54 were also identified by iTRAQ, and were compared with the iTRAQ quantification list. Of these, 25 were distinctly regulated and all showed agreement, i.e. were shown by iTRAQ to be up-regulated in whichever growth condition the label-free method had exclusively assigned. This has been represented as a comparative table in Supplementary Data, Table 7. The 29 identifications which fall outside the iTRAQ filtering parameters for accepted regulation levels, as described earlier, have been highlighted.

Comparison of experimental approaches

A summary of the methodology for all three experimental systems and the results obtained from each can be seen in **Table 2**. There is a stark difference in the total amount of protein required for the three setups: up to 1 mg for iTRAQ, 14 μg for the gel-based method, and less than 1 μg for label-free. Although the injection amount for the LC-MS/MS analysis is comparable between all three techniques, this does not relate to the total amount of protein required for an adequate dataset. In the gel-based and iTRAQ approaches, the amount indicated is necessary to generate enough peptides over 30-60 fractions for MS analysis. With the employed label-free method, the amount loaded directly for LC-MS analysis is sufficient for a full qualitative and quantitative dataset. Sample requirement can be an important consideration when performing proteomic studies, as it can be a challenge to generate a suitable amount from biological systems. If less sample is required for a single experiment, additional analyses can be carried out, which will add confidence to the results obtained [41]. It has previously been shown that even three replicate MudPIT experiments may not provide full coverage of all the proteins within a sample [42].

An ideal method for proteomic analysis would enable comprehensive and high-throughput studies, making experimental and instrumentation time an important factor when considering which approach to

utilise. Both the gel-based and iTRAQ setups require up to 60 hours of MS data acquisition time, based upon our chosen number of bands cut from the gel or fractions from the strong cation exchange chromatography, and upon the gradient setup in the reversed-phase chromatography. The analytical time could be shortened by choosing fewer fractions, or reducing the reversed-phase gradient, but this may also reduce peptide recovery and/or separation. The label-free experiments require 6 hours of instrument time (2 hours per replicate). In addition to this, preparing samples for iTRAQ requires a number of days, including overnight steps. This issue can make the approach less suitable for a routine analysis setup when compared to the label-free method.

The average number of peptides identified per confident protein assignment for the gel-based and iTRAQ analyses is 5, compared to an average of 12 for the label-free method. The gel-based approach gives an average sequence coverage of 15%, higher than the iTRAQ average of 11% which is slightly lower than previous work [2]. The average sequence coverage for the label-free data is 45%. An increased number of peptides and higher sequence coverage can confer more confidence in identifications obtained.

	1D-SDS-PAGE	**iTRAQ**	**Label-free**
Protein loading	14 µg	100 µg per iTRAQ labelling vial; 800 µg total loading	0.5 µg for each of 3 technical replicates
Number of overnight steps	2	5	1
Samples to analyse by MS	30-40 fractions	30-60 fractions	1 per growth condition
Reverse-phase LC and MS acquisition	30-40 hours	30-60 hours	2 hours
Total analysis time	4 days	6 days	Less than 3 days
Total instrument time	30-40 hours	30-60 hours	6 hours per sample
Size of data file	300 MB x 40 (1.2 GB)	300 MB x 40 (1.2 GB)	6 GB x 3 (18GB)
Number of proteins confidently identified (with more than one peptide)	235	178	421

Average number of peptides per protein (including single-peptide identifications)	5	5	12
Average sequence coverage	15%	11%	45%

Table 2: A comparison of the experimental requirements for each of the approaches, and the information obtained from the data generated.

Biological significance of results obtained

Bacteria from the methanotroph family utilise a common pathway to process methane in order to use it as a carbon and energy source, an overview of which is given in **Figure 4**. As the *Methylocella* genus has been recently identified and is relatively uncharacterised, it is difficult to make predictions about potential biochemical changes which would be seen on a growth substrate other than methane. It could, however, be suggested that some down-regulation of the enzymes in the methane oxidation pathway would be seen. Our study identified the key enzymes methane monooxygenase (MMO) and methanol dehydrogenase, with quantitative data from both iTRAQ and the label-free approach indicating a significant down-regulation when *M. silvestris* was grown on acetate. MMO is a multimeric protein with subunits α, β and γ [37]. The α and β subunits show up-regulation in the methane growth samples, as does the accessory MMO Protein B. The γ subunit shows significant up-regulation on methane when analysed by iTRAQ; using our data filtering (more than one peptide, more than one replicate for label-free) this subunit is only seen in the methane growth condition for the gel and label-free analyses, as is the accessory MMO Protein C. There is also up-regulation of the alpha and beta subunits of methanol dehydrogenase in the methane-grown samples, which is the second enzyme in the methanotroph methane oxidation pathway.

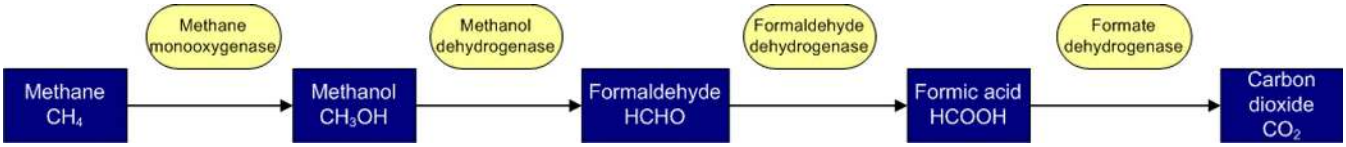
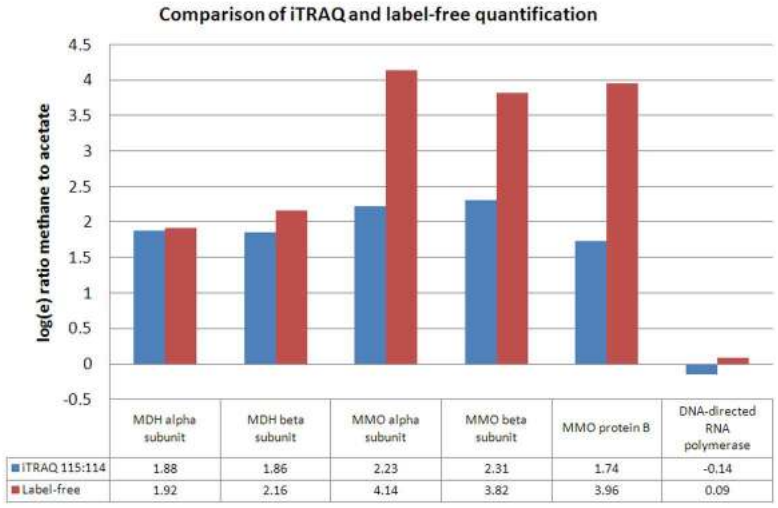


Figure 4: The pathway of methane oxidation in methanotrophic bacteria.

The quantification data relating to these enzyme identifications has been shown in **Figure 5**. DNA-directed RNA polymerase has also been included as a housekeeping protein, and as such should not display up- or down-regulation regardless of growth substrate. Such proteins can provide a check, on the biological level, for the significance of differential proteomic data.



	MDH alpha subunit	MDH beta subunit	MMO alpha subunit	MMO beta subunit	MMO protein B	DNA-directed RNA polymerase
iTRAQ 115:114	1.88	1.86	2.23	2.31	1.74	-0.14
Label-free	1.92	2.16	4.14	3.82	3.96	0.09

Figure 5: iTRAQ and label-free quantification data for proteins identified as key metabolic enzymes within *M. silvestris*; a housekeeping protein has been included as indication of a biological marker.

Conclusions

For any integrated proteomics experiment, a number of important issues need to be considered. These include the need for qualitative (profiling) and quantitative information. Confidence in identification and quantification, reproducibility, sample size, instrument time, sample preparation, cost, and sequence

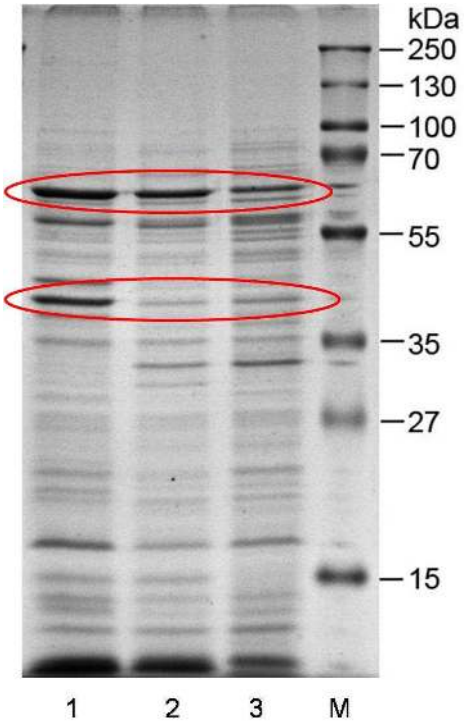
coverage are all important factors that need to be taken into account. The ability to place any changes observed into the context of the biological pathways involved remains a crucial aspect of the research. This study has evaluated the potential applicability of a number of common approaches to profiling and differential proteomics. The experiments have been restricted to a proteomics study of cytosolic proteins, and comparable technology platforms were employed. Good agreement was obtained between the commonly utilised iTRAQ labelled experiment, a gel based study and that based on a label-free LC-MS approach. At the profiling level, when considering all identifications, including those based on single peptides, the number of identified proteins was comparable for all three methods. When requiring more than one peptide for identification, the label-free approach gave superior information particularly when coverage was taken into account. Both the iTRAQ experiment and the label-free approach provided relative quantification datasets, and the agreement between the approaches was better than previously observed in comparisons between different quantitative methods [40]. This is most likely due to the use of comparable instrumentation, as each method employed high-performance liquid chromatography coupled to a Q-TOF tandem MS acquisition. The label-free experiment does, however, have advantages in terms of sample requirement, sample preparation and instrumental time requirements.

A preliminary screen of the protein regulation results for biological significance shows agreement with previous analysis of the regulation of methane monooxygenase in *Methylocella* [43]. This, together with the significant number of identifications provided by all three approaches, and the excellent agreement of two quantitative datasets, indicates the potential for further proteomic studies on this methanotroph.

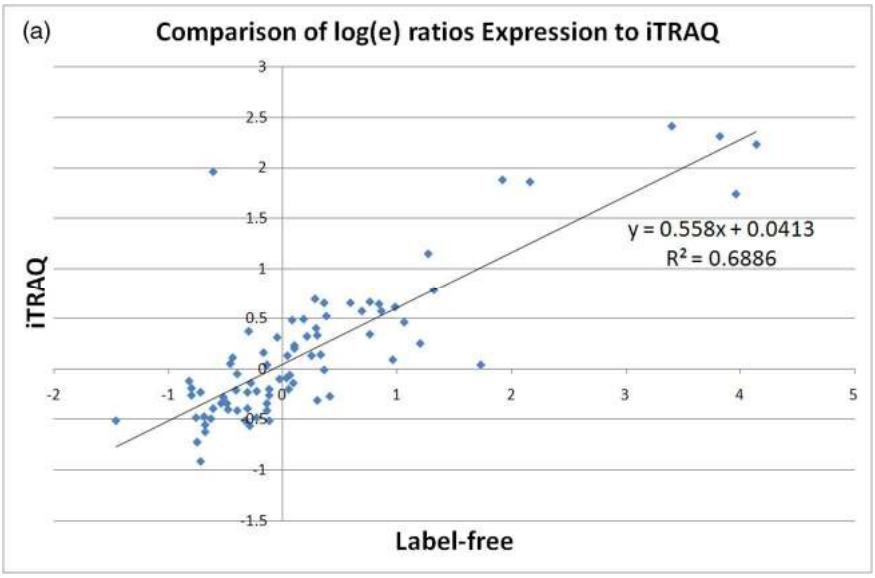
Acknowledgements

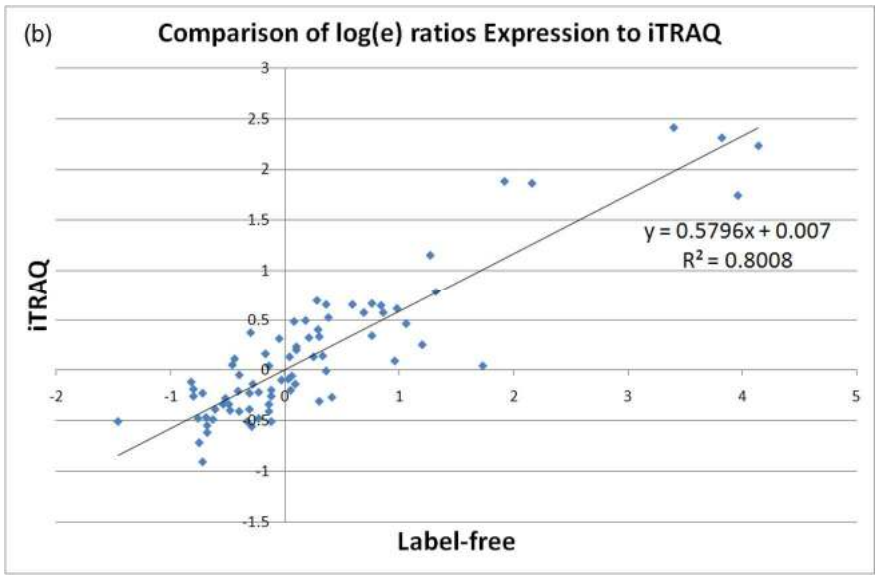
Vibhuti Patel thanks Eleanor Blatherwick (Biological Mass Spectrometry and Proteomics Facility, Department of Biological Sciences, Warwick) for assistance with experimental work. Vibhuti Patel and Konstantinos Thalassinos thank Hans Vissers (Waters) for helpful discussions regarding the manuscript.

Supplementary data

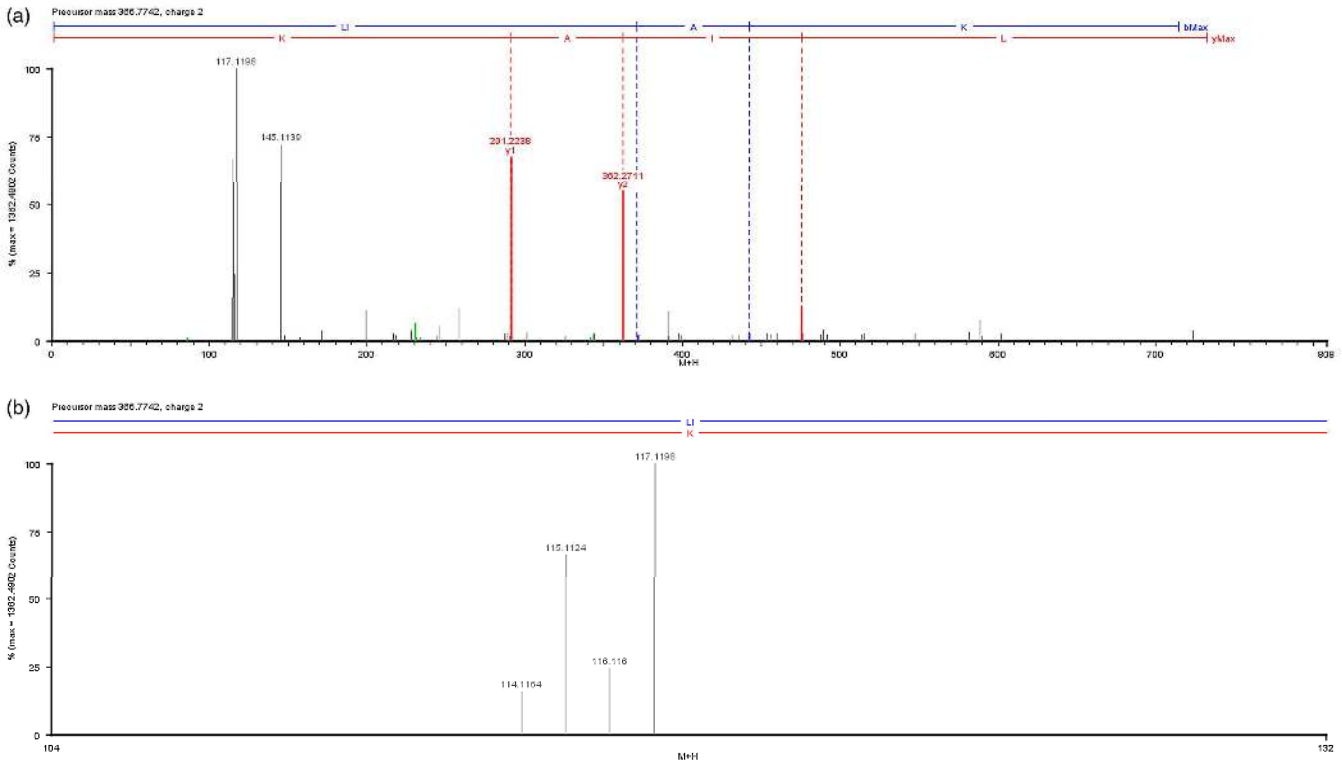


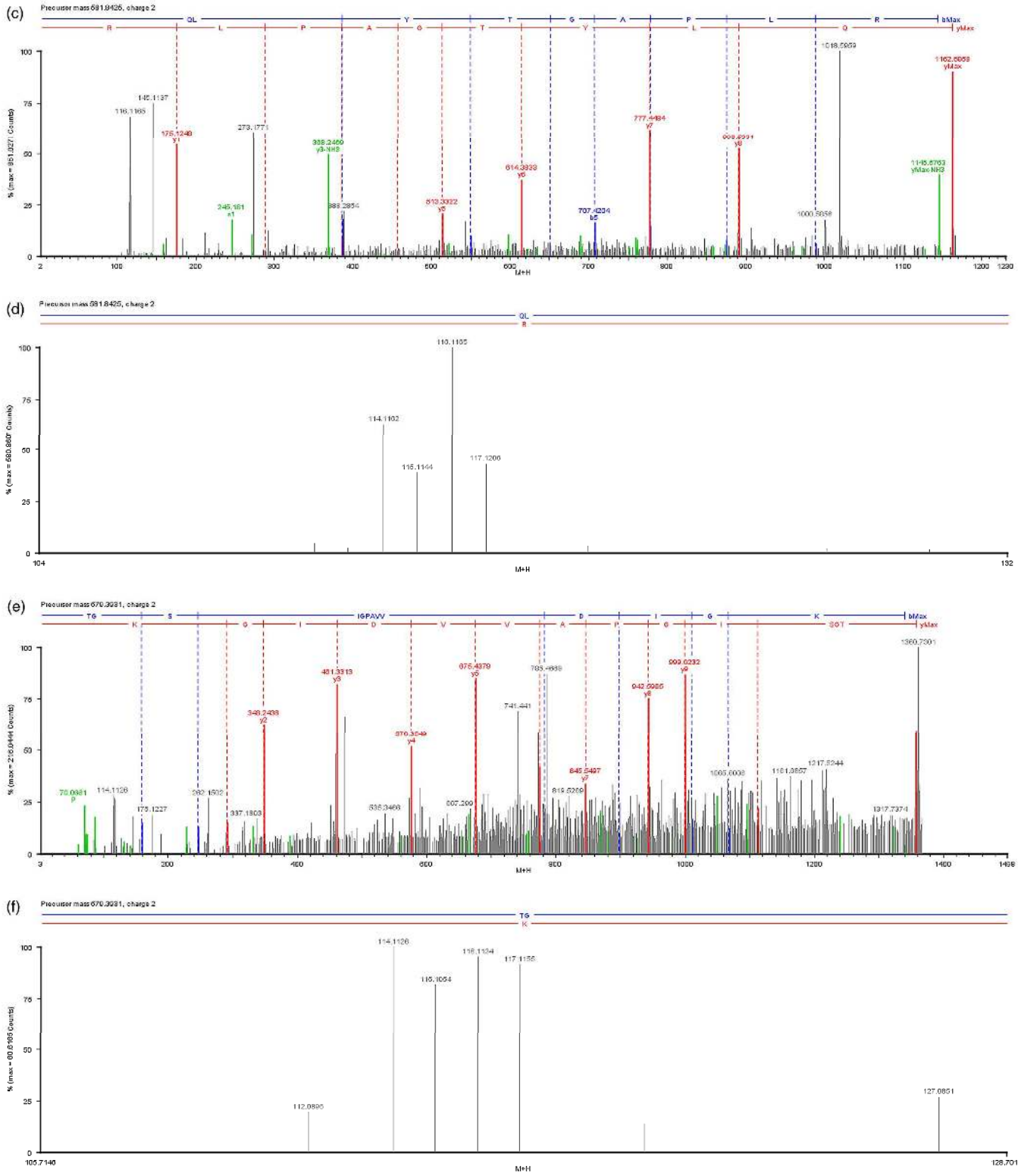
Supplementary Figure 1: 1D SDS-PAGE separation of the *M. silvestris* cytoplasmic proteome under different growth conditions; 1 - methane-grown, 2- succinate-grown (further data not shown), 3 - acetate-grown, M - molecular mass markers. The upper bands highlighted in red indicate how multiple bands can be unresolved; the lower bands may be indicative of differential expression.

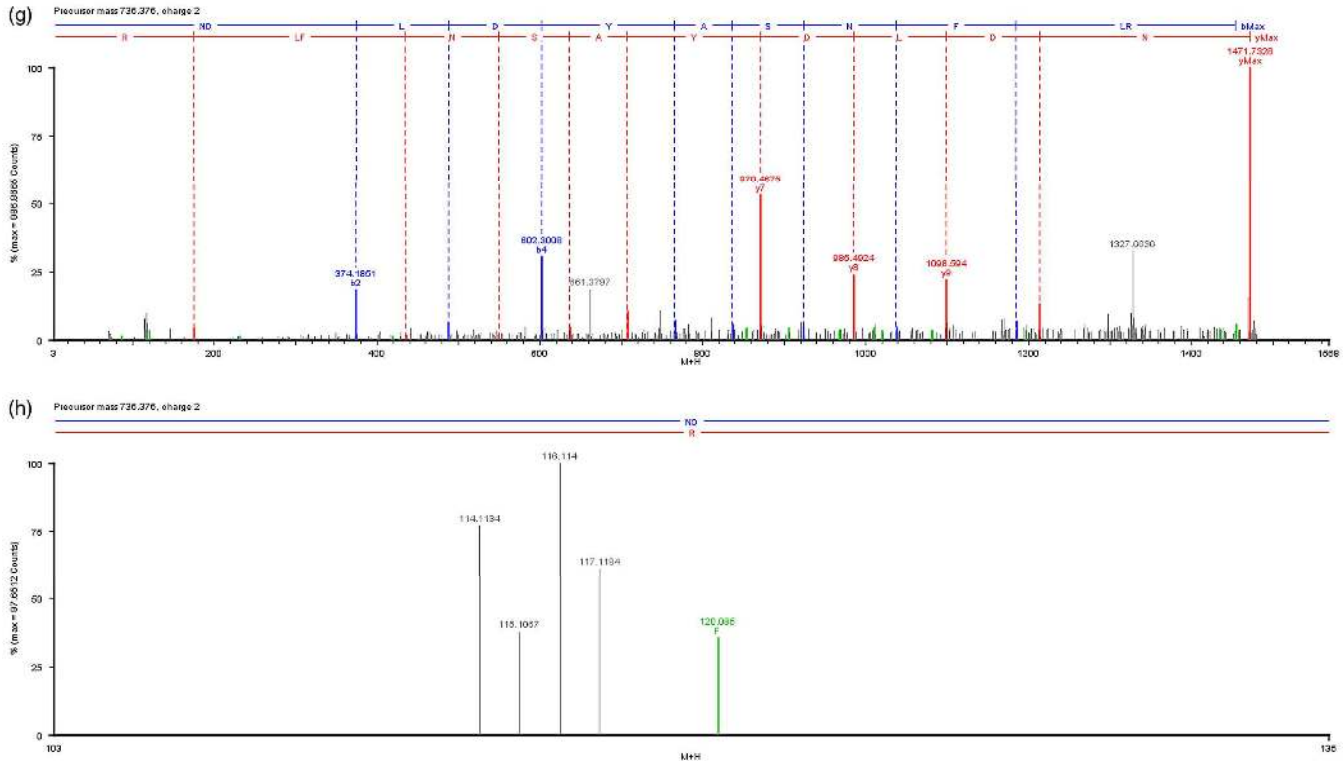




Supplementary Figure 2: (a) Correlation of quantification data from iTRAQ and the label-free method for identifications using two or more peptides; **(b)** Correlation when the outlier corresponding to citrate synthase is removed.




(c)
Precursor mass 681.8425, charge 2
QL
YTGAPLR
RLPAGTGYLQ
bMax
yMax
% (max = 951.927 Counts)
M+H
(d)
Precursor mass 681.8425, charge 2
QL
R
% (max = 580.8807 Counts)
M+H
(e)
Precursor mass 679.3981, charge 2
TG
KSGIGPAVVDVVAPGIGK
SGT
% (max = 216.0464 Counts)
M+H
(f)
Precursor mass 679.3981, charge 2
TG
K
% (max = 60.5165 Counts)
M+H



Supplementary Figure 3: Peptide sequences for precursors and corresponding isobaric reporter ions for outlying protein identification citrate synthase **(a)** Mis-assigned peptide, precursor *m/z* 366, **(b)** reporter ions for precursor 366; **(c)** precursor *m/z* 581, **(d)** reporter ions for precursor 581; **(e)** precursor *m/z* 679, **(f)** reporter ions for precursor 679; **(g)** precursor *m/z* 736, **(h)** reporter ions for precursor 736.

References

1. Pandey, A. and M. Mann, *Proteomics to study genes and genomes.* Nature, 2000. **405**(6788): p. 837-846.
2. Domon, B. and R. Aebersold, *Mass Spectrometry and Protein Analysis.* Science, 2006. **312**(5771): p. 212-217.
3. Wilkins, M.R., et al., *Guidelines for the next 10 years of proteomics.* Proteomics, 2006. **6**(1): p. 4-8.
4. Ong, S.-E. and A. Pandey, *An evaluation of the use of two-dimensional gel electrophoresis in proteomics.* Biomolecular Engineering, 2001. **18**(5): p. 195-205.
5. Washburn, M.P., D. Wolters, and J.R. Yates, *Large-scale analysis of the yeast proteome by multidimensional protein identification technology.* Nature Biotechnology, 2001. **19**(3): p. 242-247.
6. Ünlü, M., M.E. Morgan, and J.S. Minden, *Difference gel electrophoresis. A single gel method for detecting changes in protein extracts.* Electrophoresis, 1997. **18**(11): p. 2071-2077.
7. Ong, S.-E., et al., *Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics.* Molecular and Cellular Proteomics, 2002. **1**(5): p. 376-386.
8. Fairwell, T., et al., *Sequence analysis of complex protein mixtures by isotope dilution and mass spectrometry.* Biochemistry, 1970. **9**(11): p. 2260-2267.
9. Kuhn, E., et al., *Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and 13C-labeled peptide standards.* Proteomics, 2004. **4**(4): p. 1175-1186.
10. Sirlin, J.L., *On The Incorporation Of Methionine 35S Into Proteins Detectable By Autoradiography.* Journal of Histochemistry and Cytochemistry, 1958. **6**(3): p. 185-190.
11. Gerber, S.A., et al., *Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(12): p. 6940-6945.
12. Thompson, A., et al., *Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS.* Analytical Chemistry, 2003. **75**(8): p. 1895-1904.
13. Gygi, S.P., et al., *Quantitative analysis of complex protein mixtures using isotope-coded affinity tags.* Nature Biotechnology, 1999. **17**(10): p. 994-999.
14. Ross, P.L., et al., *Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-reactive Isobaric Tagging Reagents.* Molecular and Cellular Proteomics, 2004. **3**(12): p. 1154-1169.
15. Ow, S.Y., et al., *Quantitative shotgun proteomics of enriched heterocysts from Nostoc sp. PCC 7120 using 8-plex isobaric peptide tags.* Journal of Proteome Research, 2008. **7**(4): p. 1615-1628.
16. Sadowski, P.G., et al., *Quantitative proteomic approach to study subcellular localization of membrane proteins.* Nature Protocols, 2006. **1**(4): p. 1778-89.
17. Panchaud, A., et al., *Experimental and computational approaches to quantitative proteomics: Status quo and outlook.* Journal of Proteomics, 2008. **71**(1): p. 19-33.
18. Allet, N., et al., *In vitro and in silico processes to identify differentially expressed proteins.* Proteomics, 2004. **4**(8): p. 2333-2351.
19. Liu, H., R.G. Sadygov, and J.R. Yates, *A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics.* Analytical Chemistry, 2004. **76**(14): p. 4193-4201.
20. Colinge, J., et al., *Differential Proteomics via Probabilistic Peptide Identification Scores.* Analytical Chemistry, 2005. **77**(2): p. 596-606.

21. Rappsilber, J., et al., *Large-Scale Proteomic Analysis of the Human Spliceosome.* Genome Research, 2002. **12**(8): p. 1231-1245.
22. Ishihama, Y., et al., *Exponentially Modified Protein Abundance Index (emPAI) for Estimation of Absolute Protein Amount in Proteomics by the Number of Sequenced Peptides per Protein.* Molecular and Cellular Proteomics, 2005. **4**(9): p. 1265-1272.
23. Chelius, D. and P.V. Bondarenko, *Quantitative Profiling of Proteins in Complex Mixtures Using Liquid Chromatography and Mass Spectrometry.* Journal of Proteome Research, 2002. **1**(4): p. 317-323.
24. Müller, C., et al., *Ion suppression effects in liquid chromatography-electrospray-ionisation transport-region collision induced dissociation mass spectrometry with different serum extraction methods for systematic toxicological analysis with mass spectra libraries.* Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences, 2002. **773**(1): p. 47-52.
25. Wang, W., et al., *Quantification of Proteins and Metabolites by Mass Spectrometry without Isotopic Labeling or Spiked Standards.* Analytical Chemistry, 2003. **75**(18): p. 4818-4826.
26. Radulovic, D., et al., *Informatics Platform for Global Proteomic Profiling and Biomarker Discovery Using Liquid Chromatography-Tandem Mass Spectrometry.* Molecular and Cellular Proteomics, 2004. **3**(10): p. 984-997.
27. Silva, J.C., et al., *Quantitative Proteomic Analysis by Accurate Mass Retention Time Pairs.* Analytical Chemistry, 2005. **77**(7): p. 2187-2200.
28. Silva, J.C., et al., *Absolute Quantification of Proteins by LCMSE: A Virtue of Parallel ms Acquisition.* Molecular and Cellular Proteomics, 2006. **5**(1): p. 144-156.
29. Silva, J.C., et al., *Simultaneous Qualitative and Quantitative Analysis of theEscherichia coli Proteome: A Sweet Tale.* Molecular and Cellular Proteomics, 2006. **5**(4): p. 589-607.
30. Cheng, F.-Y., et al., *Absolute Protein Quantification by LC/MSE for Global Analysis of Salicylic Acid-Induced Plant Protein Secretion Responses.* Journal of Proteome Research, 2009. **8**(1): p. 82-93.
31. Theisen, A.R., et al., *Regulation of methane oxidation in the facultative methanotroph Methylocella silvestris BL2.* Molecular Microbiology, 2005. **58**(3): p. 682-692.
32. Link, A.J., et al., *Direct analysis of protein complexes using mass spectrometry.* Nature Biotechnology, 1999. **17**(7): p. 676-682.
33. Geromanos, S.J., et al., *The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS.* Proteomics, 2009. **9**(6): p. 1683-1695.
34. Li, G.-Z., et al., *Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures.* Proteomics, 2009. **9**(6): p. 1696-1719.
35. Veenstra, T.D., T.P. Conrads, and H.J. Issaq, *What to do with one-hit wonders?* Electrophoresis, 2004. **25**(9): p. 1278-1279.
36. Breci, L., et al., *Comprehensive proteomics in yeast using chromatographic fractionation, gas phase fractionation, protein gel electrophoresis, and isoelectric focusing.* Proteomics, 2005. **5**(8): p. 2018-2028.
37. Colby, J. and H. Dalton, *Some properties of a soluble methane mono-oxygenase from Methylococcus capsulatus strain Bath.* Biochemical Journal, 1976. **157**(2): p. 495-497.
38. Vissers, J.P.C., J.I. Langridge, and J.M.F.G. Aerts, *Analysis and Quantification of Diagnostic Serum Markers and Protein Signatures for Gaucher Disease.* Molecular and Cellular Proteomics, 2007. **6**(5): p. 755-766.
39. Chambery, A., et al., *Qualitative and Quantitative Proteomic Profiling of Embryonic Stem Cells by Means of Accurate Mass LC-MS Analysis.* Journal of Proteome Research, 2009. **8**(2): p. 1047-1058.

40. Usaite, R., et al., *Characterization of Global Yeast Quantitative Proteome Data Generated from the Wild-Type and Glucose Repression Saccharomyces cerevisiae Strains: The Comparison of Two Quantitative Methods.* Journal of Proteome Research, 2008. **7**(1): p. 266-275.
41. Karp, N.A. and K.S. Lilley, *Design and analysis issues in quantitative proteomics studies.* Proteomics, 2007. **7 Suppl 1**: p. 42-50.
42. Durr, E., et al., *Direct proteomic mapping of the lung microvascular endothelial cell surface in vivo and in cell culture.* Nature Biotechnology, 2004. **22**(8): p. 985-992.
43. Trotsenko, Y.A., et al., *Metabolic Aspects of Aerobic Obligate Methanotrophy*, in *Advances in Applied Microbiology*. 2008, Academic Press. p. 183-229.

Table of contents synopsis

Three profiling and quantitative proteomics approaches have been utilised to study the recently discovered bacterium *Methylocella silvestris* when grown under two different substrates. Experimental conditions have been compared, and the data obtained has been evaulated with respect to number of proteins identified, confidence in identification, sequence coverage and agreement of regulated proteins. A preliminary screen of the protein regulation results for biological significance has also been performed.