EuPA Open Proteomics 8 (2015) 68–77

Contents lists available at ScienceDirect

EuPA Open Proteomics

journal homepage: www.elsevier.com/locate/euprot

A Breast Cell Atlas: Organelle analysis of the MDA-MB-231 cell line by density-gradient fractionation using isotopic marking and label-free analysis

Marianne Sandin[1], Linn Antberg[1], Fredrik Levander, Peter James*

Department of Immunotechnology, CREATE Health, House 406, Medicon Village, 223 81 Lund, Sweden,

ARTICLE INFO

Article history:
Received 6 February 2015
Received in revised form 30 May 2015
Accepted 10 June 2015
Available online 26 June 2015

Keywords:
LOPIT (location of organelle proteins by isotopic tagging)
Label-free quantification
Isotopic-labelling
Organelle
Subcellular location

ABSTRACT

Protein translocation between organelles in the cell is an important process that regulates many cellular functions. However, organelles can rarely be isolated to purity so several methods have been developed to analyse the fractions obtained by density gradient centrifugation. We present an analysis of the distribution of proteins amongst organelles in the human breast cell line, MDA-MB-231 using two approaches: an isotopic labelling and a label-free approach.

1. Introduction

After the completion of the human genome sequencing in 2001, the efforts in molecular and cellular biology research have turned to the delineation of the proteome. Such delineation involves not only assessing protein abundance but also consideration of the spatial organisation of proteins, as knowledge of subcellular location brings insight into the function and interaction partners of proteins. Moreover, external stimuli or a certain disease can have implications for the organelle location of proteins and may lead to translocations of proteins between organelles. In a recent study, the trafficking of proteins between the nucleus and the cytoplasm was found to have a crucial role in cell cycle regulation [1].

Traditional methods to study organelle location are either focused on individual proteins visualised by microscopy, or use mass spectrometry to identify proteins in an isolated organelle. However, different types of problems are associated with these methods. The standard methods involving microscopy are typically low-throughput while the more high-throughput techniques, using mass spectrometry characterisation of isolated organelles, require an organelle preparation free of contamination. This is both time-consuming and impossible to achieve for some organelles, such as the endoplasmic reticulum that forms a continuum with the nuclei and Golgi vesicles. Even in a highly purified such organelle fraction, it is difficult to discriminate contaminating proteins from the true residents of the organelle [2].

Recently, partial separation of organelles by density-gradient centrifugation followed by proteomic quantification by mass spectrometry has evolved as a valuable tool for organelle proteomics. Simultaneous analysis of fractions from the whole cell makes it both easier to distinguish contaminants from true residents and gives the possibility to use the approach for tracking movement of proteins between organelles, for instance as the effect of different stimuli. In this method, each protein is quantified across the gradient fractions, creating a distribution profile. Proteins residing in the same organelle are assumed to co-fractionate, thereby obtaining the same profile. The profiles can therefore be compared to the ones of known organelle markers. An organelle marker is defined as a single or several proteins that are known to reside in a certain organelle, which creates a blueprint for unknown protein profiles acquired from the analytical set-up, and comparison to a marker therefore lets proteins with a similar profile to be assigned to their respective organelles. Consequently,

Abbreviations: LOPIT, location of protein by isotopic tagging; PCP, protein correlation profiling; PCA, principle component analysis; SVM, support vector machine; PLS-DA, partial least squares discriminant analysis; TMT, tandem mass tag; DOC, deoxycholate; TCA, trichloroacetic acid; ICAT, isotope coded affinity tagging; GFP, green fluorescent protein.

* Corresponding author.
E-mail address: peter.james@immun.lth.se (P. James).

[1] These authors have contributed equally to the work and are shared first authors.

http://dx.doi.org/10.1016/j.euprot.2015.06.002
1876-3820/

high-throughput organelle characterisation is possible without the need for highly purified organelle preparations. Density organelle centrifugation is commonly performed using sucrose gradients, but self-generating Iodixanol gradients have emerged as an attractive alternative since they are very reproducible and simple to handle [3].

Different proteomics techniques exist for the quantification of proteins; isotopic labelling and label-free being two frequently used. Several groups have successfully used stable isotope labelling with amino acids in cell cultures (SILAC) for subcellular proteomic analysis [4,5]. The capability of the SILAC-technique for multiplexing is limited due to few available heavy labelled amino acids, if not combined with other labelling techniques [6], wherefore other options must be considered for density-gradient purposes [7]. Dunkley et al. [3] have developed the LOPIT (localization of organelle proteins by isotope tagging) technique and it has been applied on samples from plant, fruit fly and chicken [8–10] and Foster et al. [4] used label-free quantification in combination with the PCP (protein correlation profiling) technique to identify the organelle proteome of mouse.

PCP infers protein location by a χ^2 score, defined as the normalized squared deviation of a protein profile from known markers [11]. The advantage of PCP is a relative ease of computation, as well as requiring very few organelle markers. The χ^2 score is however computed fraction-wise, resulting in increased technical variation and reproducibility requirements [12]. It should be noted that the score is peptide based, resulting in a χ^2 score distribution for every protein. Andersen et al. [11] used the median of the peptide χ^2 values to represent the protein score, where an upper limit of 0.05 was set for organelle allocation [4]. A measure of score dispersion, such as the MAD (median absolute deviation) could however be interesting to take into account, giving a more complete picture of the protein profile. The distribution information could also aid in selecting a representative set of peptides used for PCP, as the protein will be characterized by only one peptide when using the median value, which can differ depending on the identifications in the experiment.

LOPIT employs clustering assessment by PCA followed by supervised classification using e.g. PLS-DA (partial least squares discriminant analysis) [8–10] or support vector machines (SVM) [13]. The technique simultaneously makes use of the entire protein profile, but necessitates a larger marker set. This, however, also puts less emphasis on individual markers. Furthermore, the predictive power resulting from the multivariate supervised classification facilitates generalization of the analysis. Interestingly, label-free data has traditionally been analysed by PCP, while LOPIT has been used for labelled data [12], although the former MS quantification strategy in general results in larger numbers of identified proteins [14–16].

In this study, we have performed LOPIT analysis on both a peptide labelling and label-free analysis of density-gradient fractions from a human breast cancer cell line. Principal component analysis (PCA) was used for visualization and clustering assessment and organelle affiliation prediction was done by training a support vector machine (SVM) on the marker profiles An SVM classifier with a non-linear kernel was selected over PLS-DA as the latter is based on an assumption of linearity. SVM has furthermore been shown to have higher predictive accuracy as well as robustness to noise in a comparison between the techniques [17]. Equivalent clustering and organelle affiliation results, with roughly a quarter of the identified proteins assigned to organelles, were obtained using either quantification method, with the advantage of label-free quantification resulting in larger numbers of identifications. Furthermore, consistent results were found between using manually selected organelle markers and GO annotations followed by a variance-based outlier removal strategy. This, in combination with the rapidly increasing viability of label-free quantification for biomarker discovery [18], shows that organelle separation techniques will in the near future be an important factor for biological inference in high-throughout, third-generation proteomics experiments.

2. Materials and methods

2.1. Materials

Acrylamide, urea, Tris, magnesium acetate, DTT and iodoacetamide were from Sigma–Aldrich (Stockholm, Sweden). 12.5% Criterion Precast SDS-PAGE gels were from BioRad (Hercules, CA, USA). Sequencing-grade modified trypsin was purchased from Pierce (SDS diagnostics, Falkenberg, Sweden). The Micro-Lowry Protein Assay Kit was from Sigma Diagnostics (Stockholm, Sweden). All HPLC solvents were from Fluka (Sigma–Aldrich, Stockholm, Sweden). UltraMicroSpin C_{18} and UltraMacroSpin C_{18} columns were from the NestGroup (Southborough, MA, USA).

2.2. Cell culture

The human breast cancer cell line MDA-MB-231 was cultured in RPMI-1640 supplemented with 10% foetal bovine serum (FBS) and 1 % L-glutamine (Invitrogen, Carlsbad, CA) at 37 °C in a humidified atmosphere at 5% CO_2.

2.3. Subcellular fractionation by density-gradient centrifugation

MDA-MB-231 cells were cultured as described above and subsequently washed twice with PBS and then resuspended in a hypotonic homogenisation buffer (10 mM HEPES pH 7.9, 10 mM KCl, 1.5 mM $MgCl_2$, 0.5 mM DTT, Complete Mini EDTA-free protease inhibitor). 5 milllion cells were used for each separation. Samples for each of the two conditions were run in biological duplicate. The cells were kept on ice for 10 min and then disrupted using 35 strokes with a tight pestle in a 7 mL Dounce homogeniser until 90% of the cells were broken as determined by microscopy. Subsequently, nuclei were pelleted and saved in the freezer. The post-nuclear supernatant was diluted in homogenisation buffer and centrifuged at 100,000 × g for 1 h at 4 °C to pellet the organelles. The organelles were resuspended in 0.5 mL PBS and layered on top of a preformed 5–45% continuous iodixanol gradient. The gradient was centrifuged at 200,000 × g in a Beckman SW41 rotor at 4 °C for 2 h. Subsequently, 10 fractions of 1.2 mL were collected from the top of the gradient and the refractive index of each fraction was measured using a refractometer (Atago, Tokyo, Japan). The fractions were washed twice in 20 mM Na_2CO_3 and once with water. The membranes were collected by centrifugation at 100,000 × g for 1 h. The membrane pellets were resuspended in 10 mM Tris–HCl, 2% SDS and the protein concentrations were determined using the Micro-Lowry Assay Kit (Sigma–Aldrich, Stockholm, Sweden).

2.4. Separation of chromatin-associated and nucleoplasm fractions

Nuclear pellets were thawed on ice and washed once with ice-cold PBS and once with extraction buffer (15 mM Tris–HCl pH 7.4, 1 mM EDTA, 400 mM $MgCl_2$, 10% glycerol, 10 mM β-mercaptoethanol) before resuspending in 500 μL extraction buffer and incubation 1 h on ice. Samples were then centrifuged for 30 min at 16,000 × g and the supernatant was diluted in hypotonic buffer (10 mM Tris–HCl pH 7.4, 10 mM KCl, 1.5 mM $MgCl_2$, 10 mM β-mercaptoethanol) and saved as the nucleoplasm fraction. Subsequently, pellets were washed and resuspended in

Table 1
Minimum SVM scores necessary for a protein to be allocated to an organelle. N/A indicates not applicable.

	Label-free	TMT
Cytosol/cytoplasm	0.87666	0.89528
Endoplasmic reticulum	0.93127	0.61553
Golgi apparatus	0.35416	N/A
Mitochondrion	0.5331	0.43353
Plasma membrane/cell membrane	0.87022	N/A
Ribonucleoprotein complexes	0.28922	0.54973

micrococcal nuclease buffer (20 mM Tris–HCl, 10 mM KCl, 2 mM $MgCl_2$, 1 mM $CaCl_2$, 300 mM sucrose, protease inhibitor cocktail). Three units of S7 nuclease were added and incubated with the samples for 30 min on ice with mixing of the tubes every 4 min. The nuclease was inactivated by the addition of 10 μL 0.5 M EDTA and the samples were centrifuged for 3 min at 1500 g. The supernatant was diluted in hypotonic buffer and saved as chromatin-associated fraction. The two nuclear fractions were precipitated by adding DOC to a concentration of 0.08% (w/v) and incubating 1 hour on ice, followed by an addition of TCA to a concentration of 2.5% (w/v) and 30 min incubation on ice. The precipitated proteins were collected by centrifugation at 4000 g for 30 min and the pellet was washed with acetone. Proteins were then resuspended in 10 mM HEPES and the protein concentration was determined using the Micro-Lowry Assay Kit.

2.5. Label-free experimental design

Fractions 8, 9 and 10 were pooled in order to obtain the required protein amount and 25 μg of protein from each fraction was mixed with sample buffer (0.05 M Tris–HCl, 0.05 M SDS, 5% v/v glycerol, 0.1% DTT) and heated at 98 °C for 5 min. The samples were separated on a 12. % Criterion Precast SDS-PAGE gel (BioRad, Hercules, CA, USA) at 25 °C. The gel was stained using Gel Code Blue Stain Reagent (Pierce). Each of the lanes was cut into 10 slices that were destained in 50% acetonitrile and 25 mM NH_4HCO_3. The gel slices were reduced with 10 mM TCEP in 50 mM NH_4HCO_3 at 56 °C for 1 h. Alkylation was performed by adding 55 mM iodoacetamide acid in 50 mM NH_4HCO_3 and incubating for 45 min at RT in the dark. The slices were then washed once with 100 mM NH_4HCO_3 and several times using ACN and ddH_2O. Trypsin (20 μL of 12.5 μg/mL trypsin in 100 mM TEAB) was added to the dehydrated gel slices. The samples were left for 30 min at 4 °C prior to incubation at 37 °C overnight (approximately 18 h). The peptides were extracted from the gel by adding 1% TFA in 75% ACN and incubating at RT for 30 min. This was repeated once and the liquid from the two extractions were pooled together. Subsequently, the volume was reduced using a Speed vac and the samples were then dissolved in 10 μL 0.1% formic acid.

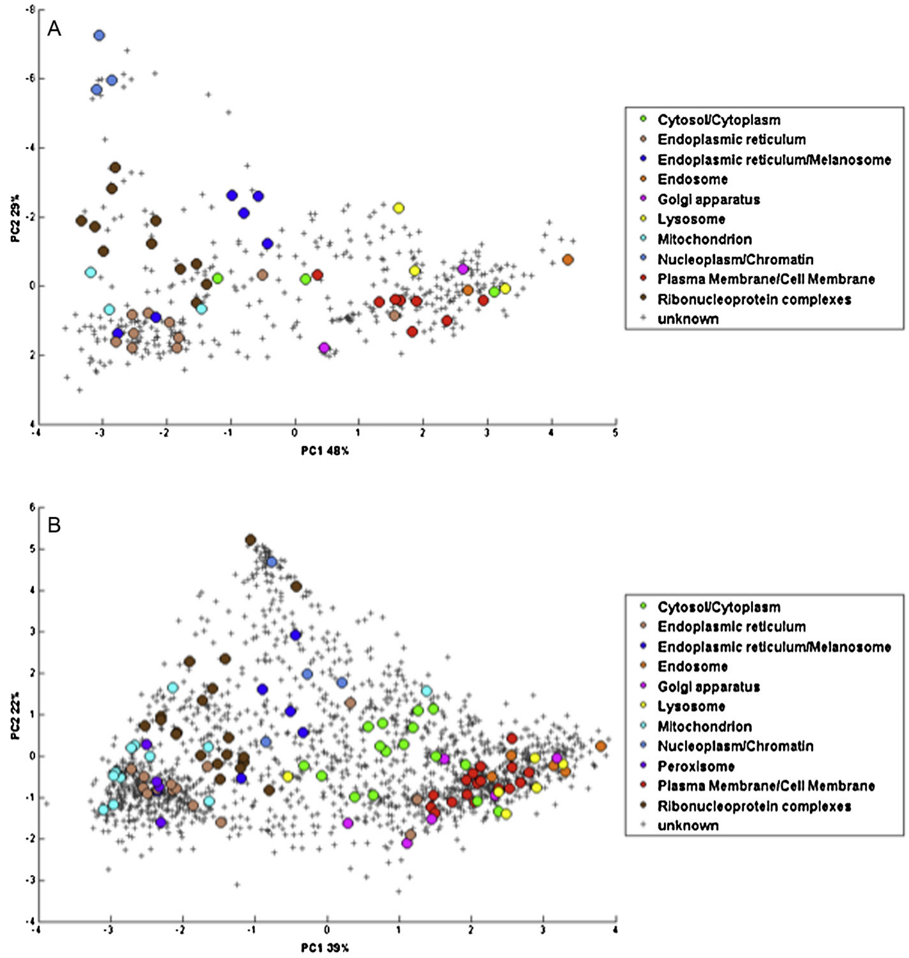

Fig. 1. PCA analysis of protein distribution using only manual markers. No marker outlier removal has been performed. Panel A shows the 439 proteins from the TMT dataset with 51 manual markers identified and Panel B the 1696 proteins from the label-free dataset with 111 markers identified.

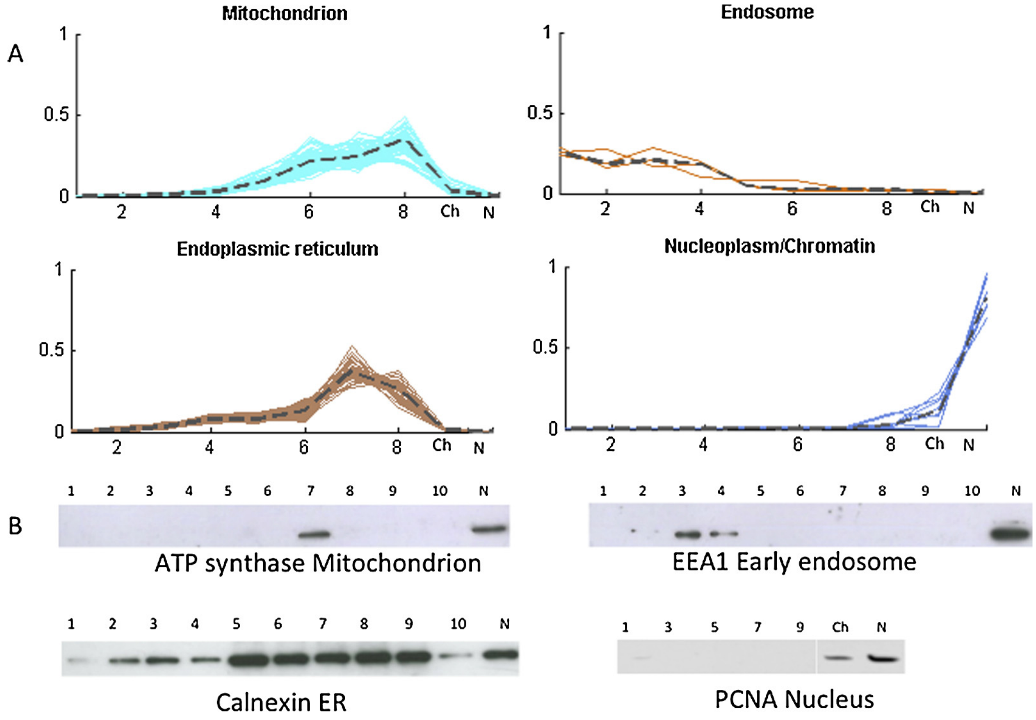

Fig. 2. Panel A shows the distribution profiles of organelle markers by label-free quantification, with normalized protein abundance (intensity) on the *Y*-axis and fraction number on the *X*-axis. The chromosomal associated protein fraction is labelled Ch and the nucleoplasm is labelled N. There was not enough material in fraction 8,9, and 10 so these were combined and run as one. Panel B shows the Western blotting of the fractions from density gradient centrifugation. The markers are: ATP synthase (mitochondria), EEA1 (early endosome), calnexin (endoplasmic reticulum) and PCNA (nucleus).

2.6. TMT experimental design

25 μg of protein from each fraction was prepared as described for samples for label-free analysis, except that the gel was run for shorter time and the whole sample was kept in one slice instead of 10. The peptides extracted from the gel were resuspended in 150 μL 200 mM TEAB and a reference pool was created by mixing 25 μL of each fraction. A TMT labelling kit (Thermo Scientific) was used to label the samples with isotopic tags and the labelling was performed according to the manufacturers protocol. The fractions were labelled as follows: fraction 1 and 6, 126; fraction 2 and 7, 127; fraction 3 and 8/9/10, 128; fraction 4 and chromatin fraction, 129; fraction 5 and nucleoplasm fraction, 130; reference pool, 131. The labelled peptides were combined into two samples: Sample A (Fraction 1,2,3,4,5,reference pool) and Sample B (fraction 6,7,8/9/10, chromatin, nucleoplasm, reference pool). Sample A and B were subsequently cleaned and fractionated on an SCX column (ICAT Strong Cation Exchange Cartridges (Applied Biosystems, Foster City, CA, USA)) at a flow rate of 100 μL/min. The samples were loaded on the pre-equilibrated cartridge and then eluted in 500 μL fractions by injecting KCl at increasing concentrations (0, 10, 25, 40, 60, 80, 100, 150, 250, 400 and 600 mM) in 5 mM KH_2PO_4, 25% ACN. The volume of the fractions was then reduced to less than 10 μL using a speed vac. The fractions were desalted using Ultra-MicroSpin C18 columns (the NestGroup, Southborough, MA, USA) before eluting the peptides in 100 μL 0.1% formic acid, 60% ACN. Subsequently they were dried in a speed vac until less than 5 μL remained, resuspended in 0.1% FA and analysed by RP-HPLC-MS/MS.

2.7. Western blotting

10 μg from each fraction was run on NuPAGE 10% Bis–Tris gels (Invitrogen, Carlsbad, CA, USA) under reducing conditions for ∼45 min at 130 V. Separated proteins were blotted on PVDF iBlot Transfer stacks in the iBlot gel transfer device (both Invitrogen). The PVDF membranes were blocked with 5% milk-PBS buffer for 2 h and probed with the following antibodies: PCNA (Chemicon, 1:1000), EEA1 (AbCAM, 1:2000), ATP Synthase (AbCAM, 1:4000), Calnexin (AbCAM, 1:2000). HRP-conjugated swine anti-rabbit antibody (DAKO) was used as secondary antibody and the blots were developed with SuperSignal West Femto Max Sensitivity Substrate (Pierce Biotechnology Inc., Rockford, IL), according to the protocol of the manufacturer. Detection was made with a CCD-camera (Odyssey FC Imager from LI-COR Biosciences UK Ltd (Cambridge, England)) and analysed in the Quantity One software (Hercules, CA, USA).

2.8. LC–MS/MS analysis

The fractions were run (TMT in duplicate) on an Eksigent 2D NanoLC system (Eksigent Technologies, Dublin, CA, USA) coupled to an LTQ OrbitrapXL (Thermo Electron, Bremen, Germany). Peptides were loaded with a constant flow of 10 μL/min onto a pre-column (Acclaim PepMap 100, C18, 5 μm, 5 mm × 0.3 mm, Thermo Fischer Scientific, Hägersten, Sweden) and subsequently separated on a 10 μm fused silica emitter, 75 μm × 16 cm (PicoTip Emitter, New Objective, Inc., Woburn, MA) packed in-house with Reprosil-Pur C18-AQ resin (3 μm Dr. Maisch GmbH) at a flow rate of 400 nL/min. The peptides were eluted with a 55 min linear gradient of 5–40 % ACN (0.1% FA) followed by a 5 min linear gradient from 40 to 80% ACN (0.1% FA). The LTQ-Orbitrap was operated in a data-dependent mode, simultaneously acquiring MS spectra in the Orbitrap (from m/z 400 to 2000) and MS/MS spectra in the LTQ. The ion trap loading was set to 30,000 with an MS/MS threshold of 500 counts using a normalised collision energy set to 35%. Each Orbitrap-MS scan was acquired at 60,000 FWHM nominal resolution settings using the lock mass option (m/z 445.120025) for internal calibration. The dynamic exclusion list was restricted to 500 entries using a repeat count of two with a

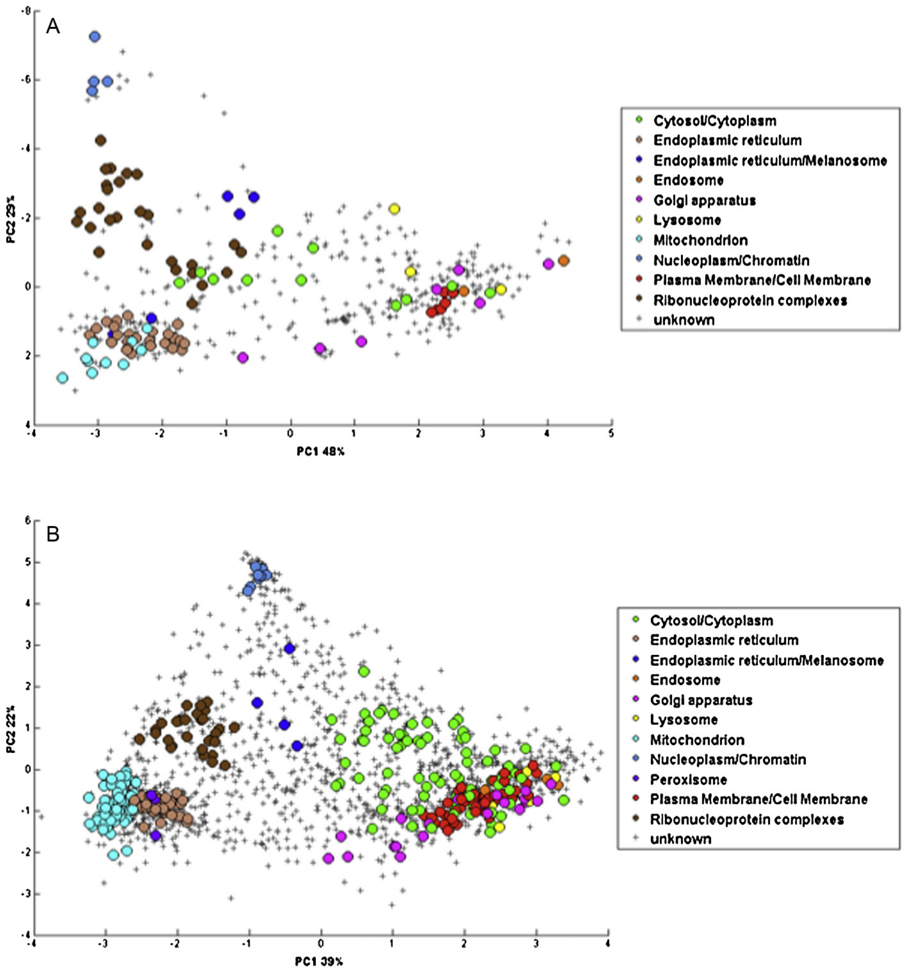

Fig. 3. PCA plots showing the markers acquired through combination of manual markers and GO annotations. Panel A TMT with 101 markers and Panel B label-free with 302 markers.

repeat duration of 20 s, and with a maximum retention period of 120 s. Precursor ion charge state screening was enabled to select for ions with at least two charges and rejecting ions with undetermined charge state. The seven most intense ions were selected for fragmentation with a dynamic exclusion window of 2 min. For TMT labelled samples the top three ions were subject to CID and HCD fragmentation.

2.9. Data processing

Initial data processing was done in the Proteios Software Environment [19] version 2.18.0. Raw spectra were converted to mzML [20] and Mascot Generic Format (MGF) using Proteowizard [21]. From the Proteios workspace, the mzML files and MGF files for the TMT and label-free data, respectively, were independently searched against the human part of SwissProt as of 2011-08-17, containing separate isoform entries and a reverse sequence database of equal size, totalling 71,324 entries. Searches were performed using X!Tandem (version TORNADO (2008.12.01.1)) with both native and K scoring [22], as well as Mascot version 2.3.01. The mass tolerance was set to 3 ppm for parent ions and 0.4 Da for fragment ions and one missed protease cleavage was allowed. Cys carbamidomethylation was set as fixed modification, with TMT 6plex (K- and N-terminal) added for the TMT data, and Met oxidation as variable modification. Searches were combined and the False Discovery Rate (FDR) computed using reverse sequences [23]. msInspect [24] feature detection with default settings was used for label-free quantification, followed by an in-house developed alignment algorithm for retention time correction and sequence propagation [25]. An FDR of < 0.01 was used on both peptide and protein level.

2.10. Location analysis

Quantitative protein reports were exported from Proteios and processed using MATLAB R2011a version 7.12.0.635 (www.mathworks.org). The TMT data was normalized to the reference pools and Sample A and B combined. Profiles of proteins identified in both technical replicates were averaged. For the label-free data, only profiles having no missing values were extracted. For both types of data, the profiles were normalized to the sum of all fractions.

Organelle location was inferred both from manually acquired markers as described and used in [26] (Supplementary information Table 1) and GO annotations. Here, a protein is referred to as a marker if it was identified either in the list of manually selected markers or a GO annotated marker which passed the outlier removal process described below. If a protein differed in subcellular location between the manual markers and GO annotations, the manual marker information was used. The marker profiles were subsequently subjected to outlier removal. A profile was considered as an outlier if it deviated more than 3 times the interquartile range from the median in at least one fraction or 1.5 times the interquartile range in 2 or more fractions. This process was iterated until no more outliers

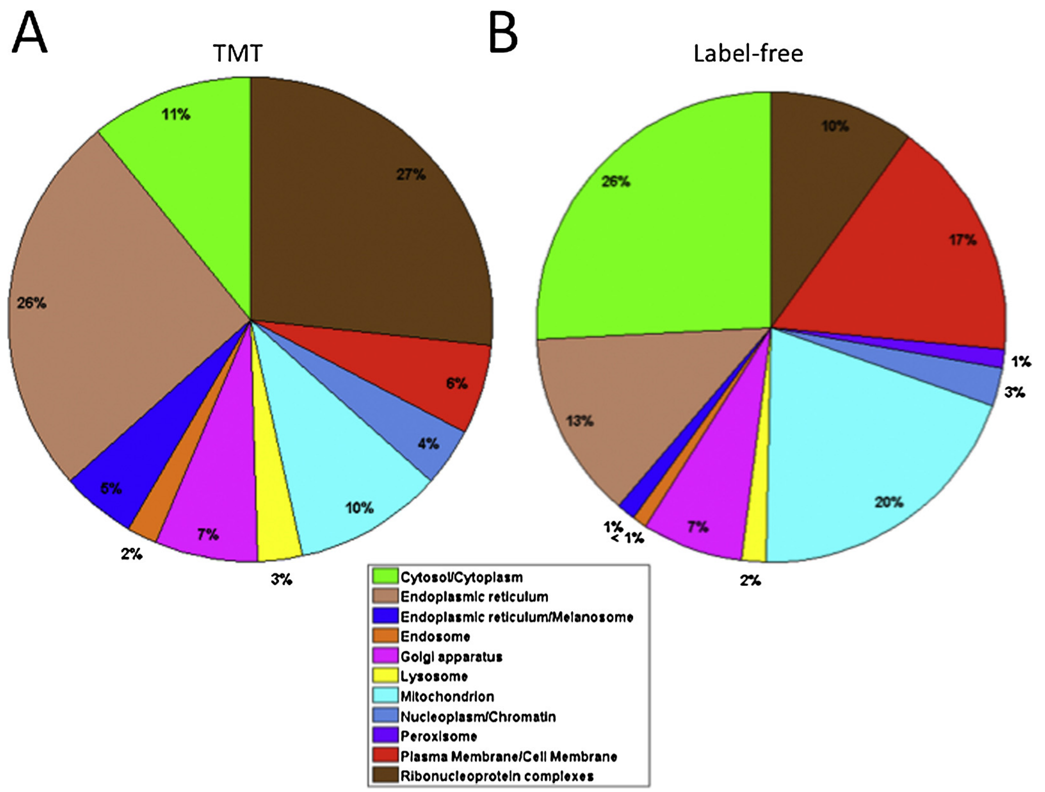

Fig. 4. Pie chart showing the distribution of proteins amongst the organelles.

were found. Only organelles represented by at least 10 markers were considered for further classification. In addition, 10% of the markers for each organelle were set aside for assessing prediction score limits after classification.

Supplementary material related to this article found, in the online version, at http://dx.doi.org/10.1016/j.plantsci.2004.08.011.

Protein profiles as well as markers were subsequently imported into pRoloc version 0.99.8 [27]. R package version 0.99.8 (2013) was used, where an SVM with a radial basis function (RBF) kernel was trained on the markers. Weights corresponding to the inverse of the class sizes were used. First, parameter settings were optimized by partitioning data, where 80% was used for stratified 5-fold cross-validation in combination with parameter grid search and the remaining 20% to evaluate the best parameter set using the F1 score. The F1 score is defined as the harmonic mean of precision = TP/(TP + FP) = 1-FDR and recall (sensitivity) = TP/(TP + FN), where TP – true positive, FP – false positive and FN – false negative. This process was repeated 100 times to acquire an F1 distribution from where the optimal parameters could be extracted. The final SVM was subsequently trained on the entire marker set, using the acquired settings. Organelle classification and corresponding SVM prediction scores were extracted to MATLAB, where an automated organelle score limit based on maximizing precision was computed. A true positive was defined as a marker not used for classification that had been predicted to belong to the correct organelle (from the 10% excluded markers) and a false positive an incorrectly classified marker. Proteins that had no previously assigned organelle location were disregarded in the calculations.

3. RESULTS

In this study, we describe the analysis of the subcellular locations of proteins from the human breast cancer cell line MDA-MB-231. Organelles were isolated by gentle disruption of the cell membrane, spin-down of the nucleus and ultracentrifugation of the post-nuclear supernatant to pellet the organelles. The organelles were layered on top of a continuous iodixanol gradient and separated according to their density by ultracentrifugation. The resultant distribution of proteins was analysed by two different MS quantification methods: Label-free and TMT labelling.

3.1. Initial cluster analysis of the label-free and isotopic labelling approaches

Fig. 1A and B show PCA plots for the two experiments, totalling 439 proteins for TMT and 1696 for the label-free dataset, respectively. Only manually acquired markers are annotated (51 identified for the TMT dataset and 111 for the label-free dataset). Although roughly equivalent clustering is seen, the label-free analysis showed a greater spread of the manual markers. Despite the clear separation seen in the TMT PCA plot, the low number of identified markers necessitated adding GO annotated markers. This was performed for both datasets, including marker outlier removal as described above, resulting in 101 and 302 markers for the TMT and label-free dataset, respectively. Fig. 2A shows the combined GO and manual protein marker profiles after outlier removal, while Fig. 2B shows the corresponding antibody markers used to monitor the extent of organelle separation. Partly overlapping distributions were seen for the ER and mitochondrion, while the endosome and nucleus had very distinct profiles (Figure 2B). This is reflected in the PCA plots in Fig. 3A and 3B, where the ER and mitochondrion markers are closely clustered (at least from what can be inferred from the first two principal components). The same can be seen for the Golgi and plasma membrane markers, while the nucleoplasm/chromatin cluster is clearly separated. Marker profiles showed considerable agreement between the two quantification techniques and profiles for all investigated organelles can be found in Supplementary

Fig. 1A and B. In Fig. 4, the marker organelle distributions are presented as pie charts. Interestingly, the distributions differ somewhat between the two quantification methods, with a higher percentage of endoplasmic reticulum (ER) proteins and ribonucleoprotein complexes in the labelled approach and more plasma membrane (PM) and cytosolic proteins in the label-free data. Close inspection of the marker profiles (Supplementary Fig. 1) showed that cytosolic proteins in the labelled data showed a small peak corresponding to the ER marker, which could indicate some cytosolic proteins being misclassified as ER. This is not seen in the label-free profiles and as the percentage of ER and cytosolic proteins combined are virtually the same (37% for TMT and 39% for label-free data, respectively), the distribution is therefore most likely closer to what is seen in the label-free case. The contrary can however be said for the ribonucleoprotein complexes and the mitochondria markers for the label-free dataset. Evidently, resolution needs improvement for separation of organelles in both experiments, which is elaborated on in the Discussion section.

Supplementary material related to this article found, in the online version, at http://dx.doi.org/10.1016/j.plantsci.2004.08.011.

3.2. Multivariate classification

Organelles represented by at least 10 markers were selected for further analysis that resulted in six organelle groups for the label-free data (cytosol/cytoplasm, ER, Golgi, mitochondrion, PM and ribonucleoprotein complexes) and four for the TMT data (cytosol/cytoplasm, ER, mitochondrion and ribonucleoprotein complexes). In pRoloc, the svmOptimize function was run with default settings and added class weights, resulting in a gamma value of 0.1 for both datasets and a cost of 8 and 16 for the TMT and label-free data, respectively. These parameters were subsequently input to the svmClassify function, again with class weights. Resulting organelle classifications and their respective scores were analysed based on precision, resulting in the minimum scores presented in Table 1. 461 proteins were assigned an organelle location for the label-free and 113 for the TMT dataset, respectively. Corresponding PCA plots can be seen in Fig. 5A and B, where triangular markers are classified proteins. Although not used for classification, all groups are represented in the PCA plots for comparison with Figs. 1 and 3 as can be seen, no immediate misclassifications are apparent. Fig. 6 shows the distribution of the organelle allocations, which roughly correspond to the marker distributions seen in Fig. 4. This can be expected, assuming the markers are representative of the organelle distribution of the respective datasets. The classifications were subsequently compared to GO annotations, were a large fraction of proteins with multiple locations were detected (Fig. 7).

4. DISCUSSION

Previous proteomic studies of organelle locations have mainly been focused on one compartment, thus suffering from the inability of assessing contamination. In this study, we perform quantification of all fractions from an organelle density-gradient

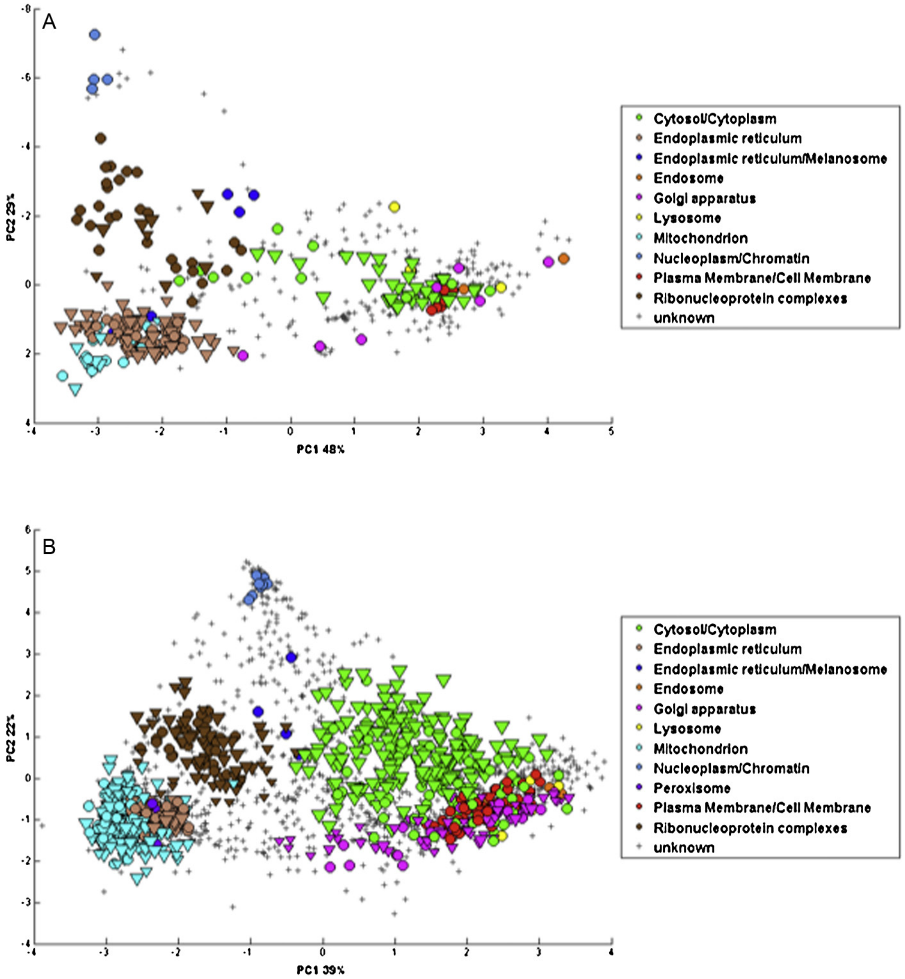

Fig. 5. Classification results from the SVM. Triangular markers denote classified proteins, the smaller the triangle, the lower the score. (A) TMT with the 101 original markers and 113 classified proteins. (B) Label-free with the 302 markers and 416 classified proteins.

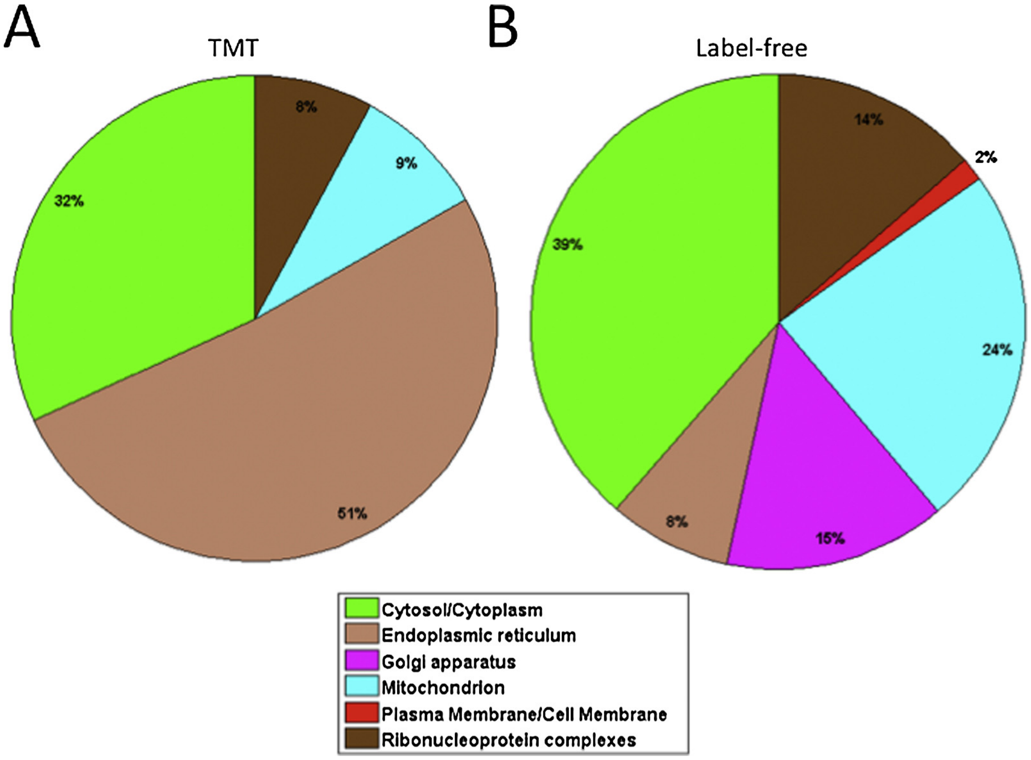

Fig. 6. Pie chart showing the distribution of all markers amongst the organelles.

and in total identify more than 1000 proteins from a number of organelles. This indicates that density-gradient experiments are valuable resources for studying multiple locations of proteins, which is one of the complications when working with organelles in the dynamic environment of a cell with its proteins trafficking between different compartments.

We have here successfully performed both a labelling experiment and a label-free analysis of the density-gradient fractions. 1697 proteins were identified with the label-free method

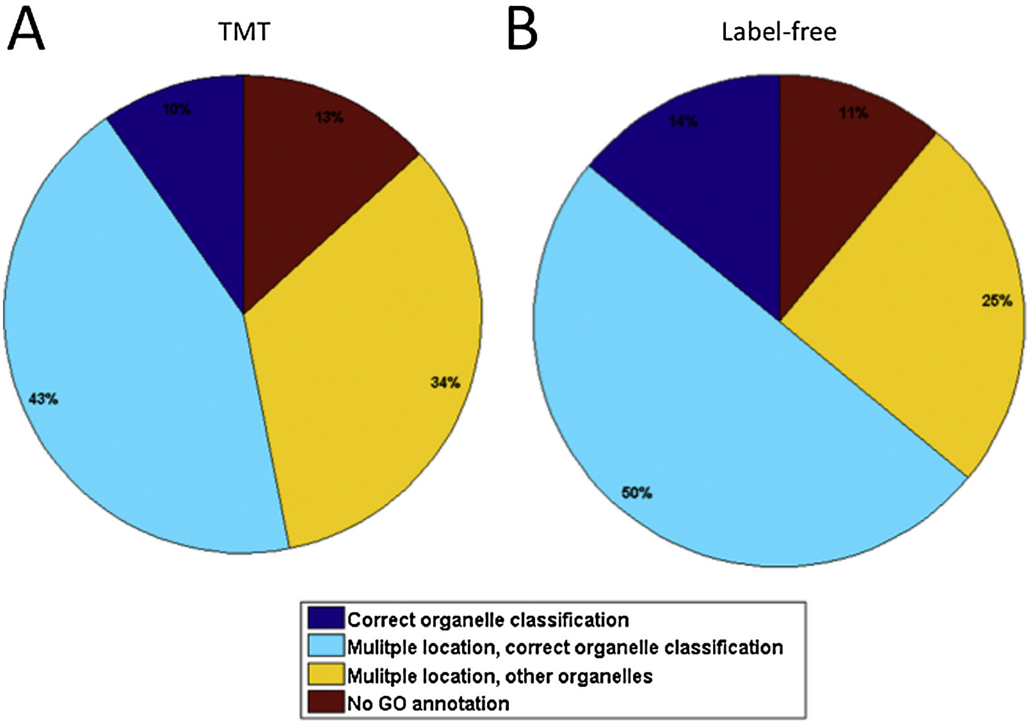

Fig. 7. Pie chart summarising the types of classifications obtained. Correct organelle classification: our classification concordant with GO annotation. Multiple location, correct organelle classification: our classification concordant with one out of several GO annotations. Multiple location, other organelles: our classification not concordant with GO annotation. No GO annotation: protein that had no previous GO annotation.

compared to only 439 with TMT labelling, consistent with previous studies in which label-free quantification gives a larger number of quantified proteins [14–16]. In our case, this is partly due to the experimental setup, in which we had two groups of 5 samples each for the labelling experiment. In order for a protein to obtain a distribution profile over the entire gradient, it had to be identified in both groups and was otherwise discarded. However, in general, quantification through labelling is thought to be more reliable in terms of accuracy and precision. This is consistent with the clearer separation of the manual markers seen in Fig. 1A compared to the label-free dataset in Fig. 1B. The lower number of identified markers did however result in less material for classification, limiting the applicability of this quantification technique for a LOPIT analysis. Increasing the marker cohort with GO annotations did nevertheless result in an equivalent ratio of organelle allocations; roughly a quarter of the original datasets were assigned subcellular location, demonstrating that for this type of organelle proteomic experiment, both quantification methods are able to distinguish between the major organelle proteins. The viability of using GO annotated markers has been discussed further in [28], where manually selected markers proved preferable, but selecting GO annotations based on uniqueness and experimental evidence showed similar results, albeit with some additional noise. Naturally, the number and especially quality of markers are extremely important for a successful experiment. Here, as mentioned above, the label-free quantification strategy produced considerably more markers, even though only proteins without missing values were selected for further analysis from the original dataset. More markers as well as a larger dataset could be acquired by e.g. a nearest neighbour imputation strategy [28], followed by a more advanced data filtering procedure for clearer separation and classification since comparing the assigned organelle locations to GO annotations (Figure 7) showed a large ratio of proteins with multiple locations. This is consistent with other studies estimating the multiple locations to be more than 50% of the proteins [29]. There was, however, a considerable portion of allocations not consistent with existing GO annotations. While this is not an error per se, manual investigation of the GO information showed a considerable amount of proteins having been classified to an organelle closely clustered in the PCA plots. The density of label-free data in particular could be alleviated by implementing a more elaborate data filtering approach, where only profiles containing pronounced peaks are analysed, such as in [4], where peaks were extracted by considering a fraction's intensity deviation from neighbouring fractions. Also, a more sophisticated outlier removal than presented here could be used for the GO markers, such as a clustering technique for removal of deviating profiles and/or including abundance estimates into the selection as higher abundance proteins are generally more reliable in terms of quantification.

In addition, the initial separation is crucial and density-gradient experimental designs often involve collection of at least 20 fractions [4,8]. Recently, a combination of gradients has been proposed to increase separation [13]. Nevertheless, this indicates that density gradient experiments are valuable resources for studying multiple locations of proteins when working with organelles in the dynamic environment of the cell with its proteins trafficking between different locations.

With the present design of (most) density-gradients proteomic experiments including defining organelle cut-off values from marker proteins, it is obvious that the data analysis is very dependent on the reliability of the marker proteins. This can be questioned in several ways, since there are major problems associated with discrepancies in organelle location databases. These discrepancies are due to the variance in type of sample, experimental conditions and techniques; leading to many false assignments. Many of the existing annotations today originate from GFP tagging experiments or antibody staining experiments. Tagging of proteins might change the location of a protein resulting in misleading organelle annotations [30] and since the antibody staining of a protein is often seen in dual locations, but to different extents, it can be difficult to determine the major location of a protein using antibody staining. Proteomic quantification by mass spectrometry on the other hand, is a fast and accurate technique without those previously mentioned limitations.

In conclusion, we argue that density-gradient approaches for organelle proteomics have the potential to become implemented in a majority of proteomic experiments in combination with careful quantification and data analysis. We argue that label-free quantification (high number of identifications) in combination with automatic marker selection (and processing) has the necessary prerequisites for biological inference in high-throughput proteomics experiments. The technique is easy to use with several options existing for protein quantification and can be adapted for different conditions, thereby generating important information that adds to the sought-after complete picture of the dynamic human cell.

Author contributions

LA carried out the experiments; MS the statistical analysis; FL the MS data analysis and PJ initiated and designed the study. The manuscript was written through contributions of all authors and all have given approval to the final version of the manuscript.

Funding sources

This work was supported by grants from the Knut and Alice Wallenberg Foundation, the Swedish Research Council, Vetenskapsrådet and the Swedish Strategic Research Council to CREATE Health.

Conflict of interest

The authors declare that they have no competing financial interests.

Supporting information

This material is available free of charge via the internet at http://pubs.acs.org. All MS data has been deposited at the EBI PRIDE database and the Swestore repository.

References

[1] C.M. Mulvey, S. Tudzarova, M. Crawford, G.H. Williams, K. Stoeber, J. Godovac-Zimmermann, Subcellular proteomics reveals a role for nucleo-cytoplasmic trafficking at the DNA replication origin activation checkpoint, J. Proteome Res. 12 (2013) 1436–1453, doi:http://dx.doi.org/10.1021/pr3010919.

[2] L.A. Huber, K. Pfaller, I. Vietor, Organelle proteomics: implications for subcellular fractionation in proteomics, Circ. Res. 92 (2003) 962–968, doi: http://dx.doi.org/10.1161/01.RES.0000071748.48338.25.

[3] T.P.J. Dunkley, R. Watson, J.L. Griffin, P. Dupree, K.S. Lilley, Localization of organelle proteins by isotope tagging (LOPIT), Mol. Cell. Proteomics 3 (2004) 1128–1134, doi:http://dx.doi.org/10.1074/mcp.T400009-MCP200.

[4] L.J. Foster, C.L. de Hoog, Y. Zhang, Y. Zhang, X. Xie, V.K. Mootha, et al., A mammalian organelle map by protein correlation profiling, Cell 125 (2006) 187–199, doi:http://dx.doi.org/10.1016/j.cell.2006.03.022.

[5] F.-M. Boisvert, Y. Ahmad, M. Gierliński, F. Charrière, D. Lamont, M. Scott, et al., A quantitative spatial proteomics analysis of proteome turnover in human cells, Mol. Cell. Proteomics 11 (2012) , doi:http://dx.doi.org/10.1074/mcp. M111.011429 M111.011429.

[6] M. Bantscheff, B. Kuster, Quantitative mass spectrometry in proteomics, Anal. Bioanal. Chem. 404 (2012) 937–938, doi:http://dx.doi.org/10.1007/s00216-012-6261-7.

[7] N. Dephoure, S.P. Gygi, Hyperplexing: method for higher-order multiplexed quantitative proteomics provides a map of the dynamic response to rapamycin

in yeast, Sci. Signal 5 (2012) rs2, doi:http://dx.doi.org/10.1126/scisignal.2002548.

[8] T.P.J. Dunkley, S. Hester, I.P. Shadforth, J. Runions, T. Weimar, S.L. Hanton, et al., Mapping the *Arabidopsis* organelle proteome, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 6518–6523, doi:http://dx.doi.org/10.1073/pnas.0506958103.

[9] D.J.L. Tan, H. Dvinge, A. Christoforou, P. Bertone, A. Martinez Arias, K.S. Lilley, Mapping organelle proteins and protein complexes in *Drosophila melanogaster*, J. Proteome Res. 8 (2009) 2667–2678, doi:http://dx.doi.org/10.1021/pr800866n.

[10] S.L. Hall, S. Hester, J.L. Griffin, K.S. Lilley, A.P. Jackson, The organelle proteome of the DT40 lymphocyte cell line, Mol. Cell Proteomics 8 (2009) 1295–1305, doi:http://dx.doi.org/10.1074/mcp.M800394-MCP200.

[11] J.S. Andersen, C.J. Wilkinson, T. Mayor, P. Mortensen, E.A. Nigg, M. Mann, Proteomic characterization of the human centrosome by protein correlation profiling, Nature (2003) 570–574, doi:http://dx.doi.org/10.1038/nature02166.

[12] L. Gatto, J.A. Vizcaíno, H. Hermjakob, W. Huber, K.S. Lilley, Organelle proteomics experimental designs and analysis, Proteomics 10 (2010) 3957–3969, doi:http://dx.doi.org/10.1002/pmic.201000244.

[13] M.W.B. Trotter, P.G. Sadowski, T.P.J. Dunkley, A.J. Groen, K.S. Lilley, Improved sub-cellular resolution via simultaneous analysis of organelle proteomics data across varied experimental conditions, Proteomics 10 (2010) 4213–4219, doi:http://dx.doi.org/10.1002/pmic.201000359.

[14] Z. Li, R.M. Adams, K. Chourey, G.B. Hurst, R.L. Hettich, C. Pan, Systematic comparison of label-free, metabolic labeling, and isobaric chemical labeling for quantitative proteomics on LTQ Orbitrap Velos, J. Proteome Res. 11 (2012) 1582–1590, doi:http://dx.doi.org/10.1021/pr200748h.

[15] V.J. Patel, K. Thalassinos, S.E. Slade, J.B. Connolly, A. Crombie, J.C. Murrell, et al., A comparison of labeling and label-free mass spectrometry-based proteomics approaches, J. Proteome Res. 8 (2009) 3752–3759, doi:http://dx.doi.org/10.1021/pr900080y.

[16] H. Wang, S. Alvarez, L.M. Hicks, Comprehensive comparison of iTRAQ and label-free LC-based quantitative proteomics approaches using two *Chlamydomonas reinhardtii* strains of interest for biofuels engineering, J. Proteome Res. 11 (2012) 487–501, doi:http://dx.doi.org/10.1021/pr2008225.

[17] S. Mahadevan, S.L. Shah, T.J. Marrie, C.M. Slupsky, Analysis of metabolomic data using support vector machines, Anal. Chem. 80 (2008) 7562–7570, doi:http://dx.doi.org/10.1021/ac800954c.

[18] M. Sandin, A. Chawade, F. Levander, Is label-free LC–MS/MS ready for biomarker discovery? Proteomics Clin. Appl. 9 (2015) 289–294, doi:http://dx.doi.org/10.1002/prca.201400202.

[19] J. Häkkinen, G. Vincic, O. Månsson, K. Wårell, F. Levander, The proteios software environment: an extensible multiuser platform for management and analysis of proteomics data, J. Proteome Res. 8 (2009) 3037–3043, doi:http://dx.doi.org/10.1021/pr900189c.

[20] L. Martens, M. Chambers, M. Sturm, D. Kessner, F. Levander, J. Shofstahl, et al., mzML – a community standard for mass spectrometry data, Mol. Cell. Proteomics 10 (R110) (2011) 133, doi:http://dx.doi.org/10.1074/mcp.R110.000133.

[21] D. Kessner, M. Chambers, R. Burke, D. Agus, Mallick P. ProteoWizard, Open source software for rapid proteomics tools development, Bioinformatics 24 (2008) 2534–2536, doi:http://dx.doi.org/10.1093/bioinformatics/btn323.

[22] R. Craig, R.C. Beavis, TANDEM: matching proteins with tandem mass spectra, Bioinformatics 20 (2004) 1466–1467, doi:http://dx.doi.org/10.1093/bioinformatics/bth092.

[23] F. Levander, M. Krogh, K. Wårell, P. Gärdén, P. James, J. Häkkinen, Automated reporting from gel-based proteomics experiments using the open source Proteios database application, Proteomics 7 (2007) 668–674, doi:http://dx.doi.org/10.1002/pmic.200600814.

[24] M. Bellew, M. Coram, M. Fitzgibbon, M. Igra, T. Randolph, P. Wang, et al., A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC–MS, Bioinformatics 22 (2006) 1902–1909, doi:http://dx.doi.org/10.1093/bioinformatics/btl276.

[25] M. Sandin, A. Ali, K. Hansson, O. Månsson, E. Andreasson, S. Resjö, et al., An adaptive alignment algorithm for quality-controlled label-free LC–MS, Mol. Cell. Proteomics 12 (2013) 1407–1420, doi:http://dx.doi.org/10.1074/mcp.O112.021907.

[26] L.M. Breckels, L. Gatto, A. Christoforou, A.J. Groen, K.S. Lilley, M.W.B. Trotter, The effect of organelle discovery upon sub-cellular protein localisation, J Proteomics 88 (2013) 129–140, doi:http://dx.doi.org/10.1016/j.jprot.2013.02.019.

[27] L. Gatto, L.M. Breckels, S. Wieczorek, T. Burger, K.S. Lilley, Mass-spectrometry-based spatial proteomics data analysis using pRoloc and pRolocdata, J. Gerontol. 30 (2014) 1322–1324, doi:http://dx.doi.org/10.1093/bioinformatics/btu013.

[28] L. Gatto, L.M. Breckels, T. Burger, D.J.H. Nightingale, A.J. Groen, C. Campbell, et al., A foundation for reliable spatial proteomics data analysis, Mol. Cell. Proteomics 13 (2014) 1937–1952, doi:http://dx.doi.org/10.1074/mcp.M113.036350.

[29] A.T. Qattan, C. Mulvey, M. Crawford, D.A. Natale, J. Godovac-Zimmermann, Quantitative organelle proteomics of MCF-7 breast cancer cells reveals multiple subcellular locations for proteins in cellular functional processes, J. Proteome Res. 9 (2010) 495–508, doi:http://dx.doi.org/10.1021/pr9008332.

[30] G.J. Phillips, Green fluorescent protein – a bright idea for the study of bacterial protein localization, FEMS Microbiol. Lett. 204 (2000) 9–18, doi:http://dx.doi.org/10.1016/S0378-1097(01)00358-5.