Review

Organelle proteomics experimental designs and analysis

Laurent Gatto[1], *Juan Antonio Vizcaíno*[2], *Henning Hermjakob*[2], *Wolfgang Huber*[2,3] *and Kathryn S. Lilley*[1]

[1] Cambridge Centre for Proteomics, Cambridge Systems Biology Centre, Department of Biochemistry, University of Cambridge, Cambridge, UK
[2] EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
[3] EMBL Genome Biology Unit, Heidelberg, Germany

In biology, localisation is function: knowledge of the localisation of proteins is of paramount importance to assess and study their function. This supports the need for reliable protein sub-cellular localisation assignment. Concomitant with recent technological advances in organelle proteomics, there is a requirement for more rigorous experimental and analysis design planning and description. In this review, we present an overview of current experimental designs in qualitative and quantitative organelle proteomics as well as associated data analysis. We also consider the major benefits associated with careful description and dissemination of the experiment and analysis designs, namely (i) comparison and optimisation of experimental designs and analysis pipelines, (ii) data validation, (iii) reproducible research, (iv) efficient repository submission and retrieval and (v) meta analysis. Formalisation of experimental design and analysis work flows is of direct benefit for the organelle proteomics researchers and will result in providing organelle localisation data of highest quality for the wider research community.

Received: April 14, 2010
Revised: June 12, 2010
Accepted: July 7, 2010

Keywords:
Bioinformatics / Data analysis / Data annotation / Experimental design / Metadata / Organelle proteomics

1 Introduction

To gain a full picture of the proteome of any cell, it is essential to investigate the sub-cellular location(s) of each protein (localizome [1]). Indeed, knowing the sub-cellular location of a protein is of paramount importance to biologists in the elucidation of its role and in the refinement of knowledge of cellular processes. Co-localisation of proteins within the same sub-cellular structure is necessary for them to interact, and may in many cases be the decisive factor in conferring specificity to an interaction [2]. As such, co-localisation is also an extremely effective technique to assess the function of an unknown protein [3] (functional proteomics, [4]). Organelle proteomics is a key discipline in the study of cell biology.

Organelle-associated proteins can be localised in the lumen of an organelle or at its membrane. Integral membrane proteins either have one or have more transmembrane domains that span across the lipid bilayer (transmembrane proteins) or are associated with one side of the membrane through α-helix or lipid chains. Peripheral proteins are more or less loosely attached to the membrane through non-covalent bounds or by covalent attachment to lipids. One third of the total proteome is believed to be associated with the membrane in some form or other [5–8]. The membrane proteome contains many less-abundant proteins and its study can reveal information not easily obtained when analysing a total cell lysate. Plasma membrane proteins are of particular interest: as they are located at the cell surface and form the interface with other

Correspondence: Dr. Kathryn S. Lilley, Cambridge Centre for Proteomics, Cambridge System Biology Centre University of Cambridge, Department of Biochemistry, Tennis Court Road, Cambridge, CB2 1QR, UK
E-mail: k.s.lilley@bioc.cam.ac.uk
Fax: +44-1223-760241

Abbreviations: GFP, green fluorescent protein; **LOPIT,** localisation of organelle proteins by isotope tagging; **PCA,** principle component analysis; **PCP,** protein correlation profile; **PLS-DA,** Partial least square discriminant analysis; **QC,** quality control cells, they are responsible for cell–cell interactions. Proteins in numerous different types of sub-cellular membranes act as receptors and transporters and enable and regulate traffic across membrane boundaries. Proteins in the cell membrane act as anchors for cytoskeletal proteins or the extracellular matrix.

Membrane proteins are characterised by poor solubility and exist in a wide range of concentrations, which make them particularly challenging to analyse with many proteomic techniques [8]. Moreover, contamination during organelle enrichment is a serious issue that has triggered the development of new approaches in organelle proteomics. Another confounding factor in the study of organelles proteomes is the fact that the protein content of organelles is not static. Some proteins may be in transit through an organelle or may interact transiently with resident proteins. Furthermore, the protein composition of organelles will vary depending on cell type [9–11], metabolic or environmental condition [12].

The increasing interest of the research community in organelle proteomics (reviewed among others by [13–18]) is illustrated by the annual numbers of publications since 1996 (Fig. 1). In this review, we address some of the important open questions of the field: experimental designs (Section 2) and execution (Section 3), and data analysis (Section 4). In Section 5, we conclude with opinions about future analyses and the need for and requirements of a common data repository.

2 Experimental designs

Experimental design is arguably the most crucial part of any study [3, 19]. A carefully planned design, as shown in Fig. 2, should take several factors into account including biological question, sample type and number, organelles to be screened, qualitative (cataloguing) or quantitative assessment, available resources (technological and human), and sample preparation. Moreover, where multiple samples/replicates are used, well-defined analytical strategies are required. The latter is particularly important to guarantee that the results generated and interpreted are reproducible.

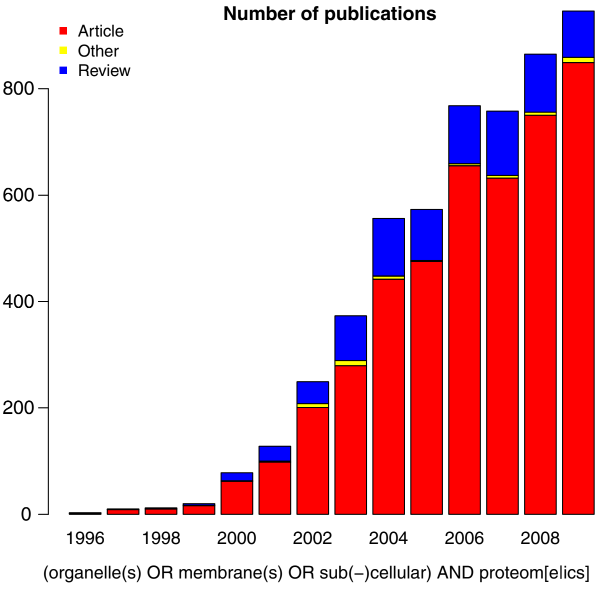

Figure 1. Number of publications *per* year from 1996 to 2009. The PubMed query is shown below the figure. ''Other'' means news articles, comments, letters, editorials, tutorials or lectures.

Techniques which allow characterisation of the protein complement of organelles fall into two main categories, those that involve tagging individual proteins followed by imaging (Section 2.1), and those which centre round the use of MS-based proteomic approaches. Figure 3 is a classification of the current most common experimental designs in organelle proteomics. Experimental designs which involve MS are further shown in Fig. 4 (for details, see Section 2.2).

2.1 Tagging approaches

The two main tagging approaches, dubbed ''tag and tell'' [20], focus on the direct observation of proteins in individual cells using green fluorescent protein (GFP)-tagged genetic constructs (fusion proteins) or fluorescently labelled antibodies specific to the protein of interest.

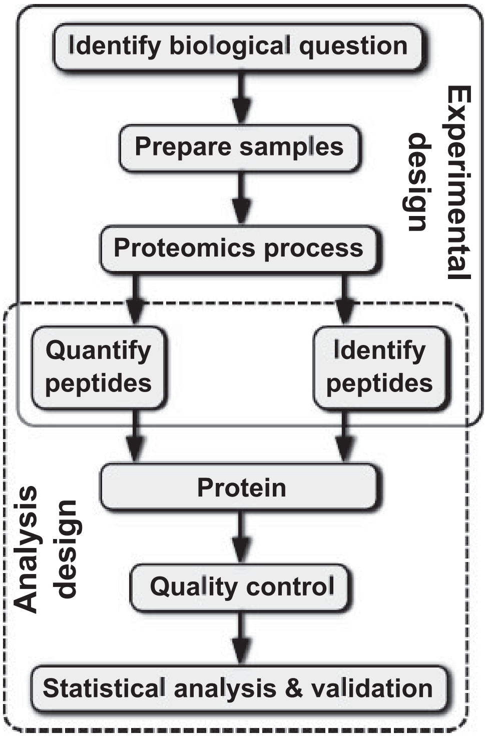

Figure 2. Schematic representation of typical experimental and analysis designs in a proteomic experiment.

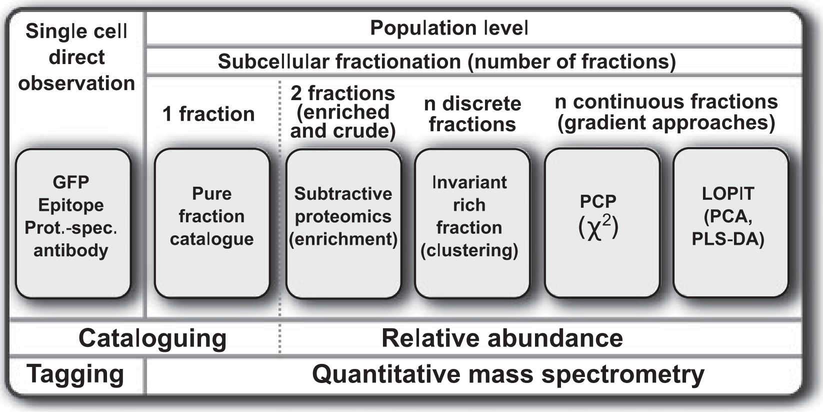
Single cell direct observation
Population level
Subcellular fractionation (number of fractions)
1 fraction
2 fractions (enriched and crude)
n discrete fractions
n continuous fractions (gradient approaches)
GFP Epitope Prot.-spec. antibody
Pure fraction catalogue
Subtractive proteomics (enrichment)
Invariant rich fraction (clustering)
PCP (χ^2)
LOPIT (PCA, PLS-DA)
Cataloguing
Relative abundance
Tagging
Quantitative mass spectrometry

Figure 3. Classification of experimental designs in organelle proteomics.

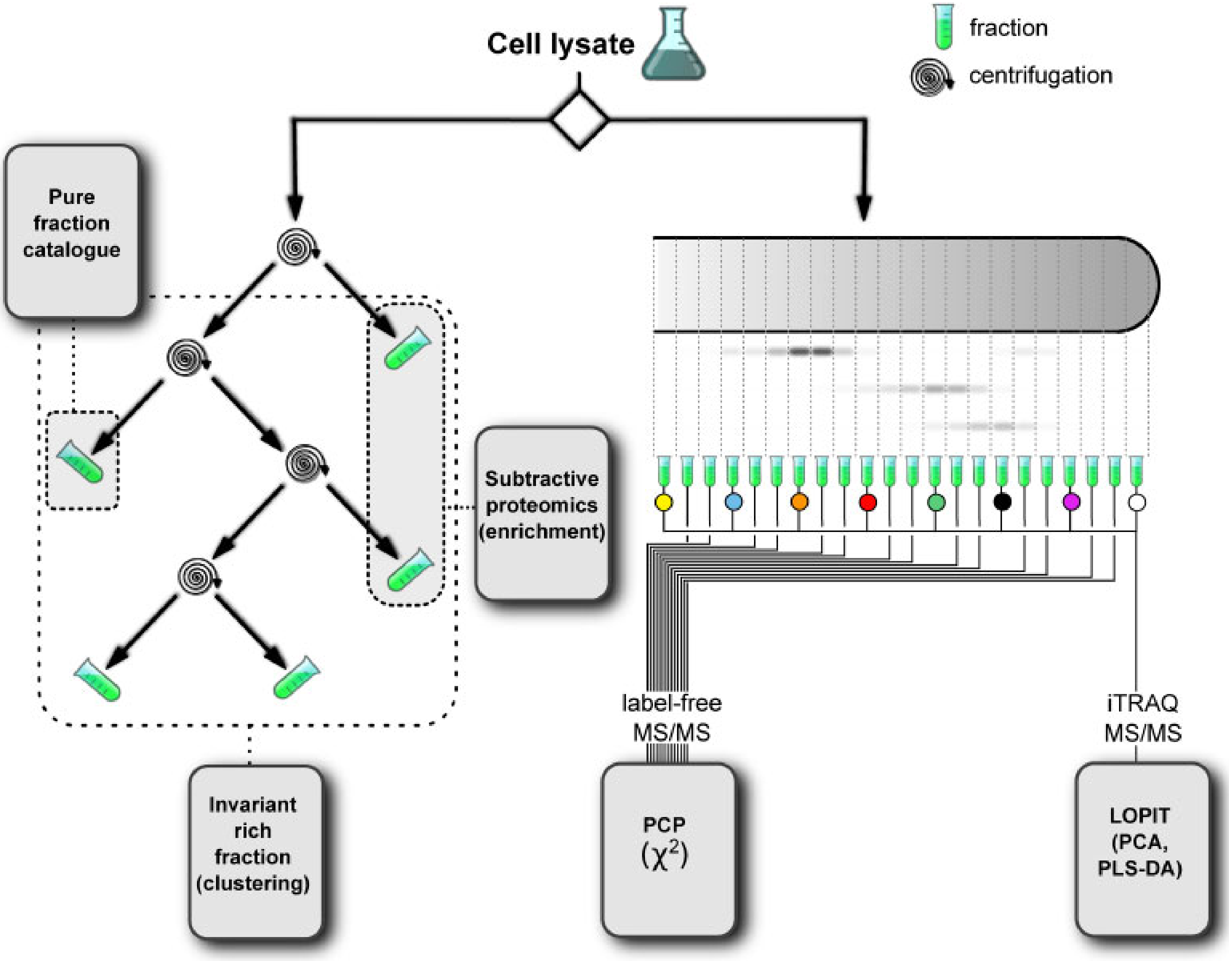

Figure 4. MS-based designs in organelle proteomics.

2.1.1 GFP tagging

Here, the GFP-coding sequence is inserted in-frame at the *C*- or *N*-terminal end of a coding sequence. The expression of the protein is generally under the control of the endogenous promoter, thus minimally perturbing the expression level of the protein. GFP tagging is most effective for soluble proteins, and difficulties may arise when used with transmembrane proteins that have to be correctly inserted within a membrane. Steric hindrance or prevention of post-translational modification due to the GFP tag can affect trafficking and result in mislocalisation. One of the most impressive applications of

GFP tagging was made by Huh *et al.* [3], who have systematically tagged the ORFs of *Saccharomyces cerevisiae* and collected localisation data for 75% of the yeast proteome. Later, Matsuyama *et al.* [21] performed a similar large-scale tagging of the *Schizosaccharomyces pombe* ORFeome with yellow fluorescent protein. Additional fluorescent proteins exist [22] that can be used to address co-localisation of multiple proteins inside a single cell.

2.1.2 Immunofluorescence and immunohistochemistry

Epitope tagging involves fusing a small immunoreactive epitope to a protein of interest using recombinant DNA technologies [23–25]. In this approach, however, partial ORF tagging, *i.e.* truncated products after transposon-mediated recombination, as well as over-expression of the tagged protein can result in mislocalisation. Alternatively, specific antibodies can be raised against a native protein [26, 27]. Barbe *et al.* [28] have targeted 466 proteins in three human cell lines using protein-specific antibodies. The latter approach does not require any genetic construct (as opposed to epitope and GFP tagging) while keeping the expression under endogenous promoter control (as does GFP). This method does require fixing and permeabilisation of cells prior to antibody staining which may lead to aberrant results [29]. Overlapping epitopes and generally off-target reactivity can also lead to mislocalisation.

From an analysis point of view, images have to be taken and individually analysed by possibly more than one scorer that assigns each protein an organelle. Advances in automatic or semi-automatic image analysis software assist in automated and high-throughput tag localisation [3, 22].

A disadvantage of the tagging methods is that they are currently very laborious and therefore limited with respect to the number of proteins covered. Various labour-intensive initiatives world wide strive to improve coverage of the proteome, but these are costly. Apart from the technical issues of the tagging techniques (genetic construction, raising-specific antibodies), the analysis of the resulting images can be challenging when differentiating background noise and multiple locations. Although these approaches have provided significant results, the technical difficulties hinder their direct application to more complex mammalian systems involving whole tissues.

2.2 MS approaches

There are four main MS-based techniques: these are the production of pure organelle fractions, subtractive proteomics, invariant-rich fraction and gradient approaches. They entail the creation of single or multiple fractions of a cell lysate to quantify and identify the protein content of a population of potentially heterogeneous cells. These approaches have been called ‘‘divide and identify’’ by Simpson and Pepperkok [20], as the cell lysate is fractionated (divided) and the proteins of different fractions are subsequently identified. The results are averaged over the different cell types within the initial sample.

2.2.1 Subcellular fractionation

The MS methods involve some form of subcellular fractionation where organelles are separated based on their physical or biological properties. The cells are first disrupted by a method which maintains the integrity of the organelles followed by separation of the organelle population [30–32]. Harsh homogenisation may result in rupturing of some organelles, thus increasing the distribution of parameters they will be separated upon, such as vesicle density, and resulting in the loss of the luminal proteins. Cytoplasmic aggregates can also form with cytoskeletal elements which in turn limit subsequent fractionation. Density gradient centrifugation is a popular approach that separates fractions based on size, density, charge and other properties [18, 33]. Other techniques that are applied alone or in concert include free-flow electrophoresis, immunoisolation, affinity chromatography or direct alteration of the physical properties of an organelle to enhance its extraction. Several organelles such as mitochondria, nuclei and plastids can be further purified to obtain nearly pure fractions. However, many endosomal compartments are similar with respect to the physical properties on which differential centrifugation is based and are as such impossible to separate. Western blotting of organelle specific markers or morphological analysis using (light, confocal or electron) microscopy can be applied to assess the purity of the fractions although lack of sensitivity of the former and the laborious nature of the latter make these suboptimal. Finally, proteins which traffic through multiple organelle compartments may be present in more than one location at any given moment, making it even more difficult to separate cargo or full-time residents from contaminants.

It is noteworthy that organelle purification is an efficient method to reduce overall protein complexity. Since protein abundance within a cell can differ up to ten orders of magnitude, lower abundant proteins such as regulatory elements (kinases, GTPases and membrane receptors) are often below the detection limit in total cell lysates. The enrichment of such proteins within sub cellular fractions aids their characterisation.

2.2.2 Pure fractions

This procedure aims at extracting a pure fraction of a given organelle and analysing it independently of other fractions to produce a catalogue of protein residents. Such approaches rarely use any type of quantitation, but simply list proteins as present in the purified sample. As mentioned above, a pure organelle fraction can rarely be achieved, with the exception of specific organelles such as mitochondria and chloroplasts, for which relatively high purity can be attained. A major disadvantage of considering a separate fraction independently from the other fractions produced as part of the purification protocol is that one is unable to assess contamination (and it is expected to be present). Multiple locations and dynamic aspects of organelle trafficking are hidden in this experimental design although these aspects are of primary biological importance. The purified organelle approach has, however, been widely used in the past and has enabled to set the first steps in the identification of organelle-specific complements [34–38].

2.2.3 Subtractive proteomics

Subtractive or differential proteomic analyses compare two fractions to assess the complement of a given organelle. This approach is used in situations where pure organelle fractions cannot be obtained, but where one fraction is highly enriched for the organelle of interest. A quantitative proteomic technique is used to measure the relative protein abundances by directly comparing enriched and crude fractions [14]. It is important that the two fractions under comparison share as much as possible of technical variability to avoid systematic biases that would invalidate the whole approach. Multiplexing the different fractions with SILAC-based isotope tagging [39], ICAT [40] or iTRAQ [41] reagents makes the results readily comparable. Schirmer *et al.* [42] successfully compared nuclear envelope fractions and fractions containing organelles known to co-fractionate with them. Marelli *et al.* [43] have identified yeast peroxisome-specific proteins after ICAT labelling. The distribution of ICAT ratios was modelled using Gaussian distributions, and a mixture model was fitted to the data to prioritise the candidates that are the most likely enriched. Wiederhold *et al.* isolated crude yeast vacuole membranes that were subjected to a second density centrifugation [44]. Both fractions were labelled with iTRAQ reagents and compared to reveal biologically relevant proteins. The relative protein abundances in the two fractions were statistically analysed to identify enriched protein clusters that were then considered to be true residents of the vacuole. Wiederhold *et al.* [44] used a modification of the rank product statistic [45] on the iTRAQ ratios to obtain a list of proteins ranked based on their enrichment. A modified version of iterative group analysis [46] was then used to estimate which proteins had co-enriched with known organelle markers.

Subtractive proteomics constitutes an improvement over the pure fraction approach as the risk and effect of contamination is explicitly included in the experimental design. It remains still necessary, however, as explained in Section 4, to compare the different statistical methods that have been applied to assess their respective efficiency, and whether they are optimally applicable for organelle proteomics. Subtractive proteomics focuses on a specific organelle of interest. A cell-wide picture of organelle proteomes is still lacking in this method and dynamic aspects or multiple locations are not taken into account.

2.2.4 Invariant rich fractions

A direct improvement of the analysis of two individual fractions is to compare several fractions in parallel. A set of discrete organelle-specific fractions are sequentially separated and purified by differential centrifugation (see Fig. 1 in [47] or [18]). It is important to note that in the process of discrete fraction sampling and purification, some material is discarded.

Gilchrist *et al.* [48] used redundant peptide counting [34] to semi-quantitatively compare proteins between smooth, rough microsome and Golgi fractions and assessed their intracellular distributions. The proteins were clustered based on their relative abundance in the different fractions and subsequently assigned a localisation based on matching the distribution to that of organelle marker proteins. This approach offers a means to identify relevant protein patterns across several fractions (and thus organelles) but it does not directly account for contamination, as it still, at least partially, relies on fraction purification. Kislinger *et al.* [10] simultaneously screened the protein content of four organelles in six mouse organs using crude organelle fractionation. Quantification was based on the peptide counting and protein clusters were identified at the organ and organelle levels.

Andreyev *et al.* [47] analysed five ''pure'' fractions and extracted sets of proteomic marker ensembles (*i.e.* a set of about 50 organelle markers) as representatives of these various sub-cellular compartments without relying on any *a priori* organelle specific marker. The inclusion of proteins to an organelle marker was based on its relative abundance in the various fractions. They then used the marker ensembles to estimate the organellar composition of the individual fractions. This methodology still relies directly on the efficiency of the organelle fractionation and thus suffers from its inherent limitations. The reliability of the marker sets is directly proportional to the fraction purity.

2.2.5 Gradient approaches

Utilising organelle-specific distribution patterns in centrifuged gradients is based on De Duve's principle, who observed that when centrifuging a cell lysate to enrich for organelles, the true residents of a given organelle were not present in a single fraction but were characterised by a distribution in several fractions, that is specific for their original sub-cellular localisation [49]. He postulated that based on the distribution of a marker that is known to be a genuine resident of an organelle (*a priori* knowledge) the other residents of the same organelle could be determined by matching their distribution to that of the marker. Gradient approaches entail quantification of proteins in as many fractions as possible to estimate their subcellular distributions. Several fractions are thus sampled along a continuous gradient, unlike the discrete fractions analysed in the invariant rich approach, and this method relies on organelle separation rather than purification. The distributions of individual proteins are compared with the distribution of known organelle markers. Matching of proteins of unknown location to organelle specific distribution patterns is carried out using statistical pattern recognition techniques.

2.2.5.1 Protein correlation profile

Protein correlation profile (PCP) is a label-free quantification gradient-based approach, first described by Andersen *et al.* [50] in a study of human centrosome. The abundance of each protein in each sucrose gradient fraction was calculated from the area of the extracted ion currents (label-free quantitative proteomics [51]) produced from peptides generated by trypsinolysis of proteins in each fraction. These abundances were used to calculate peptide-specific distributions which were then matched against a centrosome-specific profile, namely γ-tubulin. This approach has been applied by the same authors [15, 52] and others [53] for other sub-cellular proteomic investigations. Particular noteworthy is the study by Foster *et al.* [52], who mapped 1404 proteins to as many as ten subcellular compartments (including nucleus and cytosol) from 32 sucrose gradient fractions. Several co-localising proteins were subsequently visually confirmed by immunofluorescence and genome-wide functional genomics analyses revealed co-expression of genes and regulatory elements that reside in identical organelles.

In PCP, the normalised profiles along gradient fractions (Fig. 5, correlation profile plots) are compared with the reference profiles (or set of profiles) by computing χ^2-values as $\chi^2 = \Sigma i \ (x_i – x_p)^2/x_p$ where x_i is the normalised value of the peptide in fraction i and x_p is the value of the marker [53]. The protein χ^2 is then computed as the median of the peptide χ^2-values. Peptides and proteins with similar profiles to the markers will have small χ^2-values. Andersen *et al.* [50] extracted organelle-specific candidates by setting a 0.05 threshold on the χ^2-values based on known contaminant data.

Note that all gradient fractions are processed independently in PCP, substantially increasing technical variability and thus increasing the requirements for highly reproducible operating procedures, efficient normalisation and sufficient replication.

2.2.5.2 Localisation of organelle proteins by isotope tagging

Localisation of organelle proteins by isotope tagging (LOPIT) is another gradient-based approach that uses iTRAQ labelling quantification [54]. Fractions from self-generating iodixanol density gradients are collected. A set of enriched fractions are subsequently digested and labelled separately with iTRAQ reagents, pooled and the relative abundance of the peptides in the different fractions is measured by MS/MS. Proteins or peptides are visually assigned to specific organelles by clustering proteins of unknown location with known organelle markers in the space spanned by the main axes of PCA [55]. This can be performed on quantitation data or using all possible reporter ions ratios. In addition, novel members of clusters are rigorously allocated a given organelle by partial least-square discriminant analysis (PLS-DA) [56] using a well-annotated training set of proteins for each organelle (Fig. 5, PCA plot). A detailed protocol has been published [57]. LOPIT studies have been published on *Arabidopsis* [58–61], *Drosophila* [62] and DT40 lymphocyte cells [63].

Historically, χ^2 calculation has been associated with label-free quantification in PCP and iTRAQ data have been analysed by PCA and PLS-DA in LOPIT. However, these methodologies can be applied to any organelle quantification data, irrespective of the quantification methodology used.

One important aspect that should be underlined is that although the initial assignment of organelles is based on a limited number of markers, the accumulation of data, appropriately combined, increases the resolution and confidence of the clustering. Unless the first marker(s) were wrongly localised and subsequent clustering proteins where blindly assigned without biological interpretation of the results, the confidence of the localisation improves when several data sets are combined.

The main conceptual property that differentiates the tagging and quantitative proteomic approaches is the direct observation of individual cells under a microscope. Also, such approaches do not necessitate cell lysis followed by organelle fractionation, thus reducing a significant source of technical variability. As cells are not disrupted, it is possible to directly differentiate membrane and lumen proteins, although as already emphasised difficulties may arise in the case of transmembrane proteins.

3 Important aspects that have to be taken into account when executing an organelle proteomic experiment

Sensitivity and specificity are important factors that must be taken into account in organelle proteomics. These are to a large extent dictated by the methodology and the experimental designs that have been used, but also the statistical methodologies applied to filter and normalise the data play an important role.

3.1 Contaminants and specificity: reliability of the data

One of the main issues with organelle proteomics is contamination of organelle-specific fractions. Contamination can be systematic, for instance endoplasmic reticulum contaminants in nuclear membrane fractions and *vice versa*, or random. Technical replicates allow for the measurement of the latter. Unavoidable systematic contamination has led

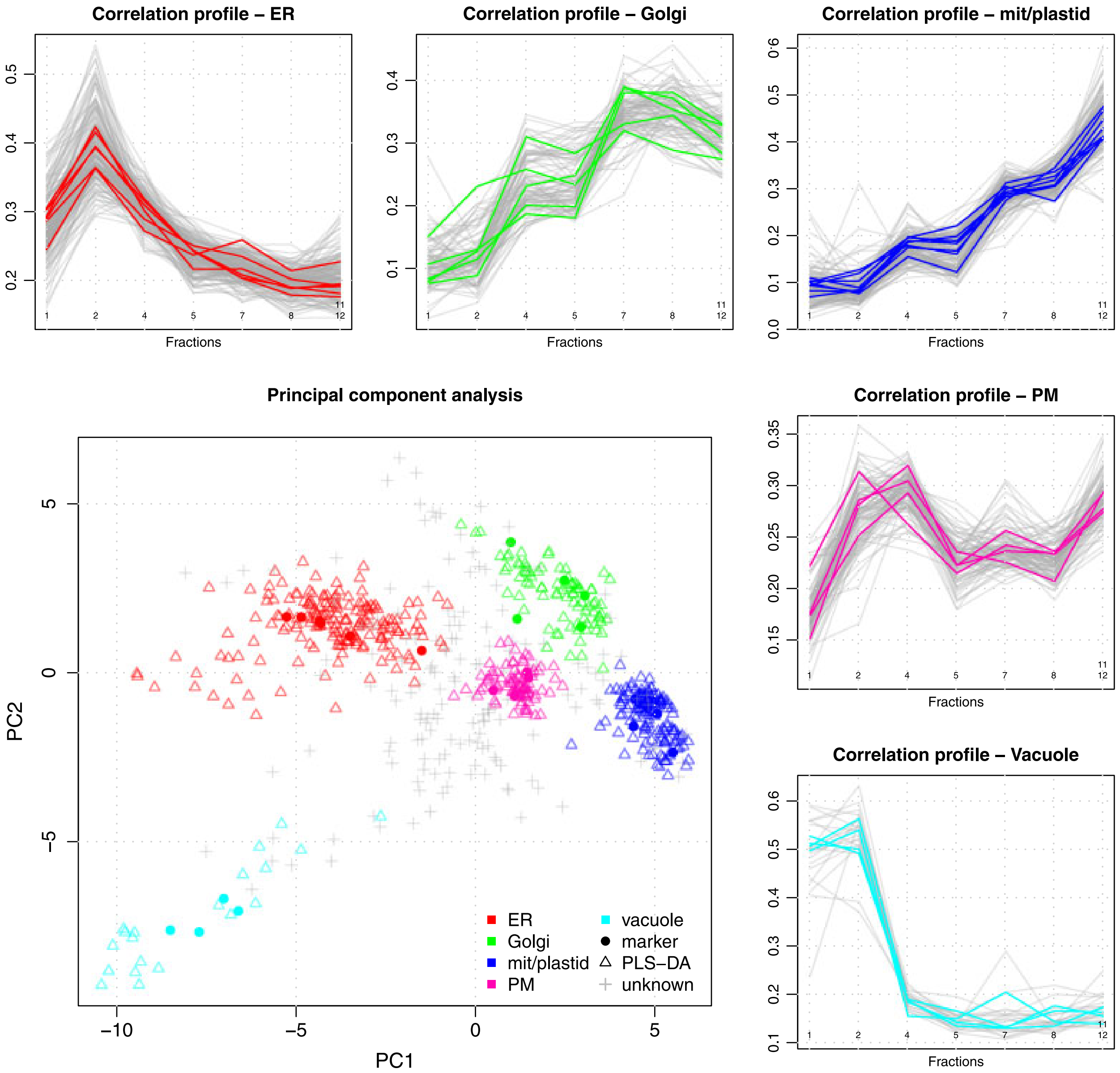

Figure 5. Analysis of data from Dunkley *et al.* (Table 3 and Supporting Information of [58]) using PCA and PLS-DA (bottom-left plot) and correlation profiles (CP) (top and right plots). The assignment of individual proteins is based on the marker proteins (circles on the PCA plot and coloured lines on CP plots) using PLS-DA (see [56, 58] for details). Note that proteins with unknown localisation are not illustrated on the CP plots. ER, endoplasmic reticulum; mit, mitochondria; PM, plasma membrane.

to more elaborated experimental designs that *de facto* deal with this by analysing several fractions. Au *et al.* [64] compared results from [48] (invariant rich fraction) and [52] (PCP) in terms of contaminants. These two studies reached only 28% agreement in the localisation of Golgi residents. This exemplifies that different experimental designs have their own respective strength and weaknesses.

Gradient approaches have a holistic approach of organelle proteomics that includes contamination in the experimental design; instead of avoiding contamination, it is accounted for. These approaches can also be used to study the dynamic aspects of organelle proteomes as multiple organelles are co-analysed, thus if a protein moves from one location to another upon a given perturbation, this movement will be captured in the analysis of multiple experiments.

3.2 Coverage and sensitivity: completeness of the complement

The definition of the proteome complement of an organelle is not merely a matter of protein identification; it is also highly dependent on quantification, assessment of contamination (for pure fraction or enrichment approaches), reliability of organelle marker (for PCP and LOPIT) and normalisation. Once the data have been quality controlled, efficient background correction algorithms are required to (i) objectively remove noise in the data and (ii) avoid the arbitrary discarding of important information based on the subjective thresholds. Then and only then it is possible to efficiently and reliably identify peptides and subsequently proteins.

Full coverage may be achieved only with sensitive instruments and multiple technologies. Several proteins such as lumen and peripheral proteins may be lost during organelle preparation. The sensitivity of the analytical methods (from protein extraction, peptide generation, separation and identification/quantification) will also define if less-abundant molecules are observed. Even in the absence of contamination, the dynamic nature of protein localisation and transient protein–protein interactions may not allow one to assess coverage in an absolute manner.

A significant amount of proteins are found in multiple locations. Andreyev *et al.* [47] claim that most proteins are located in multiple compartments. Foster *et al.* [52] estimate 40% of proteins are present in multiple locations. Hall *et al.* [63] report 70% of the identified proteins are not localised to a single organelle.

It can be argued that the multiple locations of proteins can be best uncovered by gradient approaches such as LOPIT and PCP that give information about several organelles simultaneously and provide a cell-wide localisation picture. It remains challenging, however, to identify multiply located proteins with those that either did not migrate with their corresponding organelle of origin during the fractionation or are part of a sub-cellular structure with no markers measured to aid annotation.

Despite the encouraging results of recent studies (see for instance [52] that analysed ten different organelles), not all organelles are characterised by any of the common approaches, certainly not by LOPIT and PCP. Some organelles are present in too low abundance to be identified by current technologies and some of the more abundant organelles do not resolve during sample preparation. In gradient-based approaches for instance, one density gradient will resolve only some organelles and different gradient conditions may be required to fully resolve other organelles [17].

4 Analysing organelle proteomics data

Biological interpretation of the data is of paramount importance in order to assess the strength of the results. Without this, suboptimal data sets may pollute the literature and databases, resulting in wasted efforts in validating such poor data.

4.1 Quality assessment and normalisation

A critical aspect of using technologies that generate vast amounts of data is that if technical biases are encountered but not taken into account, all the data generated will lead to misleading conclusions. As such, thorough quality control (QC) and assessment are essential prior to any decision is made based on the data. It is necessary to make sure that no systematic bias or correlation with any non-relevant analytical or technical variables confound the biological signal [65]. Several factors should be screened systematically, such as processing date (case 1 in [66]), sampling practice (case 2 in [66]), the experimenter and instrument used and instrument calibration (case 3 in [66]). It essential to avoid any confounding factors [19, 66–68], where overlapping sources of variation (technical or biological) would prevent the researcher to assign the origin of the observed variability. This can be achieved by blocking, *i.e.* randomly or deterministically assigning samples to groups (blocks), to reduce known or suspected but unwanted variability and thus increase the precision of subsequent analyses [19, 68]. Although the above has received much attention in the nucleic acid community, only a few similar studies have been published for proteomics, and most notably in biomarker discovery [69–72]. Correctness of the results can further be assured by appropriate controls, ideally chosen before starting the experiment and sometimes also *post hoc* by comparing them with previously validated data (Section 4.2). Incorporation of technical replicates may allow to explicitly assess the extent of non-relevant variability [65], which has to be less than the biological variability of interest. Standard protocols and the experience of the experimenter will greatly reduce technical variability. An organelle proteomics experiment involving four iTRAQ-labelled fractions from one gradient will be processed in one MS run and will share all technical variability. However, when multiple runs are required (as in label-free quantification or when several iTRAQ experiments are combined, see Fig. 6A), it is advised to properly assign the fractions to processing batches to avoid to overlay biological (fractions) and technical (operator, day) variability (Fig. 6B).

The main outcome of the QC should help answering the following: can the data be trusted? Are there any technical biases that have to be corrected? What objective threshold should be used before removing parts of the data? Note that a good QC procedure should be applicable to a single experiment to assess the absolute quality of individual experiments, as several comparable experiments are not always at hand. Continuous monitoring allows often to detect trends in quality metrics [72] and foresee possible issues.

Normalisation, *i.e.* identification, modelling of the data and removal of technical biases while leaving the biological signal as much as possible intact is always required before interpretation. The normalisation strategy used will depend on the exact nature of the data, but is essentially composed

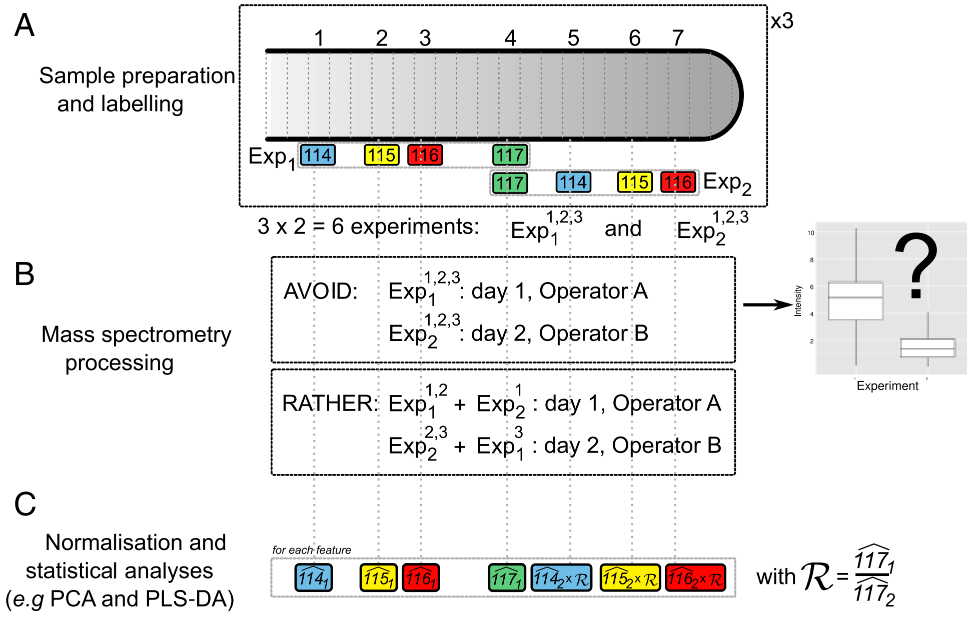
A
Sample preparation and labelling
1 2 3 4 5 6 7
x3
Exp1 114 115 116 117
117 114 115 116 Exp2
3 x 2 = 6 experiments: Exp1 1,2,3 and Exp2 1,2,3
B
Mass spectrometry processing
AVOID: Exp1 1,2,3 : day 1, Operator A
Exp2 1,2,3 : day 2, Operator B
RATHER: Exp1 1,2 + Exp2 1 : day 1, Operator A
Exp2 2,3 + Exp1 3 : day 2, Operator B
Intensity
10
8
6
4
2
?
Experiment
C
Normalisation and statistical analyses (e.g PCA and PLS-DA)
for each feature
114₁ 115₁ 116₁ 117₁ 114₂×R 115₂×R 116₂×R
with R = 117₁ / 117₂

Figure 6. (A) Example of an iTRAQ-based design involving several experiments. Seven fractions are sampled along a gradient and labelled with iTRAQ reagents. Fraction 4 is labelled once in each experiment for later normalisation. (B) A total of six experiments, fractions 1–4 and 4–7, each in triplicate are processed by two operators on different days. Operator A processes all the triplicates of experiment 1 on the same day and systematic differences in intensities are subsequently observed between the two experiments (see inlaid box plot). Where does this difference come from? Is it due to biology or a systematic technical bias? (C) Normalisation of the experiments: the mean quantity of each feature (peptide or protein) of the second experiment is linearly scaled towards the first one using the ratios of the common fraction prior to statistical analyses.

of a background correction step followed by a summarisation of the peptide data into a protein quantification. Normalisation should also make sure that independent runs are comparable. In case of the two iTRAQ experiments shown in Fig. 6, the ratios of the common fraction should be used to linearly scale one experiment towards the other (Fig. 6C).

4.2 Validation of the results

Validation of the data is typically carried out where the observed proteins are compared with a given reference. The chosen reference can be a prior study of particular interest and/or a database (see [73] for a recent review). Additional confidence can be gained when identical or similar results can be uncovered by independent and orthogonal techniques. When high-throughput MS approaches have been applied, a tagging technique such as creation of a GFP fusion protein is often used to validate the sub-cellular localisation of a set of proteins of interest [58]. Unfortunately, this generally allows only investigation of a small part of the data that has been generated. Furthermore, different techniques might suggest different locations and none can be considered a gold standard. In such cases, careful inspection of the raw data is required to eliminate false positives due to the technical biases and additional validation experiments might be necessary to eventually determine the proteins localisation(s).

Positive and negative controls, chosen prior to the experiment, can be used to partially validate the results. *Post hoc* comparison of individual results (*e.g.* proteins) and of overall patterns (*e.g.* list overlaps or gene set enrichment) with other data sets can also be applied. Bioinformatics approaches such as database mining, protein–protein interaction networks [74], ontology analysis or *in silico* sub-cellular localisation prediction [75–77] can also be used to support the experimental findings. Some computational tools are now able to assess dual locations and contaminants.

4.3 Meta analyses of proteomics data

The need for comparisons of different analytical methodologies and their implementations is highlighted by the diversity of analyses in organelle proteomics. We cited mixture-model fitting [43] as well as rank product and iterative group analysis [45, 46] that were applied in subtractive proteomic approaches [43, 44]. How do these two approaches compare with each other? Can they be applied to other similar studies? Are the implementations used for PCP by the original authors [50] and Wiese *et al.* [53] identical? How does the χ^2 approach compare with protein profile similarity [78] or PLS-DA in general? Is it possible for an independent researcher to reproduce the analysis of a data set, from raw data to results or apply a published analytical strategy to new data? The analysis pipelines used in organelle proteomics are often implemented using custom undistributed scripts, which makes reproduction of the results and comparison of the methodologies difficult – across laboratories and sometimes even within laboratory.

Meta analyses of proteomics data will hopefully be able to uncover additional layers of biological or technical complexity. To make these meta analyses possible, the first requisite is clearly that proteomic data should routinely be made publicly available for the scientific community. However, in contrast to the existing situation in other biological disciplines, comparatively few MS proteomic data sets are currently available in the public domain [79]. Despite this unfortunate situation, there are several publicly available databases for proteomics data (reviewed in [73]).

Such meta analyses of proteomics data have been already performed successfully, but not involving organelle data. In one of them, a mathematical approach called ''Latent Semantic Analysis'' was used to extract new knowledge from data coming from the HUPO Plasma Proteome Project [80]. Mueller *et al.* [81] demonstrated that it was possible to reuse data from the HUPO Brain Proteome Project in order to detect specific splice isoforms in the brain or the cerebrospinal fluid. More recently, new other types of studies integrating proteomic data with other types of biological information have also been performed. For instance, proteomic data have been reused to confirm the existence of genes in the human genome which had evolved *de novo* from the chimp [82]. The amount of this type of studies is expected to grow significantly and shows that sharing proteomic data is not only a good scientific practice, but also a potential new source of biological knowledge.

The correct and sufficient reporting of the experimental metadata is the other key point in making possible the reuse of the deposited data. In order to make easier the interchange of data between the community, several data formats have been developed in the last years by the HUPO Proteomics Standards Initiative (http://www.psidev.info), such as mzData and the new mzML (for MS data) [83], mzIdentML (for protein and peptide identifications coming from MS experiments) and PSI-MI XML (for protein interaction data) [84]. Each data standard includes a specific MIAPE (Minimal Information About a Proteomics Experiment) guidelines document [85], which states the desiderable minimal information that should be reported *per* type of experiment. In our opinion, models and ontologies for organelle proteomic experiments have to be defined and implemented to capture to majority of past, present and future studies. This in turn will allow to easily combine these state-of-the-art studies.

5 Conclusion and perspective

In biology, localisation is function and organelle proteomics is, as such, of paramount importance to understand cell biology. Major data sets have already been produced, but there are still limitations associated with the common approaches, one of them being the ability to capture the dynamic nature of protein localisation, another the need to capture the exquisite resolution of different sub-cellular compartments.

In this review, we have presented and discussed the most significant experimental designs. Among the important factors, we highlighted that minimising the technical variability between experiments (homogenate fractionation, subsequent purifications and quantification) plays a critical role as do QC, data normalisation and robust statistical analyses. Consolidation of all the aspects of a complex experiment is required to extract the most relevant information of their combination. To reach this ambitious goal, it is, however, essential to structure and describe all the decisions that are taken from sample preparation to data production and analysis. High-throughput data sets which are publicly available are only as useful as the associated metadata allows. The methods, software and parameters that are used should be clearly described along the analysis design. Such data sets are of high utility and are usually only partially mined by the researchers who generated them. If such data are to be further mined by the community structured data and metadata annotation must be available.

To facilitate sharing of organelle proteomic data, it is of vital importance that an interface is created which allows submission of data plus associated metadata to publicly accessible repositories. Only when the annotated data are made accessible, the conditions and designs under which they have been produced described, the pipeline it has undergone made explicit, will it be possible to make organelle research reproducible [86, 87]. This in turn will enable objective comparison of designs, combination of data sets and development of more complex and better controlled studies.

LG is funded by BBSRC grant BB/GO24618/1.

The authors have declared no conflict of interest.

6 References

[1] Simpson, J. C., Pepperkok, R., Localizing the proteome. *Genome Biol.* 2003, *4*, 240.

[2] Gibson, T. J., Cell regulation: determined to signal discrete cooperation. *Trends Biochem. Sci.* 2009, *34*, 471–482.

[3] Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S. *et al.*, Global analysis of protein localization in budding yeast. *Nature* 2003, *425*, 686–691.

[4] Hubbard, M. J., Functional proteomics: the goalposts are moving. *Proteomics* 2002, *2*, 1069–1078.

[5] Wallin, E., von Heijne, G., Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.* 1998, *7*, 1029–1038.

[6] Stevens, T. J., Arkin, I. T., Do more complex organisms have a greater proportion of membrane proteins in their genomes? *Proteins* 2000, *39*, 417–420.

[7] Kanapin, A., Batalov, S., Davis, M. J., Gough, J. *et al.*, Mouse proteome analysis. *Genome Res.* 2003, *13*, 1335–1344.

[8] Wu, C. C., Yates, J. R., 3rd. The application of mass spectrometry to membrane proteomics. *Nat. Biotechnol.* 2003, *21*, 262–267.

[9] Mootha, V. K., Bunkenborg, J., Olsen, J. V., Hjerrild, M. *et al.*, Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell* 2003, *115*, 629–640.

[10] Kislinger, T., Cox, B., Kannan, A., Chung, C. *et al.*, Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. *Cell* 2006, *125*, 173–186.

[11] Ponten, F., Gry, M., Fagerberg, L., Lundberg, E. *et al.*, A global view of protein expression in human cells, tissues, and organs. *Mol. Syst. Biol.* 2009, *5*, 337.

[12] Van den Oever, M. C., Goriounova, N. A., Li, K. W., Van der Schors, R. C. *et al.*, Prefrontal cortex AMPA receptor plasticity is crucial for cue-induced relapse to heroin-seeking. *Nat. Neurosci.* 2008, *11*, 1053–1058.

[13] Taylor, S. W., Fahy, E., Ghosh, S. S., Global organellar proteomics. *Trends Biotechnol.* 2003, *21*, 82–88.

[14] Yates, J. R., 3rd. , Gilchrist, A., Howell, K. E., Bergeron, J. J., Proteomics of organelles and large cellular structures. *Nat. Rev. Mol. Cell. Biol.* 2005, *6*, 702–714.

[15] Andersen, J. S., Mann, M., Organellar proteomics: turning inventories into insights. *EMBO Rep.* 2006, *7*, 874–879.

[16] Tan, S., Tan, H. T., Chung, M. C., Membrane proteins and membrane proteomics. *Proteomics* 2008, *8*, 3924–3932.

[17] Sadowski, P. G., Groen, A. J., Dupree, P., Lilley, K. S., Subcellular localization of membrane proteins. *Proteomics* 2008, *8*, 3991–4011.

[18] Wiederhold, E., Veenhoff, L. M., Poolman, B., Slotboom, D. J., Proteomics of *Saccharomyces cerevisiae* organelles. *Mol. Cell. Proteomics* 2010, *9*, 431–445.

[19] Oberg, A. L., Vitek, O., Statistical design of quantitative mass spectrometry-based proteomic experiments. *J. Proteome Res.* 2009, *8*, 2144–2156.

[20] Simpson, J. C., Pepperkok, R., The subcellular localization of the mammalian proteome comes a fraction closer. *Genome Biol.* 2006, *7*, 222.

[21] Matsuyama, A., Arai, R., Yashiroda, Y., Shirai, A. *et al.*, ORFeome cloning and global analysis of protein localization in the fission yeast *Schizosaccharomyces pombe*. *Nat. Biotechnol.* 2006, *24*, 841–847.

[22] Davis, T. N., Protein localization in proteomics. *Curr. Opin. Chem. Biol.* 2004, *8*, 49–53.

[23] Jarvik, J. W., Telmer, C. A., Epitope tagging. *Annu. Rev. Genet.* 1998, *32*, 601–618.

[24] Ross-Macdonald, P., Coelho, P. S., Roemer, T., Agarwal, S. *et al.*, Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* 1999, *402*, 413–418.

[25] Kumar, A., Agarwal, S., Heyman, J. A., Matson, S. *et al.*, Subcellular localization of the yeast proteome. *Genes Dev.* 2002, *16*, 707–719.

[26] Berglund, L., Andrade, J., Odeberg, J., Uhlen, M., The epitope space of the human proteome. *Protein Sci.* 2008, *17*, 606–613.

[27] Berglund, L., Bjorling, E., Oksvold, P., Fagerberg, L. *et al.*, A genecentric Human Protein Atlas for expression profiles based on antibodies. *Mol. Cell. Proteomics* 2008, *7*, 2019–2027.

[28] Barbe, L., Lundberg, E., Oksvold, P., Stenius, A. *et al.*, Toward a confocal subcellular atlas of the human proteome. *Mol. Cell. Proteomics* 2008, *7*, 499–508.

[29] Stadler, C., Skogs, M., Brismar, H., Uhlen, M., Lundberg, E., A single fixation protocol for proteome-wide immunofluorescence localization studies. *J. Proteomics* 2009, *73*, 1067–1078.

[30] Brunet, S., Thibault, P., Gagnon, E., Kearney, P. *et al.*, Organelle proteomics: looking at less to see more. *Trends Cell. Biol.* 2003, *13*, 629–638.

[31] Huber, L. A., Pfaller, K., Vietor, I., Organelle proteomics: implications for subcellular fractionation in proteomics. *Circ. Res.* 2003, *92*, 962–968.

[32] Stasyk, T., Huber, L. A., Zooming in: fractionation strategies in proteomics. *Proteomics* 2004, *4*, 3704–3716.

[33] Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., Kuster, B., Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* 2007, *389*, 1017–1031.

[34] Blondeau, F., Ritter, B., Allaire, P. D., Wasiak, S. *et al.*, Tandem MS analysis of brain clathrin-coated vesicles reveals their critical involvement in synaptic vesicle recycling. *Proc. Natl. Acad. Sci. USA* 2004, *101*, 3833–3838.

[35] Andersen, J. S., Lyon, C. E., Fox, A. H., Leung, A. K. *et al.*, Directed proteomic analysis of the human nucleolus. *Curr. Biol.* 2002, *12*, 1–11.

[36] Cronshaw, J. M., Krutchinsky, A. N., Zhang, W., Chait, B. T., Matunis, M. J., Proteomic analysis of the mammalian nuclear pore complex. *J. Cell Biol.* 2002, *158*, 915–927.

[37] Taylor, S. W., Fahy, E., Zhang, B., Glenn, G. M. *et al.*, Characterization of the human heart mitochondrial proteome. *Nat. Biotechnol.* 2003, *21*, 281–286.

[38] Rezaul, K., Wu, L., Mayya, V., Hwang, S. I., Han, D., A systematic characterization of mitochondrial proteome from human T leukemia cells. *Mol. Cell. Proteomics* 2005, *4*, 169–181.

[39] Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B. *et al.*, Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* 2002, *1*, 376–386.

[40] Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F. *et al.*, Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* 1999, *17*, 994–999.

[41] Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B. *et al.*, Multiplexed protein quantitation in *Saccharomyces*

cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* 2004, *3*, 1154–1169.

[42] Schirmer, E. C., Florens, L., Guan, T., Yates, J. R., 3rd., Gerace, L., Nuclear membrane proteins with potential disease links found by subtractive proteomics. *Science* 2003, *301*, 1380–1382.

[43] Marelli, M., Smith, J. J., Jung, S., Yi, E. *et al.*, Quantitative mass spectrometry reveals a role for the GTPase Rho1p in actin organization on the peroxisome membrane. *J. Cell Biol.* 2004, *167*, 1099–1112.

[44] Wiederhold, E., Gandhi, T., Permentier, H. P., Breitling, R. *et al.*, The yeast vacuolar membrane proteome. *Mol. Cell. Proteomics* 2009, *8*, 380–392.

[45] Breitling, R., Armengaud, P., Amtmann, A., Herzyk, P., Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.* 2004, *573*, 83–92.

[46] Breitling, R., Amtmann, A., Herzyk, P., Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *Biomed. Chromatogr. Bioinformatics* 2004, *5*, 34.

[47] Andreyev, A. Y., Shen, Z., Guan, Z., Ryan, A. *et al.*, Application of proteomic marker ensembles to subcellular organelle identification. *Mol. Cell. Proteomics* 2010, *9*, 388–402.

[48] Gilchrist, A., Au, C. E., Hiding, J., Bell, A. W. *et al.*, Quantitative proteomics analysis of the secretory pathway. *Cell* 2006, *127*, 1265–1281.

[49] De Duve, C., Tissue fractionation. Past and present. *J. Cell Biol.* 1971, *50*, 20d–55d.

[50] Andersen, J. S., Wilkinson, C. J., Mayor, T., Mortensen, P. *et al.*, Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* 2003, *426*, 570–574.

[51] Zhu, W., Smith, J. W., Huang, C. M., Mass spectrometry-based label-free quantitative proteomics. *J. Biomed. Biotechnol.* 2010, *2010*, 840518.

[52] Foster, L. J., de Hoog, C. L., Zhang, Y., Xie, X. *et al.*, A mammalian organelle map by protein correlation profiling. *Cell* 2006, *125*, 187–199.

[53] Wiese, S., Gronemeyer, T., Ofman, R., Kunze, M. *et al.*, Proteomics characterization of mouse kidney peroxisomes by tandem mass spectrometry and protein correlation profiling. *Mol. Cell. Proteomics* 2007, *6*, 2045–2057.

[54] Dunkley, T. P., Watson, R., Griffin, J. L., Dupree, P., Lilley, K. S., Localization of organelle proteins by isotope tagging (LOPIT). *Mol. Cell. Proteomics* 2004, *3*, 1128–1134.

[55] Pearson, K., *Philosophical Magazine* 1901, pp. 559–572.

[56] Karp, N. A., Griffin, J. L., Lilley, K. S., Application of partial least squares discriminant analysis to two-dimensional difference gel studies in expression proteomics. *Proteomics* 2005, *5*, 81–90.

[57] Sadowski, P. G., Dunkley, T. P., Shadforth, I. P., Dupree, P. *et al.*, Quantitative proteomic approach to study subcellular localization of membrane proteins. *Nat. Protoc.* 2006, *1*, 1778–1789.

[58] Dunkley, T. P., Hester, S., Shadforth, I. P., Runions, J. *et al.*, Mapping the Arabidopsis organelle proteome. *Proc. Natl. Acad. Sci. USA* 2006, *103*, 6518–6523.

[59] Lilley, K. S., Dunkley, T. P., Determination of genuine residents of plant endomembrane organelles using isotope tagging and multivariate statistics. *Methods Mol. Biol.* 2008, *432*, 373–387.

[60] Lilley, K. S., Dupree, P., Methods of quantitative proteomics and their application to plant organelle characterization. *J. Exp. Bot.* 2006, *57*, 1493–1499.

[61] Lilley, K. S., Dupree, P., Plant organelle proteomics. *Curr. Opin. Plant Biol.* 2007, *10*, 594–599.

[62] Tan, D. J., Dvinge, H., Christoforou, A., Bertone, P. *et al.*, Mapping organelle proteins and protein complexes in *Drosophila melanogaster*. *J. Proteome Res.* 2009, *8*, 2667–2678.

[63] Hall, S. L., Hester, S., Griffin, J. L., Lilley, K. S., Jackson, A. P., The organelle proteome of the DT40 lymphocyte cell line. *Mol. Cell. Proteomics* 2009, *8*, 1295–1305.

[64] Au, C. E., Bell, A. W., Gilchrist, A., Hiding, J. *et al.*, Organellar proteomics to create the cell map. *Curr. Opin. Cell. Biol.* 2007, *19*, 376–385.

[65] Old, W. M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K. G. *et al.*, Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell. Proteomics* 2005, *4*, 1487–1502.

[66] Hu, J., Coombes, K. R., Morris, J. S., Baggerly, K. A., The importance of experimental design in proteomic mass spectrometry experiments: some cautionary tales. *Brief Funct. Genomic. Proteomic.* 2005, *3*, 322–331.

[67] White, C. N., Chan, D. W., Zhang, Z., Bioinformatics strategies for proteomic profiling. *Clin. Biochem.* 2004, *37*, 636–641.

[68] Chich, J. F., David, O., Villers, F., Schaeffer, B. *et al.*, Statistics for proteomics: experimental design and 2-DE differential analysis. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 2007, *849*, 261–272.

[69] Coombes, K. R., Fritsche, H. A., Jr., Clarke, C., Chen, J. N. *et al.*, Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. *Clin. Chem.* 2003, *49*, 1615–1623.

[70] White, C. N., Zhang, Z., Chan, D. W., Quality control for SELDI analysis. *Clin. Chem. Lab. Med.* 2005, *43*, 125–126.

[71] Hong, H., Dragan, Y., Epstein, J., Teitel, C. *et al.*, Quality control and quality assessment of data from surface-enhanced laser desorption/ionization (SELDI) time-of flight (TOF) mass spectrometry (MS). *Biomed. Chromatogr. Bioinformatics* 2005, *6*, S5.

[72] Cairns, D. A., Perkins, D. N., Stanley, A. J., Thompson, D. *et al.*, Integrated multi-level quality control for proteomic profiling studies using mass spectrometry. *Biomed. Chromatogr. Bioinformatics* 2008, *9*, 519.

[73] Mead, J. A., Bianco, L., Bessant, C., Recent developments in public proteomic MS repositories and pipelines. *Proteomics* 2009, *9*, 861–881.

[74] Shin, C. J., Wong, S., Davis, M. J., Ragan, M. A., Protein-protein interaction as a predictor of subcellular location. *Biomed. Chromatogr. Syst. Biol.* 2009, *3*, 28.

[75] Donnes, P., Hoglund, A., Predicting protein subcellular localization: past, present, and future. *Genomics Proteomics Bioinformatics* 2004, *2*, 209–215.

[76] Rey, S., Gardy, J. L., Brinkman, F. S., Assessing the precision of high-throughput computational and laboratory approaches for the genome-wide identification of protein subcellular localization in bacteria. *Biomed. Chromatogr. Genomics* 2005, *6*, 162.

[77] Sprenger, J., Fink, J. L., Teasdale, R. D., Evaluation and comparison of mammalian subcellular localization prediction methods. *Biomed. Chromatogr. Bioinformatics* 2006, *7*, S3.

[78] Kirchner, M., Renard, B. Y., Kothe, U., Pappin, D. J. *et al.*, Computational protein profile similarity screening for quantitative mass spectrometry experiments. *Bioinformatics* 2010, *26*, 77–83.

[79] Editors, Credit where credit is overdue. *Nat. Biotechnol.* 2009, *27*, 579.

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

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

[82] Knowles, D. G., McLysaght, A., Recent de novo origin of human protein-coding genes. *Genome Res.* 2009, *19*, 1752–1759.

[83] Deutsch, E., mzML: a single, unifying data format for mass spectrometer output. *Proteomics* 2008, *8*, 2776–2777.

[84] Kerrien, S., Orchard, S., Montecchi-Palazzi, L., Aranda, B. *et al.*, Broadening the horizon – level 2.5 of the HUPO-PSI format for molecular interactions. *Biomed. Chromatogr. Biol.* 2007, *5*, 44.

[85] Taylor, C. F., Paton, N. W., Lilley, K. S., Binz, P. A. *et al.*, The minimum information about a proteomics experiment (MIAPE). *Nat. Biotechnol.* 2007, *25*, 887–893.

[86] Gentleman, R., Reproducible research: a bioinformatics case study. *Stat. Appl. Genet. Mol. Biol.* 2005, *4*, Article2.

[87] Gentleman, R., Lang, D. T., Bioconductor Project Working Papers. Working Paper 2 2004.